REPORT
Image Based Deepfake Detection System

https://github.com/Pmalik24/DeepfakeDetectionSystem

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 also incorporate Swin Transformers, which build upon the ViT architecture but offer superior performance through a hierarchical approach that captures both local and global features, enhancing the model's ability to detect complex manipulations. By comparing these models with Convolutional Neural Networks (CNNs), we assess performance across an expanded dataset, including Face Forensic ++ dataset and images generated by Generative Adversarial Networks (GANs). We also deployed the system to enable real-world application of these detection techniques. 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 Modeling

2.1.1. Swin Transformer – What & Why

The SWIN Transformer is an advanced vision transformer designed for efficient high-resolution image processing through a hierarchical structure. Unlike traditional transformers like ViT, which use a flat architecture, SWIN divides image processing into four stages, capturing features at varying scales. This design allows it to focus on both local details, such as texture inconsistencies, and global contexts, such as lighting or alignment anomalies, which are crucial for deepfake detection.

A key innovation of SWIN is its shifted window mechanism as shown in Fig.1, where self-attention is computed within non-overlapping windows that shift across layers, enabling feature sharing between neighboring regions. This approach is computationally efficient while capturing cross-regional relationships, essential for identifying subtle deepfake artifacts. Compared to ViT, which struggles to balance local and global information, SWIN integrates both efficiently, making it more effective for deepfake detection. Its adaptability and efficiency make it a superior choice for identifying manipulated content.

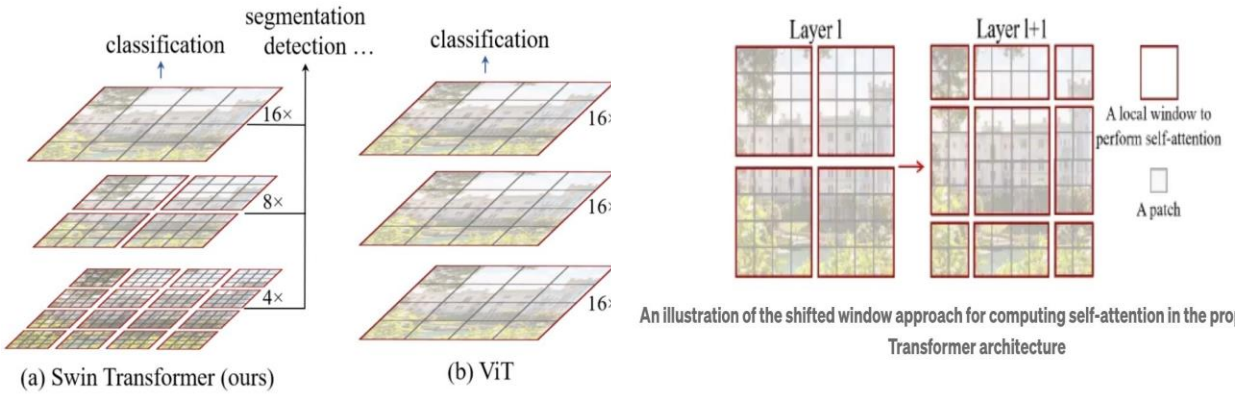

Fig 1. Swin Transformer and ViT Architectures

2.1.2 SWIN Transformer – Architecture and Modifications

The SWIN Transformer uses a hierarchical structure with four stages that progressively reduce resolution while increasing feature complexity. In the first two stages, low- and mid-level features like edges and textures are extracted, while the later stages focus on high-level, task-specific features for deepfake detection. Patch merging between stages ensures efficient processing as shown in Fig.2.

For deepfake detection, the early stages were frozen to retain pretrained ImageNet features, ensuring computational efficiency and preventing overfitting. The later stages were fine-tuned to detect anomalies like pixel distortions and lighting irregularities. The classification head was adapted for binary classification, aligning the output with the deepfake detection task. Key aspects, such as the hierarchical structure and patch merging, were retained to preserve the model's ability to capture multi-scale features, ensuring an efficient adaptation for deepfake detection.

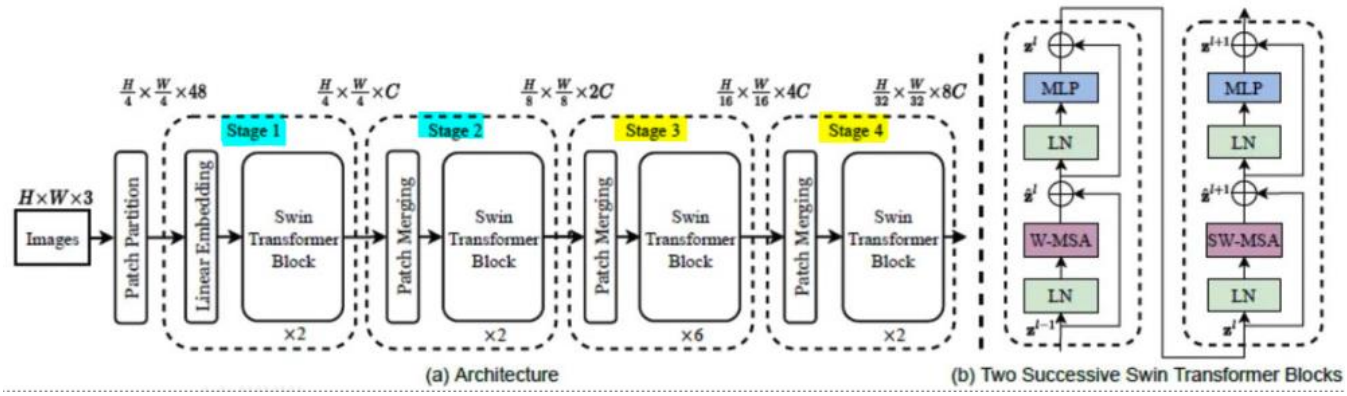

Fig 2. Swin Transformer Architecture

2.2. Training Performance Analysis: CNN vs. ViT vs. Swin

The performance of CNN, ViT, and Swin models during training provides key insights into their respective architectures and suitability for deepfake detection. The analysis focuses on the stability of loss curves, generalization capabilities, and computational efficiency, revealing the strengths and limitations of each model.

2.2.1. CNN: Lightweight but Prone to Overfitting

The CNN model, while efficient, shows a notable gap between training and validation loss as shown in Fig.3, indicating overfitting to training data. This limits its ability to generalize, especially in tasks like deepfake detection, where nuanced manipulations demand capturing fine-grained details. Its unstable validation performance further highlights its inadequacy for complex datasets, reducing reliability in real-world scenarios.

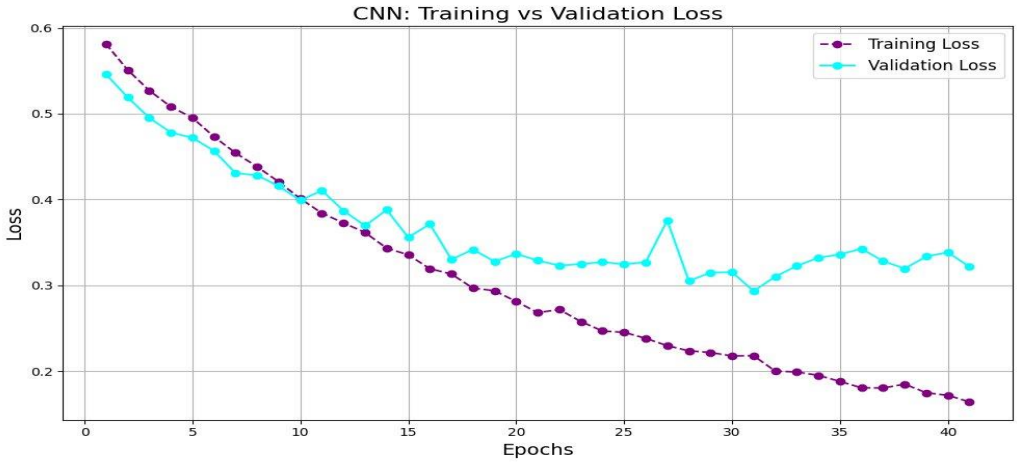

Fig 3. CNN: Training vs Validation Loss

2.2.2 ViT: Improved Generalization through Self-Attention

ViT utilizes self-attention to capture long-range dependencies, showing improved alignment in training and validation loss compared to CNN, indicating better generalization, as shown in Fig.4. However, its focus on recall limits balanced performance across metrics like precision and F1 score. The moderate gap in loss curves and lack of localized context awareness, essential for detecting subtle deepfake artifacts, leave it trailing behind the Swin model.

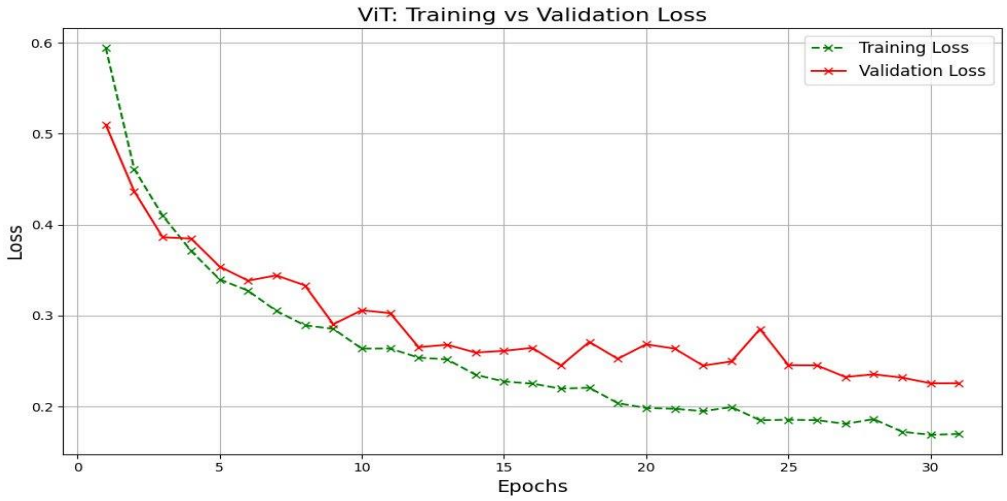

Fig 4. ViT: Training vs Validation Loss

2.2.3 Swin Transformer: A Hierarchical Transformer for Superior Performance

The Swin Transformer outperforms CNN and ViT due to its hierarchical structure with shifted windows, allowing it to capture both global and local features. This makes it ideal for deepfake detection, where fine details are crucial. Its stable training and validation loss curves reflect strong generalization, as shown in Fig.5, while dual optimization of F1 score and recall improves overall

performance. Despite a longer training time, the Swin model's superior accuracy justifies the extra computational cost.

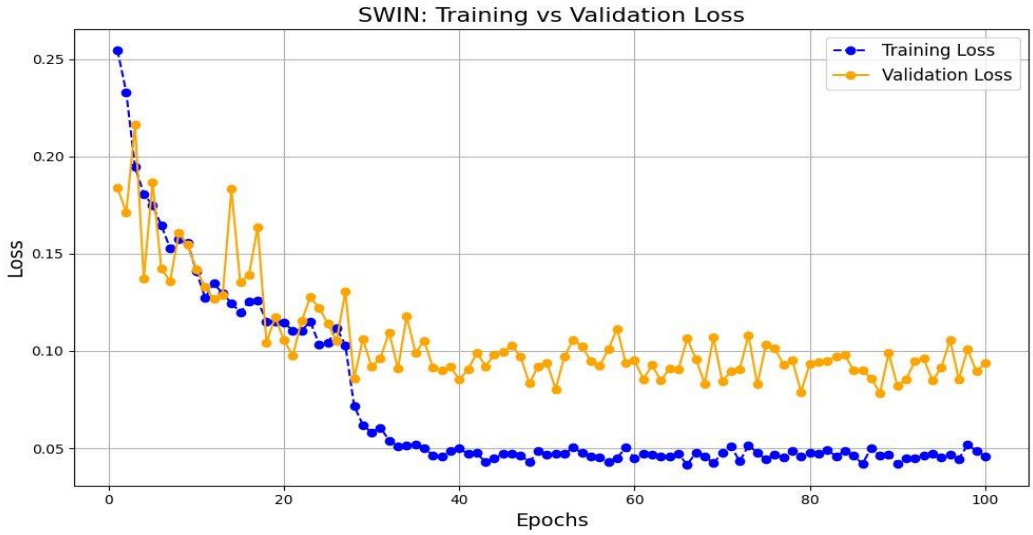

Fig 5. SWIN: Training vs Validation Loss

In conclusion, while CNN and ViT offer efficiency, the Swin Transformer's architectural strengths and robust performance make it the best choice for deepfake detection, even with higher computational demands

2.3. Inference Pipeline

The inference pipeline is designed to process input images, pass them through one of the selected models (CNN, ViT, or Swin Transformer), and return predictions on whether the image is real or fake. The pipeline follows these steps:

- **Image Preprocessing:**
  Each image is resized to the required dimensions (e.g., 224x224 pixels) and normalized using ImageNet's mean and standard deviation values ([0.485, 0.456, 0.406] and [0.229, 0.224, 0.225]) to match the model's expectations.
- **Model Loading:**
  Pre-trained weights for each model (CNN, ViT, and Swin Transformer) are loaded from saved paths. The models are moved to the appropriate device (GPU or CPU) and set to evaluation mode to disable training-specific layers.
- **Model Inference:**
  The preprocessed image is passed through the model, generating an output vector representing the likelihood of each class. The class with the highest probability is selected as the model's prediction (real or fake).
- **Post-Processing:**
  The output is analyzed to determine the final label. For binary classification, the highest probability class is chosen, with "real" (1) or "fake" (0) being the result.

3. Results

In phase 1, both the CNN and ViT models were trained for deepfake detection. While CNN minimized false positives, it struggled with missed detections. The ViT model, on the other hand, showed better recall and a stronger ability to identify fake images, though it had more false positives. Overall, ViT proved to be the better choice for deepfake detection.

The Swin model's performance and its advantages over ViT will be discussed in this section.

3.1. F1 Optimized SWIN Vs. Recall Optimized SWIN

In evaluating the performance of the Swin model for deepfake detection, two optimization strategies—recall-optimized and F1-optimized—were compared. The following metrics, as shown in Fig.6. provide insights into the trade-offs and strengths of each approach.

Metric	Recall Optimized	F1 Optimized
Recall	97.98%	96.37%
Precision	90.34%	96.28%
F1-Score	94.01%	96.32%
Accuracy	93.75%	96.32%
ROC	.9900	.9957

Fig 6. Table showing comparison between Recall Optimized and F1 Optimized

The F1-optimized Swin model provides a more balanced performance, making it better suited for real-world deepfake detection. While the recall-optimized model slightly reduces false negatives (48 vs. 64), this comes at the cost of significantly higher false positives (201 vs. 71). The F1-optimized model achieves higher precision (0.9628 vs. 0.9034), F1 score (0.9632 vs. 0.9401), and accuracy (0.9632 vs. 0.9375), ensuring both real and fake images are classified more reliably. This balance minimizes the operational challenges of excessive false alarms while maintaining strong recall, making it the preferred choice for practical and scalable deployment.

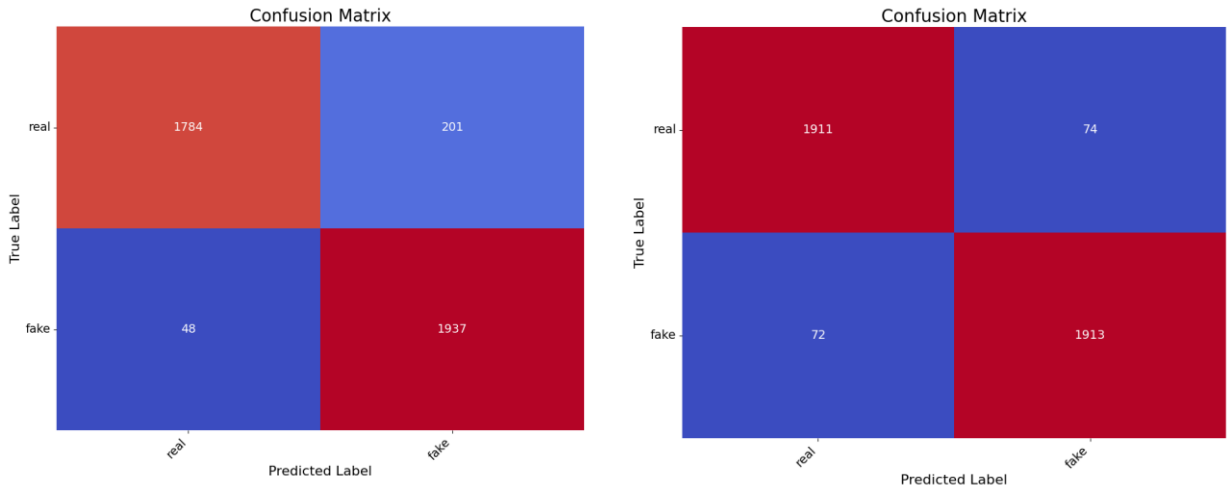

Fig 7. Recall Optimized SWIN Confusion Matrix Vs F1 Optimized SWIN Confusion Matrix

The F1-optimized Swin model outperforms the recall-optimized version, as seen in Fig 7. While the recall-optimized model reduces false negatives (48 vs. 72), it increases false positives (201 vs. 74), creating operational inefficiencies. The F1-optimized model strikes a balance by minimizing both false positives and negatives, ensuring reliable and efficient deepfake detection. This balance is crucial for accurate detection without overwhelming the system with false alarms, making the F1-optimized model the best choice.

3.2. SWIN Vs ViT: Massive Performance Gains

The comparison of Swin(F1-Optmized) and ViT models, as shown in Fig.8 reveals significant improvements in performance across various metrics, highlighting Swin's superior capabilities in deepfake detection.

Metric	ViT	SWIN
Recall	92.1%	96.37%
Precision	86.4%	96.28%
F1-Score	89.2%	96.32%

Accuracy	88.8%	96.32%
ROC	.9640	.9957

Fig 8 . Table showing comparison between ViT and Swin

The Swin model demonstrates a significant performance improvement over ViT, evident in both metrics and confusion matrices. Swin achieves a recall of 96.37% and precision of 96.28%, compared to ViT's recall of 92.1% and precision of 86.4%, marking an increase of 4.27% in recall and 9.88% in precision for Swin.

This improvement highlights Swin's ability to reduce both false positives and false negatives effectively, which is a critical factor for deepfake detection. The F1 score and accuracy of Swin (96.32% each) also far surpass ViT's scores of 89.2% and 88.8%, respectively, indicating a 7.12% increase in F1 score and a 7.44% increase in accuracy. Additionally, Swin achieves a ROC-AUC of 0.9957, outperforming ViT's 0.9640, further emphasizing its superior classification capability.

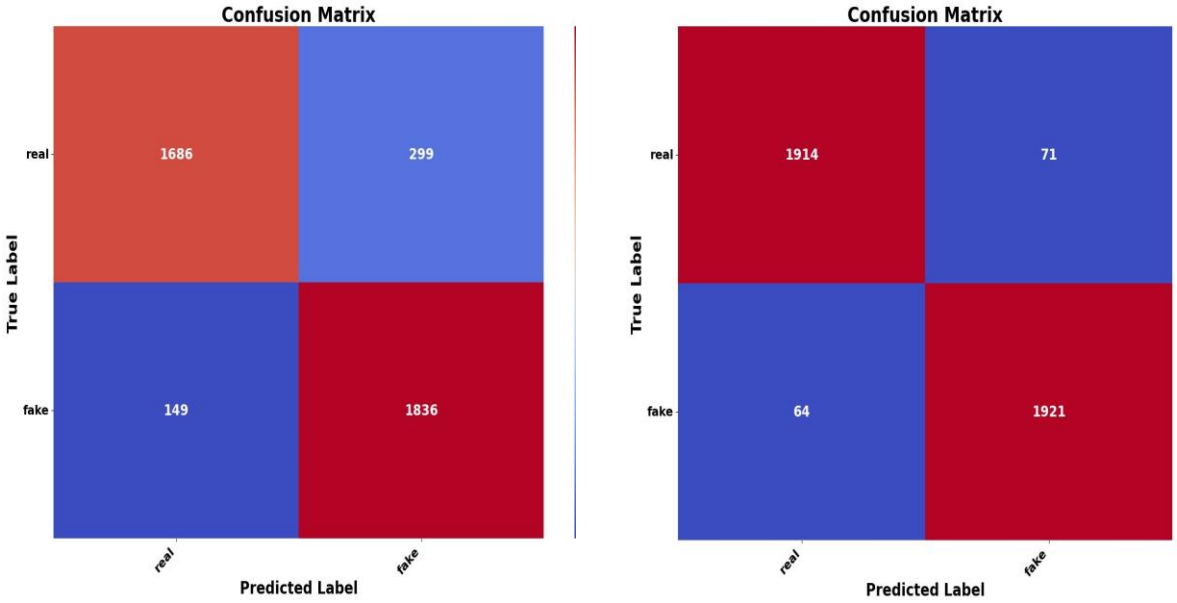

Fig 9. ViT Vs. SWIN Confusion Matrices

The confusion matrices highlight Swin's superiority, reducing false positives from 299 (ViT) to 71 and false negatives from 149 to 64, as shown in Fig.9 This improvement is due to Swin's hierarchical architecture, which combines localized self-attention with shifted windows, effectively capturing both fine-grained and global features. In contrast, ViT's reliance on global self-attention limits its ability to detect subtle deepfake artifacts. Swin's enhanced generalization and reduced errors make it ideal for applications like fraud detection and content moderation, where precision and recall are critical.

3.3 Classification Examples of our Best Model – SWIN Transformer

The **correctly classified fake images,** as shown in Fig.10 highlight the Swin model's ability to detect deepfake artifacts, such as unnatural textures, blurred features, and inconsistent lighting. For instance, distortions in facial structure and edge blending in the third image were accurately flagged as fake, showcasing the model's effective use of localized self-attention.

Conversely, **misclassified real images,** as shown in Fig.11 reveal challenges with atypical patterns like motion blur, low resolution, or exaggerated features. For example, motion blur in one image mimicked deepfake distortions, while high-contrast makeup in another triggered a false classification. These cases highlight the model's robustness but also its vulnerability to real-world complexities resembling synthetic anomalies.

Fig 10. Correctly Classified Fake Images

Fig 11. Misclassified Real Images

The **correctly classified real images,** as shown in Fig 12 highlight the Swin model's ability to detect authentic features across diverse facial structures, lighting, and expressions. Its hierarchical attention mechanism effectively captures subtle consistencies in real images, ensuring reliable identification.

Misclassified fake images reveal, as shown in Fig 13 challenges in detecting subtle manipulations. For instance, distorted edges in one image were overlooked due to sharpness, while dynamic movement in another confused the model's focus on standard facial features. Minimal facial manipulation in a sports-related image also made it appear more authentic, demonstrating the difficulty in identifying subtle fakes.

Fig 12. Correctly Classified Real Images

Fig 13. Misclassified Fake Images

3.4. Deployment

The deployment of the Deepfake Classification Web Application involved integrating a React frontend with a Flask backend to create an interactive and user-friendly experience. The frontend, developed using React and styled with Material UI, provides a modern interface for users to upload

an image and select one of three classification models (CNN, ViT and Swin) for deepfake detection. The backend, implemented in Flask, handles the processing of user input image followed by invoking the appropriate pre-trained classification model and finally, returns the prediction results as either "Real/Fake". REST APIs is utilized to facilitate seamless communication between the frontend and backend, ensuring a smooth and efficient workflow.

The process began with setting up the Flask environment, where the Python-based models were hosted and exposed through API endpoints. Simultaneously, the React application was built, incorporating Material UI components to design a responsive layout featuring a vertical card for image previews, a dropdown for model selection, a button for image upload and a button for classifying the image as real or fake.

A few snapshots of the UI :

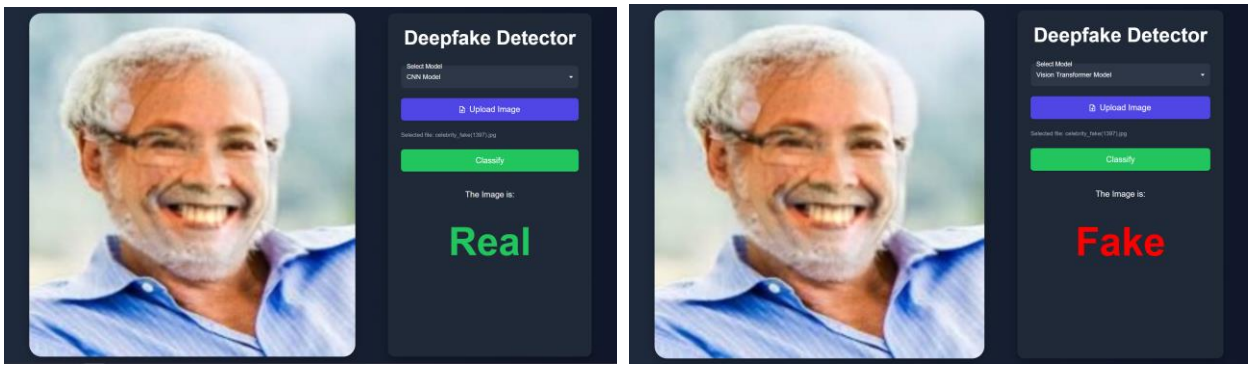

Fig . 14 Snapshots of the Deepfake Detection UI.

As shown in Figure 14, the same image was given as input to both the CNN and Vision Transformer models. The CNN model predicted the image as Fake, while the ViT model correctly identified it as Real.

3.5. Significance of Results

The outcomes of our deepfake detection efforts highlight the urgent need for effective solutions to address the growing prevalence of manipulated media across digital platforms. By evaluating the performance of CNN, ViT, and Swin Transformer models, we have provided critical insights into effective detection methods that are essential for mitigating misinformation in sensitive areas such as political campaigns and social media discourse.

Our results demonstrate that the Swin Transformer is the most effective model for deepfake detection, offering superior accuracy and efficiency. This makes it an invaluable tool for organizations aiming to preserve content authenticity and counter the impact of manipulated media. Implementing such reliable detection systems is crucial for maintaining trust in digital platforms, protecting reputations, and ensuring the public has access to accurate information.

4. Discussions

4.1 All Models Comparisons

This analysis compares the performance of three deepfake detection models—Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Swin Transformers—using key evaluation metrics, as shown in Fig 15

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**

Fig 15. Table showing comparison between CNN ,ViT and Swin

- **Recall:**
  CNN achieved a recall of 86.1%, ViT improved to 92.1%, and Swin Transformer reached the highest at 96.37%, proving its superior ability to identify deepfakes.
- **Precision:**
  While CNN recorded 87.2% and ViT slightly lower at 86.4%, Swin Transformer excelled with 96.28%, minimizing classification errors effectively.
- **F1-Score:**
  CNN's F1-score was 86.6%, ViT improved to 89.2%, and Swin Transformer achieved 96.32%, showcasing the best balance between recall and precision.
- **ROC AUC:**
  CNN scored 0.95, ViT improved to 0.9642, and Swin Transformer delivered near-perfect discrimination with 0.9957, highlighting its robustness.

Best Model

Across all metrics, the **Swin Transformer** emerged as the best-performing model, demonstrating superior recall, precision, F1-score, and ROC AUC. Its hierarchical architecture enables the capture of both local and global features, making it highly effective in detecting complex deepfake manipulations.

4.2 Strengths and Weaknesses of all the models

The strengths and weaknesses of each model are evaluated based on key metrics such as recall, precision, F1-score, and ROC AUC. These metrics provide a clear comparison of their performance in detecting deepfakes, highlighting where each model excels and where improvements are needed.

4.2.1. Convolutional Neural Network

The strengths and weaknesses of the CNN model are discussed below:

- **Strengths:**
  CNN demonstrates strong precision at 87.2%, effectively identifying real images without misclassifying them as fake. Its computational efficiency and suitability for smaller datasets are notable advantages.
- **Weaknesses:**
  With a recall of 86.1%, CNN misses more fake images compared to other models, reducing its reliability in detecting deepfakes. The F1-score of 86.6% reflects its less balanced performance between precision and recall relative to ViT and Swin.

4.2.2. Vision Transformer

The strengths and weaknesses of the ViT model are discussed below:

- **Strengths:**
  ViT outperforms CNN in recall, achieving 92.1%, which enables it to detect more fake images. Its F1-score of 89.2% indicates a better balance between precision and recall, while the ROC AUC score of 0.9642 demonstrates strong capability in distinguishing real from fake content.
- **Weaknesses:**
  ViT’s precision of 86.4% is lower than that of CNN and Swin, leading to more false positives. While its recall is strong, it still falls short of the superior performance achieved by Swin.

4.2.3. Swin Transformer

The strengths and weaknesses of the SWIN model are discussed below:

- **Strengths:**
  Swin Transformer achieves the highest recall (96.37%) and precision (96.28%), making it the most reliable model for detecting fake images and minimizing false positives. The F1-score of 96.32% highlights its excellent balance between recall and precision, and the ROC AUC score of 0.9957 underscores its superior ability to distinguish between real and manipulated content.
- **Weaknesses:**
  Despite its exceptional performance, some fake images are still misclassified, indicating room for improvement in edge cases.

5. Conclusion

In conclusion, the project effectively evaluated the performance of advanced machine learning models—CNN, ViT, and Swin Transformer—in detecting deepfakes. Among these, the Swin Transformer consistently outperformed the others across all key metrics, including recall, precision, F1-score, and ROC AUC, proving to be the most reliable model for deepfake detection. This work underscores the importance of utilizing advanced models to address the growing challenges of misinformation and ensure the authenticity of digital media. The results provide meaningful guidance for selecting the most effective model for efficient and accurate deepfake detection, particularly in high-stakes environments such as social media and political spheres.

6. Statement of Contribution

- **Parth Malik**: Was responsible for the implementation of Swin Transformers, including fine-tuning the model and optimizing it for deepfake detection tasks.
- **Shushant Ghosh:** Led the deployment process, ensuring that the trained models were effectively integrated into a functional and accessible application.
- **Suprajasai Konegari**: Focused on developing the inference pipeline and conducted a detailed comparative analysis of the three models (CNN, ViT, and Swin Transformer) to evaluate their strengths and limitations for deepfake detection.
- **Supriya Konegari**: Collaborated on the development of the inference pipeline, ensuring accurate predictions and smooth integration with the deployment process.

7. References

[1] Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo*; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, https://doi.org/10.48550/arXiv.2103.14030 .

[2] C. -M. Fan, T. -J. Liu and K. -H. Liu, "SUNet: Swin Transformer UNet for Image Denoising," *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, Austin, TX, USA, 2022, pp. 2333-2337, doi: 10.1109/ISCAS48785.2022.9937486.

[3] Hafsa Ilyas, Ali Javed, Khalid Mahmood Malik,AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection,Applied Soft Computing,Volume 136,2023,110124,ISSN 1568-4946,https://doi.org/10.1016/j.asoc.2023.110124.

[4] 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.

[5] 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.