ADVANCED DEEPFAKE DETECTION LEVERAGING SWIN TRANSFORMER TECHNOLOGY

Soumya Ranjan Mishra[1] – Hitesh Mohapatra[1*] – Seyed Ahmad Edalatpanah[2] – Mahendra Kumar Gourisaria[1]

[1]Computer Science and Engineering, KIIT (Deemed to be) University, Bhubaneswar, India
[2]Department of Applied Mathematics, Ayandegan Institute of Higher Education, Tonekabon, Iran

ARTICLE INFO

***Article history*:**
Received: 02.08.2024.
Received in revised form: 23.09.2024.
Accepted: 06.10.2024.

Keywords:
Deepfake
Image classification
SWIN trans- formers
fake image generation
image detection,
Hierarchical Representation
Transformer Block
Quadratic Complexity
SWIN Transformer blocks
Object Detection

DOI: 10.30765/er.2583

Abstract:

The widespread use of deepfake technology in recent years has made it extremely difficult to differentiate between real and fake images, usually AI-generated images. Effective detection techniques are desperately needed because one can generate fake images and spread them with ease. This research paper examines how effective the SWIN Transformer, a new transformer-based architecture, is for detecting deep fake images. The foundation of the suggested detection framework is an architecture made up of bottleneck, encoder, and decoder parts which is a type of SWIN transformer. It uses various self-attention mechanisms and advanced features to analyse the images closely whether it is a real image or a deepfake one. It relies on the concept of shifted windows during the processing of the images and is considered more effective than the traditional CNN methods. Our test results show how well the SWIN Transformer-based method performs in precisely recognizing deep fake images. The accuracy is found to be 97.91\% for CelebDF dataset and 95.715\% for FF++ dataset. The AUC for the newly modelled SWIN transformer is 0.99 and 0.9625 for CelebDF and FF++ datasets respectively. The Log Loss was calculated to be 0.034 for CelebDF dataset and 0.1573 for FF++ dataset. The proposed methodology not only enhances the accuracy of detecting manipulated images but also offers potential for scalable and efficient deployment in real-world scenarios where the proliferation of deepfakes presents significant challenges to maintaining trust and authenticity in visual media.

1 Introduction

Human faces play a crucial role in communication, association of information, and identity in human civilization. From access control and payment, to unlocking our phones, face recognition is an inevitable part of our life now. They manipulate facial images to commit fraud and pose as genuine users. This type of manipulation has become ubiquitous and raises eyebrows specifically in social media content. The level at which realism has been achieved in face synthesis is truly alarming. In recent years, advanced deep learning technologies have led to the rise of these deepfakes. These are highly realistic fake images and videos created using artificial intelligence [1]. They pose a big challenge to the credibility of digital content because they can make it seem like people are doing or saying things they never actually did. This creates an urgent need for effective ways to spot and reduce the spread of deepfake content. To tackle this problem, researchers are exploring different methods for detecting deepfakes. Deepfake is primarily a face-swapping algorithm that makes use of Neural Networks to create new images [2]. The facial features are mapped from one image to the other giving it a realistic look. The creation of deepfake includes an encoder, a bottleneck and a decoder [3]. The encoder compresses the original image by reducing its dimensions. The bottleneck produces the

* Corresponding author
E-mail address: hiteshmahapatra@gmail.com

compressed representation of our data. Following the bottleneck, we have the decoder which takes in the vector and turns it into the full-sized image. So input is taken from the encoder which is then reconstructed back. Figure.1 represents the general architecture of deepfake.

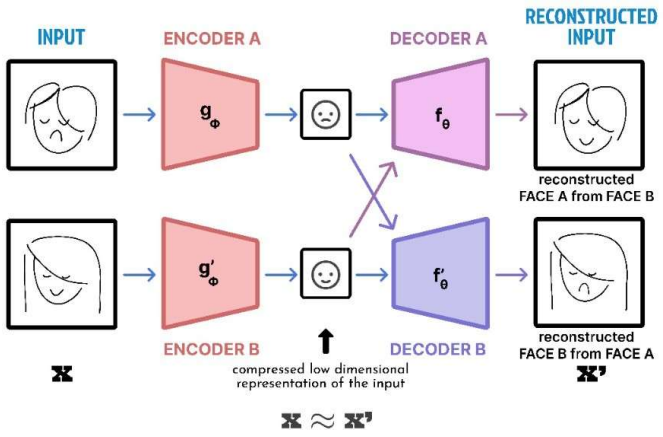

Figure 1. General DeepFake Architecture

Many research works have been proposed. Initial works detect the tampering through visual biological artifacts like inconsistent head poses and unnatural eye blinking. With the rise of learning-based methods, some studies have developed frameworks that extract features from spatial domains. These frameworks have shown excellent performance on specific datasets. A few methods detect forged faces through Spatial, Steganalysis, and Temporal features. This adds a stream of simplified Xception with a constrained convolution layer and an LSTM [4]. Many current approaches to deep fake detection oversimplify the problem by treating it as a straightforward binary classification task [5]. They focus on creating advanced feature extractors and then use a simple method to distinguish between real and fake faces. However, the photo-realistic counterfeits bring significant challenges to this binary classification framework. The deepfake detection problem has hence been redefined as a fine-grained classification problem. A promising approach is using SWIN Transformers, a type of deep learning model. They use self-attention mechanisms and advanced feature analysis to closely examine images. This helps capture both the overall context and fine details accurately. Our research focuses on understanding how SWIN Trans-formers work and how well they can identify AI-generated images, particularly deepfakes. We'll study the inner workings of the SWIN Transformer model and test how reliable it is at spotting deepfakes across different datasets and situations. The goal is to provide useful insights into computer vision and deepfake detection. By studying SWIN Transformers, we hope to give people better tools to fight against fake media and promote trust and honesty in digital platforms [7]. Here, we will proceed to critically analyse SWIN transformers, it is nothing but a significant and powerful innovation of vision transformers (ViT). Transformers 'exceptional performance has been demonstrated in various computer vision tasks, such as instance segmentation, image classification, and object detection [8]. The study uses machine learning algorithms to investigate the relationship between overall health, blood pressure and stroke risk. The study also analyses databases of stroke patients and reviews the literature to assess the impact of health indicators on stroke risk and evaluate the effectiveness of identified algorithms Findings aims to improve seizure prevention, treatment, and diagnostic tools and to help researchers understand algorithm performance for seizure prediction [26].

Traditional transformers lack the ability to process images patch by patch. This is where the SWIN Transformer comes in, it divides the image into non-overlapping shifted windows to initiate efficient and scalable computation [9]. The problem of quadratic complexity (usually found in vanilla transformers) is easily tacked by its hierarchical design whilst computing high- resolution images. SWIN Transformer is also ideal for a large and small dataset due to its adaptability as a result of its design. The image is first divided into patches in a hierarchical manner. Then, these patches are merged as the network goes deeper to capture both global and local features. The window-based self- attention and shifted windows concept reduces computation ultimately improving the performance. SWIN Transformers truly have emerged and lived up to the idea of a ground- breaking advancement in the world of computer vision and technology [10]. Its ability to be flexible,

scalable and act as an efficient solution for visual recognition tasks allows it to make way for new breakthroughs in the deep learning and computer vision space. Coupled with its capability to capture long range dependencies, without a doubt, SWIN Transformers are indeed a promising choice for modelling complex visual patterns It would not be surprising at all if, SWIN Transformers, are at the forefront of research and practical implementations in various deepfake detection, segregation and other visual imagery related issues.

Other than object and Deepfake detection, upon researching we have come across applications of SWIN transformer in across a spectrum of domains [11]. To name a few; Remote photoplethysmography for heart rate measurement, transformers in medical image segmentation, brain and vision transformers for autism spectrum disorder diagnosis and classification, air pollution measurement based on a hybrid convolutional neural network with a spatial-and-channel attention mechanism, and Earth Observation. The paper follows a structured approach: we begin with an overview of deepfake technology and the importance of effective detection methods. Next, we delve into existing research on deepfake detection and Transformer architectures in computer vision. We then introduce our model, explaining how we've adapted the SWIN Transformer for deepfake detection. After that, we detail our experimental setup, including datasets, training methods, and evaluation criteria. Following this, we present and analyse our experimental results, discussing their implications and limitations. Finally, we conclude by summarising our key findings and suggesting future research directions. This structured approach aims to make our research methodology, results, and contributions accessible to readers.

2 Literature Review

The margin at top should be set to 3.5 cm, while bottom, left and right margin should be set to 2 cm. The header position from the top should be set at 2.3 cm. The text of the paper should be arranged in sections and when necessary, into subsections. Sections should be numerated with one Arabic numeral, and subsection with two Arabic numerals e.g. 1.1, 1.2, 1.3 etc. The paper's title should be brief and informative, it must also clearly describe the paper's subject matter. The emergence of SWIN Transformers represents a pivotal advancement in bridging the gap between language and vision domains, particularly in the realm of deepfake detection. By employing a hierarchical transformer architecture with shifted windows, SWIN Transformers efficiently compute representations, facilitating multi-scale modelling with linear computational complexity. This transformative capability extends beyond deepfake detection, with applications spanning various domains. In [12], in order to improve computational efficiency, the authors devised a hierarchical Transformer with shifted windows, which limits self-attention to non-overlapping local windows. This facilitates cross-window connections, enabling flexible modelling at different scales with linear computational complexity relative to image size. In [13], the authors present SWINIR, which consists of components for high-quality im- age reconstruction, deep feature extraction, and shallow feature extraction. Multiple residual SWIN Transformer blocks (RSTBs), each with SWIN Transformer layers and a residual link, are integrated by the deep feature extraction module. Tasks including JPEG compression artifact reduction, colour and grayscale image denoising, and different types of image super-resolution—classical, lightweight, and real-world—are all covered by the model.

In [14], the authors explored scaling SWIN Transformer to 3 billion parameters, enabling training with images up to 1,536x1,536 resolution. Innovations include residual post- normalization and scaled cosine attention for model stability. They introduced a log-spaced continuous bias technique to effectively transfer pretrained models from low to higher resolution images and windows. In [15], the authors employed shifted windows with multi-head self-attention (W-MSA/SW- MSA) for texture preservation. The network comprised input modules, feature extraction modules, and output modules, with a novel multi-channel loss integrating sensitivity maps. In [16], the authors introduced DS-TransUNet, a deep medical image segmentation framework that combines a conventional U-Net design with hierarchical SWIN Transformer. By simulating multiscale contexts and non-local dependencies in medical images, it improves the quality of semantic segmentation. In [17], the authors devised a window shift scheme enhancing feature transfer for defect detection, utilizing an improved Vision Transformer. Annotated 4000+ images of metal defects, achieving superior performance in surface-defect detection. Fine-tuned the model via transfer learning for enhanced ac- curacy.

In [18] The authors introduced MoBY, a self-supervised learning method employing Vision Transformers. After 300 epochs of training, it combined MoCo v2 and BYOL to obtain high accuracy on ImageNet-1K linear evaluation: 72.8% and 75.0% top-1 accuracy with DeiT-S and SWIN-T, respectively. In [19] the authors proposed a method incorporating intra- domain fusion using self-attention and inter-domain fusion employing cross-attention to integrate long dependencies within and across domains. This enables full extraction of domain- specific information, cross-domain complementary integration, and maintenance of global intensity perspective. In [20], the authors have proposed a novel semantic segmentation framework for RS images called ST-U-shaped network (UNet), which embeds the SWIN transformer into the classical CNN- based UNet. In [21], in order to recover the low-resolution compressed image, the authors have presented the Hierarchical SWIN Transformer (HST) network, which simultaneously captures the hierarchical feature representations and improves each- scale representation using SWIN transformer. In [22], the authors proposed a cross-modality fusion model, SWINNet, with the purpose of RGB-D and RGB-T salient object detection. It is aided with the SWIN Transformer to extract the hierarchical features, boosted up by an attention mechanism which bridges the gap between two modalities, and guided by edge information to sharp the contour of salient objects. In [23], the authors have investigated key challenges including the use of transformers in different learning paradigms, improving model efficiency, and coupling with other techniques. In [6], the authors presented AVFakeNet, a deepfake detection frame- work integrating audio-visual modalities. Their unified model, Dense SWIN Transformer Net (DST-Net), consists of input, feature extraction, and output blocks. Dense layers compose the input and output blocks, while a customized SWIN Trans- former module is employed in the feature extraction block.

In this work [24], the authors introduced semantically- relevant contrastive learning (SRCL), enhancing SSL, which compares instance relevance to produce more positive pairs. In order to improve universal feature representations for histopathology problems, a hybrid model called CTransPath—which combines a CNN and multi-scale SWIN Transformer—is used to pretrained on unlabelled histopathological pictures. This model functions as a collaborative local- global feature extractor. In [25], the authors enhanced SWIN Transformer with CNN advantages, introducing Local Perception SWIN Transformer (LPSW) to boost local perception for small-scale object detection. They developed SAIEC frame- work to improve segmentation accuracy. Overall, in image pro- cessing, SWIN Transformers demonstrate remarkable efficacy in tasks such as image restoration [13] and resolution scaling [14][21]. Their versatility extends to the field of medical science, where they contribute to faster MRI processing [15], as well as enhancing medical image segmentation and analysis through integration into frameworks like U-Net [16][20][23]. Notably, the SWIN Transformer's segmentation accuracy renders it suitable for applications in salient feature detection [22] and remote sensing object detection [25]. The unique attributes of SWIN Transformers, including the shifting windows and hierarchical structure, enable the effective collection of multi-scale characteristics critical for discerning subtle discrepancies indicative of deepfake manipulation like in AVFakeNet [6]. The diverse applications of SWIN Transformers underscore their versatility and effective- ness across various domains. Their ability to capture intricate details at multiple scales positions them as valuable tools for detecting anomalies indicative of deepfake manipulation. As the threat of deepfake proliferation continues to escalate, leveraging SWIN Transformers offers a promising avenue for enhancing detection capabilities and preserving the integrity of visual media.

3 Proposed Model

This work deals with the efficacy of SWIN Transformers, a sophisticated class of deep-learning models leveraging self-attention mechanisms and advanced feature analysis. By closely scrutinizing images, they adeptly capture both overarching context and intricate details. Specifically, we investigate their effectiveness in discerning AI-generated images, with a particular emphasis on deepfakes. The details architecture of the complete module is represented in Figure.2. In Figure.2 the modules 1(a),1(b) are patch partitioning and 1(c) represents liner embedding, 1(d),1(e) and 1(f) represents SWIN block, SWIN transformer and region merging and the details diagram shown in Figure.3, Figure.4 and Figure.5.

3.1 Architectural Description

The input image first passes through the following blocks:

1. Encoder: The primary aim of encoders in classification is to look for the target region and extract contextual and required characteristics from them.
2. Patch Partitioning: Image originally being of the size 256 × 256, is further divided into patches of 4 × 4 size. This forms a grid of 64 × 64 size. Here it starts with a small patch size and then increases the patch sizes as the layers increase. Each small image portion/ patch is a coloured image with Red, Green, and Blue as its colour channels. The RGB input image is first divided into non-overlapping windows. Each patch is then handled like a token and has its feature set transformed to raw pixel values. The final feature set dimension size increase to 4 × 4 × 3 = 48.

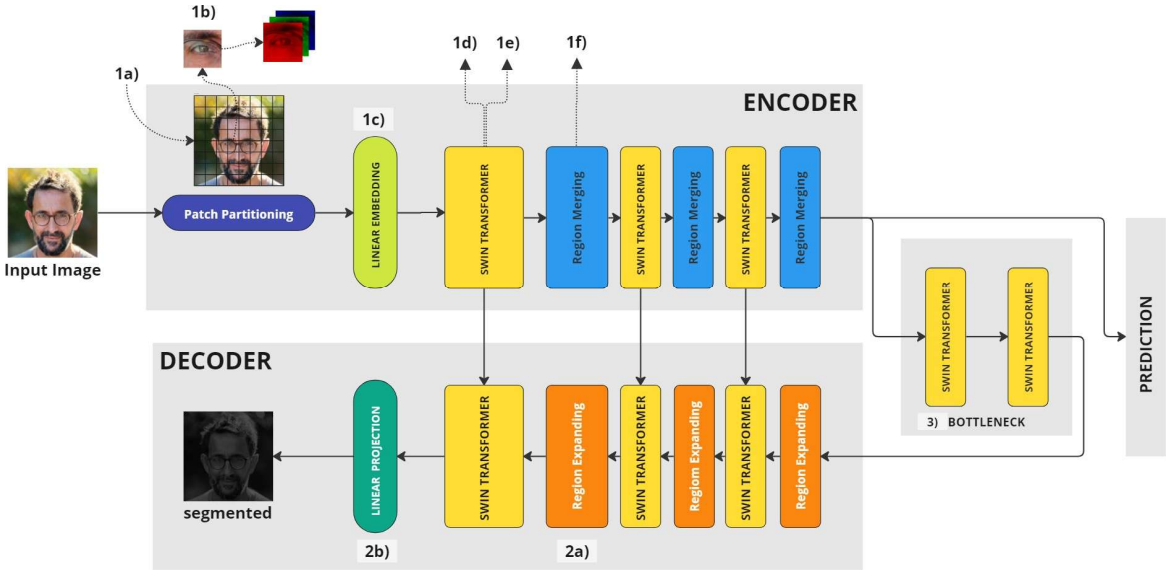

Figure 2. Detailed Architecture to Detect Deepfake Images

3.2 Linear Embedding:

Converts images to a numerical form (sequence of tokens) or AD (Arbitrary Size). As transformer works with a sequence of tokens. This helps converts a patch into a C-dimensional token (dependent on the the model size). Each token from a patch lets us calculate the attention followed by a feature extraction.

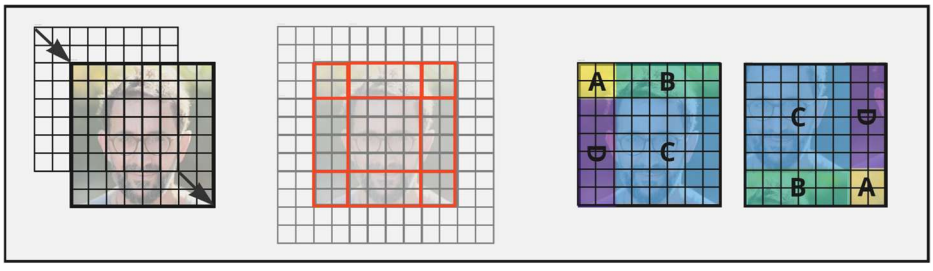

Figure 3. SWIN block in SWIN transformer

3.3 SWIN block in Transformer:

The SWIN block differs from MSA (multi-head self-attention layer) by utilising shifted windows. Both the WMSA (window-based) and SWMSA units are used in the SWIN transformer blocks. The composition of the block is depicted in the schematic diagram. Displacing window by [M/2, M/2] px from the regularly partitioned windows. (here 2 × 2 shift, M=4 patch size) Disadvantage of shifted window partitioning is that this con- figuration has more windows and some windows are smaller in second configuration as compared to the first configuration. SWIN transformer solves this problem using cyclic shifting windows, where the windows on the fringes are padded with each other. In the last portion of the image, A and C are not next to

each other in real life, hence passed through Masked MSA Displacing window by [M/2, M/2] px from the regularly partitioned windows. (here 2 × 2 shift, M=4 patch size). The disadvantage of shifted window partitioning is that this configuration has more windows and some windows are smaller in the second configuration as compared to the first configuration. SWIN transformer solves this problem using cyclic shifting windows, where the windows on the fringes are padded with each other These are not next to each other in real life, hence passed through Masked MSA. The transformer architecture of the SWIN block is shown in Figure.3.

3.4 SWIN Transformer:

Layer Normalization helps in estimating the normalization statistics without introducing any more dependencies between the training set shifted window multi-head self-attention- It takes the O/P of W-MSA shift all windows according to the parameter and compute W-MSA in shifted windows.

- Multi-Layer Perceptron: It is a dense layer which trans- forms any input dimension to the desired dimension.
- W-MSA: It uses dot product-based attention encoding for each product, w.r.t all other patches as input image. The overall architecture of SWIN transformer is shown in Fig.4.

3.5 Region Merging:

The input patches are divided into equal 4 parts combined by this layer. This boosts the feature dimension by 4 times, a linear layer later reduces the feature dimensions back to the original 2. This entire procedure is carried out three times paired with SWIN transformer blocks. SWIN transformer selectively merges adjacent patches to capture the global information properly by merging 4 patches, we keep on increasing the resolution. Fig.5 shows the region merging for boosting feature dimension.

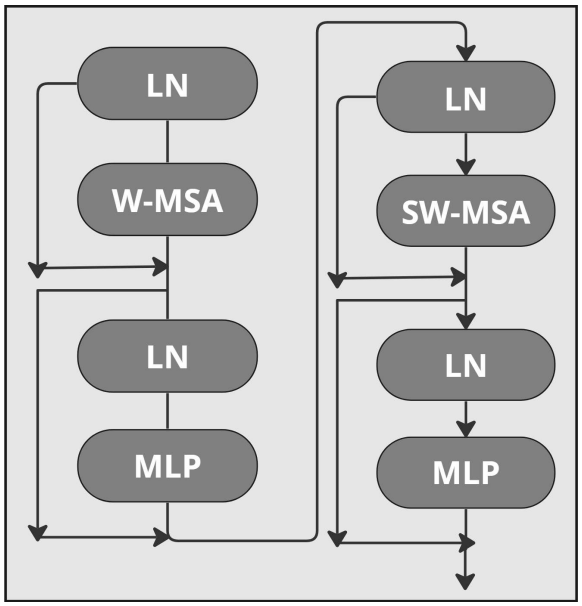

Figure 4. SWIN transformer architecture

3.6 Decoder:

Region Expansion: As part of the decoding process of the SWIN Transformer, the image is upsampled using features from the SWIN block to improve the observation of finer details.

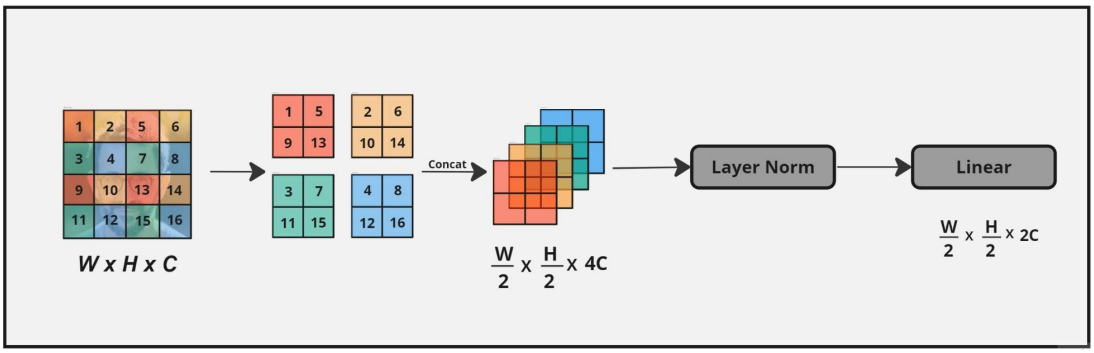

Figure 5. Region Merging for boosting feature dimension

3.7 Bottleneck Prediction:

The bottleneck block uses two successive SWIN Transformer units to overcome the difficulty of learning with deeper layers. By strategically balancing dimensionality and feature resolution, this method maximises the model's learning and representational capabilities. Together, these essential elements give the SWIN Transformer the ability to absorb and interpret visual data well for a variety of tasks. An additional key feature of the SWIN Transformer is input padding, where the model ensures the handling of images of varying dimensions, of any height or width if it's a multiple of 32. This feature increments the flexibility of the overall pre- processing. The hierarchical partitioning functionality allows the capture of both local and global features as the network deepens (layers increase), by merging the smaller patches into larger ones. The larger image and patch detect the global and local features of the image respectively. The complete block diagram of this is shown in Figure 6. This phase consists of 2 subsections:

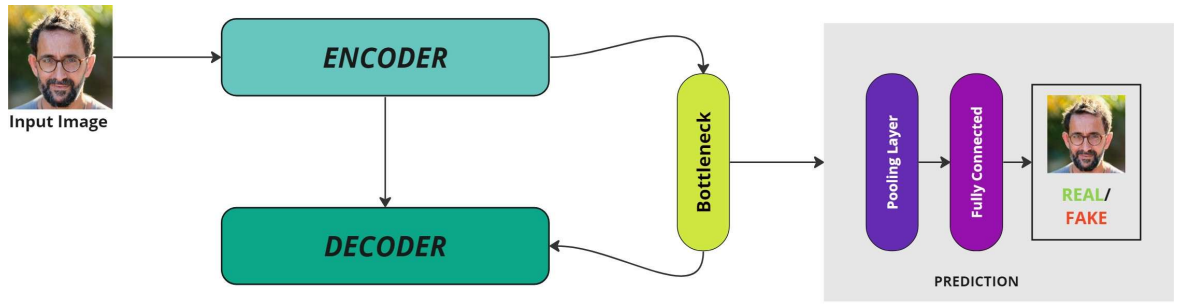

Figure 6. Complete Block Diagram

- Pooling layer: Here we witness the following procedures taking place. Dimensionality reduction - usually used to control overfitting in a dataset and decreasing the number of parameters. Here we witness the following procedures taking place. Dimensionality reduction, usually used to control overfitting in a dataset and decreasing the number of parameters. Feature Extraction aids in keeping the most relevant features and discarding the rest. Spatial Hierarchy, enables the network to go deeper and capture an increasing resolution of global and abstract features.
- Fully Connected Layer the FC layer is the final stage of this model, responsible for converting the extracted features into a format that can be easily used to make predictions and draw conclusions. It consists of one or more fully connected layers of neurons, where the number of neurons depends on the size of the input dataset and the complexity of the task. As we approach the output layer, the number of neurons gradually decreases. Since we are performing binary classification (Deep/Fake), we will use a single output neuron with the most appropriate activation function.

4 Result and Discussion

The proposed model has been assessed on Celeb-df and FaceForensics++ datasets on the basis of accuracy and AUC. Additionally, the following preprocessing could be potentially useful for our dataset to ensure that the data is suitable for training a machine-learning model and can lead to improved model performance. Determining whether a dataset of photos needs pre-processing depends on the nature of the dataset, the specific task you're aiming to perform, and the characteristics of the images. Here are some common reasons why you might consider pre-processing a dataset of photos:

- Image Quality: Check for variations in image quality, such as lighting conditions, resolution, or noise. Pre- processing may involve standardizing image quality to ensure consistency.
- Normalization: Normalize pixel values to a common scale. This is important if the images have varying
- intensity levels, ensuring the model receives consistent input.
- Noise Removal: Remove noise or artifacts from images that might interfere with model training or affect the quality of predictions.
- Data Augmentation: Apply data augmentation techniques to artificially increase the diversity of the dataset. This can involve random rotations, flips, or adjustments to brightness and contrast.
- Labelling Consistency: Ensure labelling consistency within the dataset. If labels are inaccurate or inconsistent, it can affect the model's performance.
- Outlier Detection: Identify and handle outliers, which may be images that don't conform to the typical characteristics of the dataset.
- Data Balancing: Check if the dataset is imbalanced (some classes have significantly fewer samples than others) and consider strategies like oversampling or under sampling to address this imbalance.
- Missing or Corrupt Data: Identify and handle missing or corrupt images in the dataset. We have also compared it with traditions CNN models and presented the data in the Table 1, 2, 3 and 4 below. The same also can be visualised on Celeb-df and FaceForensics++ dataset in Figure.7, Figure.8, Figure.9.

Table 1. Xception

	Parameters		
Dataset	Accuracy	AUC	Log Loss
CalebDF	97	0.99	0.0712
FaceForensic++	91.05	0.96	0.2342

Table 2. Restnet3D

	Parameters		
Dataset	Accuracy	AUC	Log Loss
CalebDF	97	0.99	0.0748
FaceForensic++	90.36	0.96	0.3224

Table 3. Res2Net-101

	Parameters		
Dataset	Accuracy	AUC	Log Loss
CalebDF	98.95	1	0.0237
FaceForensic++	93.48	0.97	0.2165

Table 4. SWIN-T

Parameters			
Dataset	Accuracy	AUC	Log Loss
CalebDF	97.91	0.99	0.034
FaceForensic++	95.715	0.9625	0.1573

The graphs denote the comparison of the various CNN models and SWIN transformer on the two datasets, Celeb-df and FF++. It can be observed that SWINT gives much better accuracy on the FaceForensics++ dataset which is a more complex dataset in comparison to Celeb-DF overshadowing its falling behind with Res2Net-101 in the Celeb-DF dataset since Celeb-DF is a simpler dataset. Hence, we can conclude that SWINT performs much better considering the complexity of the datasets. The project has achieved partial fulfilment, yielding several outcomes.

A partial solution for deepfake detection has been implemented, showcasing effectiveness in identifying manipulated content to some extent, though improvements in coverage and accuracy are needed. Valuable insights gleaned from the project have informed future research directions and enhancements in deepfake detection methodologies.

Additionally, a prototype or proof-of-concept implementation has been developed, demonstrating fundamental functionality and laying the groundwork for further refinement. The project has also identified limitations and gaps in the proposed methodology, such as scalability issues and technical challenges, providing crucial insights for future iterations. Furthermore, the partial completion of the project has set the stage for future collaborations and endeavours. Researchers can build upon the existing framework, leveraging insights gained and addressing remaining challenges to advance the field of deepfake detection. Although the project's partial fulfilment does not constitute a fully operational deepfake detection system, it has nonetheless contributed valuable knowledge and paved the way for ongoing advancements in combating synthetic media manipulation.

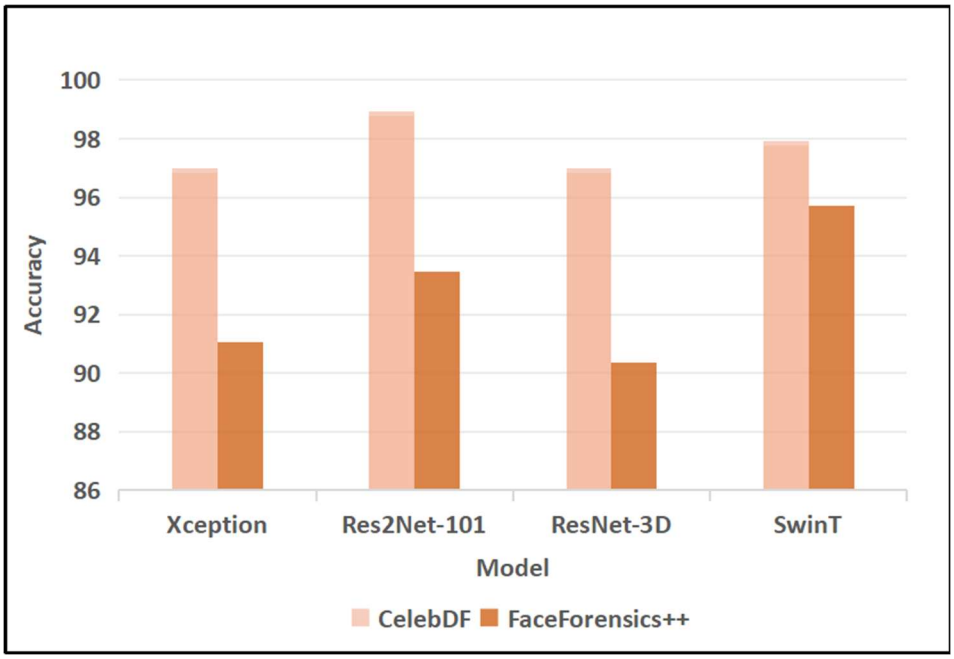

Figure 7. Model vs Accuracy

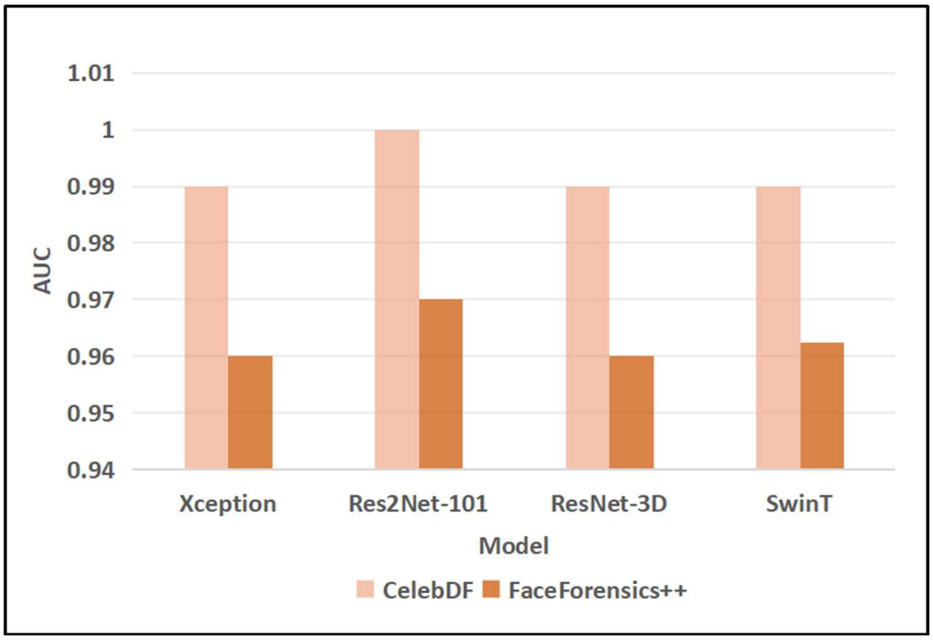

Figure 8. Model vs AUC

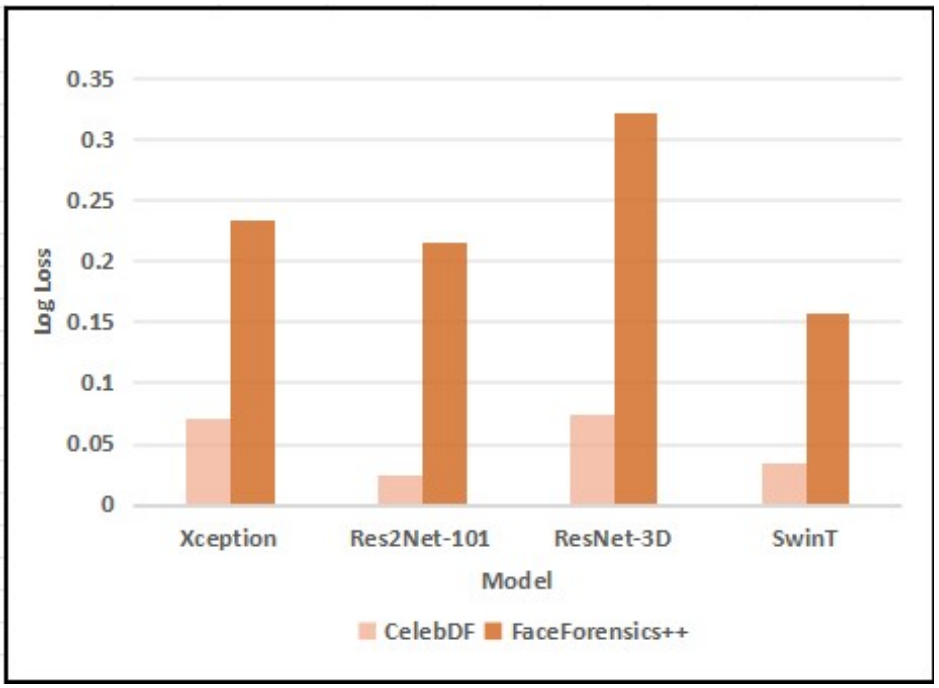

Figure 9. Model vs Log Loss

5 Conclusion

In our research, we introduced a modified SWIN Trans- former architecture tailored for the classification of deepfake images. To assess the effectiveness of our approach, we conducted evaluations using the Celeb-df and FF++ datasets, which are widely used benchmarks in the field of deepfake detection. Our results indicate that our modified SWIN Trans- former architecture exhibits promising capabilities in identifying deepfake images. Specifically, we observed that our model outperformed traditional CNN models in terms of classification accuracy and overall performance. The hierarchical structure and attention mechanisms inherent in SWIN Transformers enable better capture of spatial and contextual information, leading to more robust classification outcomes. Overall, our study underscores the potential of SWIN Transformer-based architectures

for deepfake detection tasks. The enhanced performance that we observed in our assessments underscores the effectiveness of our suggested methodology and its potential to foster progress in countering the spread of deepfake media.

Declaration

Conflict of Interest: The authors declare that they have no conflict of interest.
Funding Information: No funds, grants, or other support was received.
Author contribution: All authors contributed to the study conception and design.
Data Availability Statement: Data can be produced upon request.
Research Involving Human and /or Animals: Not Applicable. Informed Consent: Not Applicable.

References

[1] D. Pan, L. Sun, R. Wang, X. Zhang, and R. O. Sinnott, "Deepfake detection through deep learning," in 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), pp. 134–143, IEEE, 2020.

[2] A. Kohli and A. Gupta, "Detecting deepfake, faceswap and face2face facial forgeries using frequency cnn," Multimedia Tools and Applications, vol. 80, no. 12, pp. 18461–18478, 2021.

[3] D. Gu¨era and E. J. Delp, "Deepfake video detection using recurrent neural networks," in 2018 15th IEEE international conference on advanced video and signal-based surveillance (AVSS), pp. 1–6, IEEE, 2018.

[4] S. Tariq, S. Lee, and S. S. Woo, "A convolutional lstm based residual network for deepfake video detection," arXiv preprint arXiv:2009.07480, 2020.

[5] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," IEEE access, vol. 10, pp. 25494–25513, 2022.

[6] H. Ilyas, A. Javed, and K. M. Malik, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection," Applied Soft Computing, vol. 136, p. 110124, 2023.

[7] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "Deepfake detection algorithm based on improved vision transformer," Applied Intelligence, vol. 53, no. 7, pp. 7512–7527, 2023.

[8] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," arXiv preprint arXiv:2106.10270, 2021.

[9] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficient net and vision transformers for video deepfake detection," in International conference on image analysis and processing, pp. 219–229, Springer, 2022.

[10] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2tr: Multi-modal multi-scale transformers for deepfake detection," in Proceedings of the 2022 international conference on multimedia retrieval, pp. 615–623, 2022.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022, 2021.

[12] L. Gao, H. Liu, M. Yang, L. Chen, Y. Wan, Z. Xiao, and Y. Qian, "Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," IEEE journal of selected topics in applied earth observations and remote sensing, vol. 14, pp. 10990–11003, 2021.

[13] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin trans- former," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 1833–1844, 2021.

[14] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., "Swin transformer v2: Scaling up capacity and resolution," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12009–12019, 2022.

[15] J. Huang, Y. Fang, Y. Wu, H. Wu, Z. Gao, Y. Li,J. Del Ser, J. Xia, and G. Yang, "Swin transformer for fast mri," Neurocomputing, vol. 493, pp. 281–304, 2022.

[16] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "Ds-transunet: Dual swin transformer u-net for medical image segmentation," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–15, 2022.
[17] L. Gao, J. Zhang, C. Yang, and Y. Zhou, "Cas-vswin transformer: A variant swin transformer for surface- defect detection," Computers in Industry, vol. 140, p. 103689, 2022.
[18] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, "Self-supervised learning with swin transformers," arXiv preprint arXiv:2105.04553, 2021.
[19] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 7, pp. 1200–1217, 2022.
[20] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–15, 2022.
[21] B. Li, X. Li, Y. Lu, S. Liu, R. Feng, and Z. Chen, "Hst: Hierarchical swin transformer for compressed image super-resolution," in European conference on computer vision, pp. 651–668, Springer, 2022.
[22] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 7, pp. 4486– 4497, 2021.
[23] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao,Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," Intelligent Medicine, vol. 3, no. 1, pp. 59–78, 2023.
[24] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang,W. Yang, J. Huang, and X. Han, "Transformer-based unsupervised contrastive learning for histopathological image classification," Medical image analysis, vol. 81, p. 102559, 2022.
[25] X. Xu, Z. Feng, C. Cao, M. Li, J. Wu, Z. Wu, Y. Shang, and S. Ye, "An improved swin transformer-based model for remote sensing object detection and instance segmentation," Remote Sensing, vol. 13, no. 23, p. 4779, 2021.
[26] Bhowmick, R., Mishra, S.R., Tiwary, S. et al. Machine learning for brain-stroke prediction: comparative analysis and evaluation. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-20057-6.