Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 4, Issue 3, 2026

DEEPFAKESHIELD: ENHANCED VIDEO AUTHENTICITY DETECTION VIA CONVOLUTIONAL VISION TRANSFORMER

Izhar[*1], Dr. Naeem Aslam[2], Muhammad Sajid Maqbool[*3], Muqadas Nadeem[4], Hira Saleem[5]

[*1,2,*3,5]*Department of Computer Science, NFC-IET, Multan, Pakistan*
[4]*Department of Computer Science, Emerson University, Multan, Pakistan*

[*1]izharmcs1@gmail.com, [*3]sajid.maqbool@nfciet.edu.pk

DOI: https://doi.org/10.5281/zenodo.19385256

Keywords
Deepfake Detection, EfficientNet, Swin Transformer, Attention Mechanism, Convolutional Feature Extraction, Synthetic Media

Article History
Received: 31 January 2026
Accepted: 15 March 2026
Published: 31 March 2026

Corresponding Author: *
Izhar,
Muhammad Sajid Maqbool

Abstract

The way the deep learning algorithms have quickly evolved to be able to produce and recreate the extremely realistic videos, also known as Deepfakes, has caused considerable alarm about the misuse of such tools. Innovative techniques of deep learning are now capable of producing synthetic faces, swapping faces across people, changing facial expressions, modifying gender traits and manipulating facial features with incredible accuracy. Although virtual reality, digital content creation, and entertainment are lawful uses of these technologies, they have serious risks when they are used in bad intentions like misinformation, stealing of identities, and internet fraud. This study presents an effort to introduce a hybrid model combining EfficientNet with the Swin Transformer to detect Deepfakes effectively. EfficientNet is used to extract fine-grained spatial features, whereas the Swin Transformer uses hierarchical attention to capture long-range dependencies to classify authenticity. The proposed framework was also trained and tested on FaceForensics data, with the accuracy of 93.2, AUC of 0.94, and the loss rate of 0.28. A combination of EfficientNet convolutional representations and the Swin Transformer attention-based system proves to be at a better level of detecting manipulated content, and this points to the model being able to distinguish well between the manipulated and veritable videos.

1. INTRODUCTION

Technologies for modifying photos, movies, and audio are advancing swiftly [1]. Methods and technical proficiency for generating and modifying digital information are readily available. At present, one may effortlessly produce hyper-realistic digital photographs with minimal resources and readily accessible instructions online. Deepfake is a method that seeks to substitute the visage of a designated individual with that of another in a video. It is generated by

integrating a synthetic facial area into the original picture [2,4]. The phrase may also denote the ultimate result of a hyper-realistic video produced. Deepfakes can facilitate the production of hyper-realistic Computer Generated Imagery (CGI), Virtual Reality (VR), Augmented Reality (AR), education, animation, arts, and cinema.

Nonetheless, because to their inherently misleading nature, Deepfakes can be employed for nefarious ends. Following the emergence of the Deepfake phenomena, several authors have suggested various methods to distinguish authentic videos from fabricated ones. As indicated by [10], despite the strengths of each hypothesized mechanism, existing detection approaches exhibit a lack of generalizability. The authors observed that previous models mostly concentrate on the tools used for Deepfake development by analyzing their purported behaviors. According to an annual report [6] on Deepfake, deep learning researchers achieved many significant advancements in generative modeling. Computer vision researchers introduced a technique called Face2Face [7] for face re-enactment. This technique conveys facial expressions from an individual to a computer avatar in real-time. In 2017, researchers from UC Berkeley introduced CycleGAN to convert photos and movies into various genres. A separate cohort of researchers from the University of Washington suggested a technique to align lip movements in video with audio from an alternative source [9]. In November 2017, the term "Deepfake" was used to describe pornographic videos in which the faces of celebrities were superimposed onto actual footage.

Nonetheless, training a multi-attentional network is a significant challenge. This is primarily due to the fact that, in contrast to single-attentional networks [10], which may utilize video-level labels as explicit guidance and be trained in a supervised manner, the multi-attentional structure can only be taught in an unsupervised or weakly-supervised manner. Utilizing a conventional learning technique results in the degradation of multi-attention heads to a singular attention equivalent, wherein just one attention area elicits a robust response, while the other attention regions are repressed and fail to acquire valuable information [11]. To resolve this issue, we furthermore suggest a novel attention-guided data augmentation approach. During training, we will intentionally obscure some high-response attention areas (soft attention lowering) to compel the network to learn from alternative attention regions. The authors characterized generality as the consistent identification of various spoofing methods and the dependable execution of previously undetected spoofing strategies.

The studies proposed a generalized DeepFake detector (FakeCatcher), which is based on biological cues and internal image features to detect manipulated content. The model used a basic Convolutional Neural Network (CNN) consisting of three layers and trained and tested it on a dataset of around 3,000 videos. The study, however, was not clear on the preprocessing activities that were done on the data. As it has been mentioned in the previous research [31, 52, 21], more complex CNN-based models tend to outperform shallow networks in the image classification task. This fact gives reason to believe that a more powerful DeepFake detector could be developed, with a detailed data preparation pipeline and a more complex neural network architecture to detect any artifacts of manipulation and their consequences.

In these regards, we introduce a generalized Convolutional Vision Transformer (CViT) system of Deepfake video detection. The suggested CViT architecture will integrate the advantages of both Convolutional Neural Networks and Transformer-based attention to local and fully-global visual dependencies in video frames. There are three primary reasons why we believe our approach is generalized: (1) it combines CNN and Transformer modules and is capable of extracting and fusing spatial and contextual information; (2) it focuses on the significance of data preprocessing and augmentation in the process of training and classification; and (3) the strategy is trained using a large and various facial dataset that includes several conditions, environments, and viewing angles. The suggested CViT architecture is expected to provide a very flexible and scaled Deepfake detector that will be able to detect fake material created with a variety of different synthesis methods.

2. Related Work

The fast progression of Convolutional Neural Networks [4, 20], Generative Adversarial Networks (GANs) [18], and its derivatives [22] has enabled the generation of hyper-realistic pictures [32], movies [61], and audio signals [53, 15] that are increasingly difficult to identify and differentiate from authentic, unaltered audiovisuals. The capacity to generate a convincingly authentic sound, Images and videos have prompted numerous concerned parties to discourage innovations that may be used by enemies for nefarious reasons [12]. Consequently, there is a pressing demand within the research community to develop Deepfake detection techniques. The blistering development of the Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), [13] and other models has transformed the content creation process as it is possible to create hyper-realistic images, videos, and audio signals that are not always distinguishable with real ones. Although these innovations can be creatively used in various ways, it is also increasing fears of evil misuse, such as misinformation, identity theft, [14] and internet fraud. This has made the research fraternity pay attention towards creating powerful Deepfake detection systems that can detect a synthetic content produced by using advanced deep learning models.

Deepfakes are created in the form of deep generative models, namely GANs and Autoencoders (AEs) that either modify or swap facial identities on pictures and videos [15]. The most common methods of generating Deepfakes include face swapping, lipsync, face reenactment and speech synthesis. Sooner applications, like FakeApp, used dual autoencoders sharing encoders to share face features among people. Future architectures, such as StyleGAN and FSGAN, have greatly improved face realism, whereas other networks, such as CycleGAN and Face2Face, can transfer poses and expressions without paired data, allowing manipulation of faces in real time and in photorealism [16, 17]. Deepfake detection methods have also developed together with generative models and can be broadly divided into three categories: (1) methods which look at physiological and behavioral indicators (e.g., eye blinking, head motions), (2) methods looking at GAN fingerprints or biological ones like inconsistent blood flow, and (3) data-driven approaches that detect visual artifacts in the manipulated media. Models like MesoNet and MesoInception-4 are used to identify mesoscopic inconsistencies whereas other models use the

geometric and affine transformation artifacts to identify the manipulated regions. More progressive models unite time and space characteristics with hybrid CNN-RNN models to analyze video better [18,21].

Ensemble and hybrid learning strategies have also been examined in recent literature to be more robust. Indicatively, the DeepfakeStack combines several pretrained models, such as XceptionNet, InceptionV3, ResNet101 and DenseNet [22] variants to increase classification accuracy. Correspondingly, comparative analyses based on ShallowNet, VGG-16 and Xception show that more complicated architectures have an advantage over shallow networks when it comes to detecting manipulation artifacts.

Regardless of the major achievements, DeepFake detection is still considered a challenge because of the constant development of generative models. As indicated in the reviewed literature, there is an increasing demand to integrate convolutional and transformer models in a hybrid architecture as the desire to make these components generalized, data-efficient, and attention-based, which should be able to adequately represent both local and global image characteristics to provide reliable Deepfake detection [23].

3. Materials and Methods

This part explains how Deepfake video detection works. The suggested detection framework has two main parts: the preprocessing part and the detection part. The preprocessing part gets the input data ready and has two main steps: face extraction and data augmentation. In this step, facial regions are accurately taken out of video frames and improved using different augmentation methods to make the model more robust and generalize better. There are three parts to the detection component: training, validation, and testing. The Convolutional Vision Transformer (CViT) is used in both the training and validation stages. It has two powerful modules: a Feature Learning (FL) module that extracts detailed facial features from the input images, and a Vision Transformer (ViT) module that analyzes these features to determine if each video is real. Finally, the trained CViT model is used to correctly tell the difference between real and deepfake images during the testing stage. Figure 1 shows an overview of the suggested framework for detecting Deepfakes.

3.1. Data Preprocessing

The step of preprocessing is essential in framing the raw data to the stages of training, validation, and testing of the proposed CViT model. This step is used to make sure that the input data is clean, constant and well-formed so that the model can learn effectively and provide the correct output. The preprocessing pipeline is comprised of two key modules face extraction and data augmentation. Face extraction module will identify facial regions of video frames and transform them into a common 256 x 256 RGB representation. This is done to make sure that images are the same size and color representation are consistent, making model performance a stable aspect. The data augmentation module, meanwhile, adds some variations to the data, i.e. rotation, flipping, brightness adjustment, and slight scaling. These changes make the model stronger and closer to a variety of light, poses, and facial expressions. In Fig. 2 and Fig. 3, a few examples of the extracted and preprocessed facial images are shown and thereby utilized in the further training phase.faces.

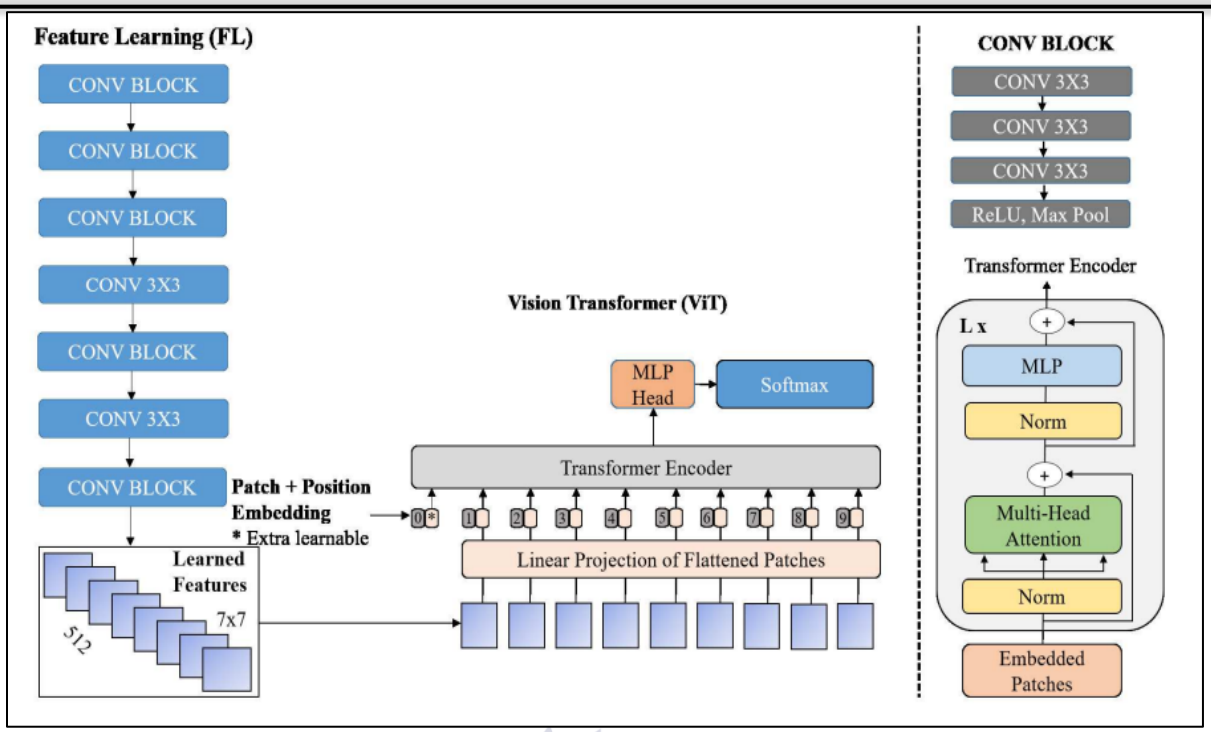

Figure 1 Convolutional vision Transformer

3.2. Identification

The Deepfake detection can be divided into three major steps, including training, validation, and testing.The training step is main aspect of proposed CViT strategy - this is the step where the real learning is carried out. Deep learning models typically take long periods and fine-tuning to be suitable in a given domain of a problem. In our scenario, the main objective is to enhance an optimized CViT architecture that has the ability to learn and discriminate the distinctive features of deepfake videos. The process of training varies parameters and hyper parameters to get the highest performance on our data set. The validation phase runs in parallel with the training phase, and it is important to fine-tune the model with this stage. It assists in measuring the performance of the CViT on unseen data, to improve the performance and avoid overfitting. The final step is the testing phase, which will make sure that the model can generalize significantly beyond the training sample and give a sound measurement of the progress in terms of accuracy and loss.The testing stage is the last phase, during which the fully trained CViT model would be used to classify and identify whether the faces extracted out of a given video are genuine or fake. This step indirectly indicates the feasibility of the offered solution in practice and produces the primary goals of the research. The suggested CViT model is based on two substantial blocks: The Feature Learning (FL) and the Vision Transformer (ViT). The FL module does the extraction of meaningful facial features in the image and the ViT module does the extraction features, transforms it to a set of image patches and does the actual classification to determine authenticity. The Feature Learning (FL) block is based on the VGG architecture but has one key difference, it lacks fully connected (dense) layers that are utilized in the first VGG model to perform classification. Rather, the FL pays attention to only convolutional operations to obtain and refine facial features representations, which in turn are sent to the ViT to undergo the final detection step. Simply put, the FL is a pure convolutional neural network (CNN)

but focused on feature extraction, but not on direct classification.

The FL component has 17 convolutional layers, each utilizing a 3 x 3 kernel. The convolutional layers extract the fundamental features of facial pictures. All convolutional layers possess a stride and padding of one. Batch normalization is utilized to standardize the output features, while the ReLU activation function is employed to introduce non-linearity across all layers. The batch normalization function standardizes variations in the distribution of preceding layers [41], as alterations between layers influence the learning process of the CNN architecture. A five max-pooling operation with a 2 x 2 pixel window with a stride of 2 is also employed. The max-pooling technique lowers the picture dimensions by fifty percent. Following each max-pooling operation, the breadth of the convolutional layer (channel) is increased by a factor of 2, starting with 32 channels in the first layer and culminating in 512 channels in the final layer.

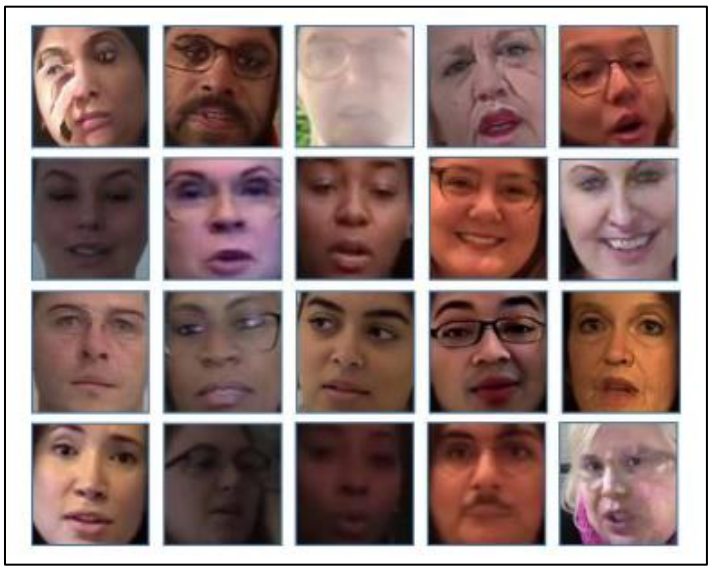

Figure 2 Fake Face Images Samples

The The Feature Learning (FL) module is an architecture of a series of convolutional layers that are intended to produce finest spatial and texture features of the facial images. Each layer has four successive convolutional operations with the last two layers having five more convolutional operations to obtain more abstract and deeper features. To make the explanation simple, the four-convolution architecture is known as the CONV Block. The FL component has about 12.5 million trainable parameters, which is a robust feature extraction feature without overfitting. The FL module takes input images that are 256 256 3 in size and carries out the convolution on the feature maps of each layer. Internal representation FL may be explained the feature map.

Finally, FL produces a 640 x 8 x 8 feature tensor representing the spatial information about the low-level features. This feature map is then passed to a high-level semantic understanding and classification module (ViT) module. ViT part is a design based on transformer that is based on recent developments in visual attention structures, making it robust and scalable to the deepfake detection task. [57].

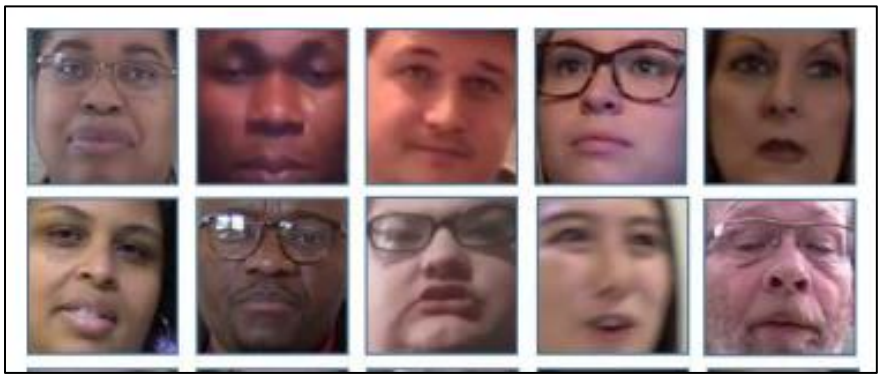

Figure 3 Real Face Images Samples

The transformer and its variations, such as GPT-3 [44], are mostly employed for natural language processing jobs. ViT expands the utilization of the transformer from the area of natural language processing to the domain of computer vision. The ViT employs identical components to the original transformer model, with minor alterations to the input signal. The FL component and the ViT component constitute our Convolutional Vision Transformer (CViT) model. We designated our model as CViT due to its foundation in both a series of convolutional operations and the ViT architecture.

ViT is unit that processes the feature maps that are extracted of the facial images. These patches of feature maps are divided into 16 patches that are flattened and projected to a 1x 768-dimensional linear sequence. These patch embeddings are then fed into the Transformer architecture with positional embeddings of 1 × 768 dimensions added to them in order to retain spatial information.

In contrast to the original Transformer, that has both an encoder and a decoder subsystems, ViT architecture implemented here has an encoder only. The MLP module is used as a Feedforward Neural Network (FFN), and Layer Normalization (Norm) is used to provide stability and consistency in the internal activations. Transformer uses twelve attention heads which means that the model is able to focus on various aspects of the picture features at the same time.

The output stage has the MLP head which comprises two fully connected layers which are triggered by the ReLU function. The initial layer has 1024 channels, Final layer contains two neurons and this neuron is the Real and Fake classes. The entire model of CViT includes 24 weighted layers and about 42 million trainable parameters, which allows it to learn finer details of high accuracy classification. Lastly, the output of the MLP head is normalised by a Softmax activation function to give the results of the classification as a range between 0 and 1, giving the result an interpretation of a probability.

4. Results and Discussion

This section delineates the tools and experimental configuration employed in the design and development of the prototype for model implementation. We will show the results obtained from the model's implementation and provide an interpretation of the experimental findings.

4.1. Data Set

Deep learning models are trained on patterns and features directly out of data; hence, it is essential to provide careful attention to the dataset preparation to enhance the learning efficiency and prediction accuracy of the deep learning model. Facial regions are detected with the help of the RetinaFace, Dlib, and the OpenCV deep neural network (DNN) module in this research based on

highly sophisticated deep learning detectors. RetinaFace and Dlib are more accurate in face localization and face landmark detection whereas OpenCV DNN is a fast and efficient algorithm in large image datasets. The combination guarantees that the extracted samples of the faces are clean, aligned and consistent, which forms very strong basis in strong model training.

Table 1 Accuracy on different Datasets

Dataset	Accuracy
Face Forensics++ FaceSwap	89%
Face Forensics++ DeepFakeDetection	92%
Face & Deepfake	95%
Face & FaceShifter	76%
FaceForensics& NeuralTextures	92.12%

The three deep learning libraries are utilized together to enhance the precision of face detection. The facial pictures are saved in JPEG format with a resolution of 224 by 224 pixels. A compression ratio of 90 percent is also implemented. We organized our datasets into training, validation, and testing sets. We utilized 162,174 photos, allocated as follows: 105,413 for training, 32,434 for validation, and 24,326 for testing, according to a 65:20:15 ratio. Both genuine and counterfeit classes contain an equal amount of photos across all datasets. We employed Albumentations for data augmentation. Albumentations is a Python package for data augmentation that encompasses a wide array of picture manipulations. Ninety percent of the facial photos were enhanced, resulting in a total dataset of 308,130 facial images.

4.2. Assessment

The binary cross-entropy loss function is used to train the CViT model. The 64 images that are in each training batch are standardized with mean [0.5, 0.5, 0.5] and standard deviation [0.25, 0.25, 0.25]. The normalized facial images are further magnified before being introduced in the CViT model in every training cycle in order to enhance features representation and the overall precision of the model.

Table 2: Comparison of accuracy on CNN,RNN and CViT

Models	Validation	Test
CNN & RNN	92.66%	90.03%
CViT	90.32%	93.2%

The This table is a comparison of accuracy of the CViT model and another hybrid deep learning model (CNN RNN GRU) using DeepFake Detection Challenge (DFDC) dataset. When doing the validation, CNN RNN GRU model is a little better as compared to the CViT. Nevertheless, the CViT obtains a similar or a little higher level of accuracy in the test phase, which proves that the CViT has a high level of real-world generalization. The outcome is that the convolutional and transformer layers can be used jointly to achieve competitive results in Deepfake detection.. The AUC represents the area encompassed by ROC curve. The AUC quantifies precision of a classifier.

Figure 4 Detection of ROI Non-Face

We show our findings utilizing accuracy, AUC score, and loss value. We evaluated the model with 398 previously unexamined DFDC films, attaining an accuracy of 91.5 percent, an AUC value of 0.91, and a loss value of 0.32. The loss value signifies the deviation of our model's forecast from the actual target value.

We utilized 29 facial photos from each video for Deepfake detection. The quantity of frame numbers utilized influences the probability of Deepfake detection. Nevertheless, accuracy may not always serve as the appropriate metric for identifying Deepfakes, as genuine face pictures may be included inside a fabricated movie (fake films might include authentic frames).

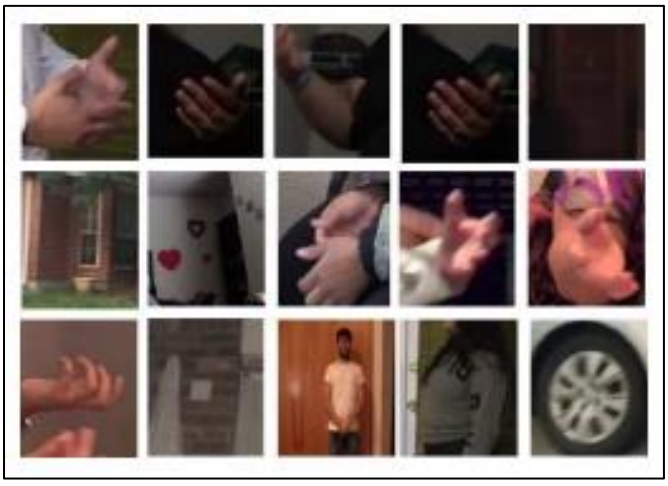

Figure 5 ROI BLAZEFACE Other than Face images

This table will compare the performance of CViT in terms of Area Under the Curve (AUC) performance on the UADFV dataset, to the MesoNet and MesoInception models. The CViT model has high validation AUC (93.75%), which indicates that it is successful in separating real and fake videos on the validation set. But in the case of FaceSwap and Face2Face datasets, its AUC scores (approximately 70) are smaller, indicating that it does not adapt well to a particular Deepfake generation style. Nevertheless, the higher validation performance of the models proves the fact that CViT extracts strong general characteristics of the data.

Table 3 AUC Performance of CViT and Other Models on UADFV Dataset

Method	Validation	FSwap	F2Face
MesoNet	84.37%	96.32%	92.96%
MesoInception	82.44%	98.20%	93.33%
CViT	93.75%	70.10%	70.32%

Figure 6 Detection of ROI on non-Face

4.3. Impacts of Data Processing in Classification

A significant potential issue impacting our model's accuracy is the intrinsic flaws present in the face detection deep learning libraries (MTCNN, BlazeFace, and face-recognition). Figures 4, 5, and 6 depict photos that were inaccurately identified by the deep learning libraries. The figures encapsulate our initial data preparation assessment conducted on 200 films randomly chosen from 10 directories. We selected our test set video encompassing all scenarios available in the DFDC dataset: indoor, outdoor, dimly lit room, brightly lit room, seated subject, standing subject, and a subject positioned at a distance from the camera This table compares three face detecting libraries BlazeFace, MTCNN, and Face Recognition with respect to its DeepFake detection accuracy in various datasets.Face Recognition is the final pre-processing pipeline that the author uses as it has better accuracy and fewer false positives.

Table 4: Comparison of Deep Learning Libraries for Face Detection

Dataset	BlazeFace	Face Recognition	MTCNN
DFDC	83.40%	90%	91.5%
FaceSwap	76%	71%	63%
FaceShifter	52%	48%	44%
NeuralTextures	60%	61%	60%
DeepFakeDetection	79%	92%	80%

5. Conclusion and Future work

The opportunities of deepfake technologies are apparent in a wide range of areas, such as digital media, virtual reality, robotics, and education. But their abuse is potentially a significant ethical and social danger that threatens authenticity and social trust. We suggest the generalized Deepfake detection architecture, the Convolutional Vision Transformer (CViT), as a combination of the advantages of Convolutional Neural Networks (CNNs) and Transformers to analyze the video. There are three main reasons why the model is referred to as generalized: (1) both CNNs and Transformers are learned synergistically, allowing to extract local and global visual representations simultaneously; (2) no special attention to the efficient data preparation is paid during both training and classification; and (3) it is trained on a large and diverse dataset to be able to adapt to a variety of manipulation techniques and scenarios. The proposed CViT architecture was optimized on a combined set of facial images based on the DeepFake Detection Challenge (DFDC) dataset and tested on 400 test videos with a 93.2 per cent accuracy. The outcomes prove that the model is efficient in identifying original and faked video material. The future work will be dedicated to increasing the dataset and incorporating other publicly accessible sources of Deepfake in the

dataset, thus improving the generalization, accuracy, and strength of the model in the real world.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A Compact Facial Video Forgery Detection Network. Pages 1–7, 2018.

Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting World Leaders Against Deep Fakes. In CVPR Workshops, 2019.

Charu C. Aggarwal. Neural Networks and Deep Learning: A Textbook. Springer International Publishing, Switzerland, 2020.

Md Zahangir Alom, Tarek M. Taha, Chris Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Mahmudul Hasan, Brian C. Van Essen, Abdul A. S. Awwal, and Vijayan K. Asari. A State-of-the-Art Survey on Deep Learning Theory and Architectures. Electronics, 8(3):292, 2019.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. arXiv preprint arXiv:1907.05047v2, 2019.

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention Augmented Convolutional Networks. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 3285–3294, 2019.

Avishek Joey Bose and Parham Aarabi. Virtual Fakes: DeepFakes for Virtual Reality. In 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), pages 1–1. IEEE, 2019.

Andrew P. Bradley. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition, 30(7):1145–1159, 1997.

John Brandon. Terrifying High-Tech Porn: Creepy 'Deepfake' Videos Are on the Rise, 2018. Available at https://www.foxnews.com/tech/terrifying-high-tech-porn-creepy-deepfake-videos-are-on-the-rise.

Joshua Brockschmidt, Jiacheng Shang, and Jie Wu. On the Generality of Facial Forgery Detection. In 2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems Workshops (MASSW), pages 43–47. IEEE, 2019.

Polychronis Charitidis, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task. arXiv preprint arXiv:2006.07084v1, 2020.

Bobby Chesney and Danielle Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, 2019. Available at https://ssrn.com/abstract=3213954.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals. arXiv preprint arXiv:1901.02212v2, 2019.

Sourabh Dhere, Suresh B. Rathod, Sanket Aarankalle, Yash Lad, and Megh Gandhi. A Review on Face Reenactment Techniques. In 2020 International Conference on Industry 4.0 Technology (I4Tech), pages 191–194, Pune, India, 2020. IEEE.

Chris Donahue, Julian J. McAuley, and Miller S. Puckette. Adversarial Audio Synthesis. In 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA, 2019. OpenReview.net.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929v1, 2020.

Adam Geitgey. The World's Simplest Facial Recognition API for Python and the Command Line. Available at https://github.com/ageitgey/face_recognition.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozairy, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, pages 2672–2680. MIT Press, 2014.

Arushi Handa, Prerna Garg, and Vijay Khare. Masked Neural Style Transfer Using Convolutional Neural Networks. In 2018 International Conference on Recent Innovations in Electrical, Electronics Communication Engineering (ICRIEECE), pages 2099–2104, 2018.

Rahul Haridas and Jyothi R. L. Convolutional Neural Networks: A Comprehensive Survey. International Journal of Applied Engineering Research (IJAER), 14(03):780–789, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778. IEEE, 2016.

Yongjun Hong, Uiwon Hwang, Jaeyoon Yoo, and Sungroh Yoon. How Generative Adversarial Networks and Their Variants Work: An Overview. Volume 52, New York, NY, USA, 2019. Association for Computing Machinery.

He Huang, Phillip S. Yu, and Changhu Wang. An Introduction to Image Synthesis with Generative Adversarial Nets. arXiv preprint arXiv:1803.04469v1, 2018.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Ming-Yu Liu. Multimodal Unsupervised Image-to-Image Translation. In Computer Vision – ECCV 2018, pages 179–196. Springer International Publishing, 2018.

TackHyun Jung, SangWon Kim, and KeeCheon Kim. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. IEEE Access, 8:83144–83154, 2020.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv preprint arXiv:1812.04948, 2018.

Hasam Khalid and Simon S. Woo. OC-FakeDect: Classifying Deepfakes Using One-Class Variational Autoencoder. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 2794–2803, 2020.

Ali Khodabakhsh, Raghavendra Ramachandra, Kiran Raja, and Pankaj Wasnik. Fake Face Detection Methods: Can They Be Generalized? In 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), pages 1–6. IEEE, 2018.

Junyaup Kim, Siho Han, and Simon S. Woo. Classifying Genuine Face Images from Disguised Face Images. In 2019 IEEE International Conference on Big Data (Big Data), pages 6248–6250, 2019.

Pavel Korshunov and Sebastien Marcel. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. arXiv preprint arXiv:1812.08685, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Commun. ACM, 60(6):84–90, 2017.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. arXiv preprint arXiv:1609.04802v5, 2017.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. arXiv preprint arXiv:1806.02877v2, 2018.

Yuezun Li and Siwei Lyu. Exposing DeepFake Videos by Detecting Face Warping Artifacts. arXiv preprint arXiv:1811.00656v3, 2019.

Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-Consistent Video-to-Video Synthesis. In Computer Vision – ECCV 2020, pages 359–378. Springer International Publishing, 2020.

Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. Automatic Analysis of Facial Actions: A Survey. IEEE Transactions on Affective Computing, 10(3):325–347, 2019.

Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2023). Optimized feature extraction and cross-lingual text reuse detection using ensemble machine learning models. Journal of Computing & Biomedical Informatics, 5(01), 26-40.

Abid, K., Aslam, N., Fuzail, M., Maqbool, M. S., & Sajid, K. (2023). An efficient deep learning approach for prediction of student performance using neural network. VFAST Transactions on Software Engineering, 11(4), 67-79.

Kanwal, F., Abid, M. K., Maqbool, M. S., Aslam, N., & Fuzail, M. (2023). Optimized classification of cardiovascular disease using machine learning paradigms. VFAST Transactions on Software Engineering, 11(2), 140-148.

Aslam, N., Meeran, M. T., Aslam, M., Maqbool, M. S., & Saeed, B. (2025). Understanding Urban Expansion Through Multi-Temporal Satellite Data Analysis. *Kashf Journal of Multidisciplinary Research*, *2*(09), 252-273.

Hasnain, M. A., Ali, Z., Maqbool, M. S., & Aziz, M. (2024). X-ray image analysis for dental disease: A deep learning approach using efficientnets. *VFAST Transactions on Software Engineering*, *12*(3), 147-165.

Fazal, U., Khan, M., Maqbool, M. S., Bibi, H., & Nazeer, R. (2023). Sentiment analysis of omicron tweets by using machine learning models. *VFAST Transactions on Software Engineering*, *11*(1), 67-75.

Maqbool, M. S., Nazeer, R., Basit, A., & Zahra, K. (2023). Automated detection and localization of fungal infections on cotton leaves using YOLO-based object detection model. *Machines and Algorithms*, 2(2), 121-136.

Malik, F., Fuzail, M., Aslam, N., Sarwar, R., Abid, K., Maqbool, M. S., & Yousaf, A. (2024). A hybrid machine learning model to predict sentiment analysis on X. *Journal of Computing & Biomedical Informatics*, 6(02), 64-79.

Aslam, N., Meeran, M. T., Aslam, M., Maqbool, M. S., & Saeed, B. (2025). Understanding Urban Expansion Through Multi-Temporal Satellite Data Analysis. Kashf Journal of Multidisciplinary Research, 2(09), 252-273.

Hasnain, M. A., Ali, S., Malik, H., Irfan, M., & Maqbool, M. S. (2023). Deep learning-based classification of dental disease using x-rays. Journal of Computing & Biomedical Informatics, 5(01), 82-95.

Basit, A., Hanif, I., Maqbool, M. S., Qayyum, W., Hasnain, M. A., & Nazeer, R. (2023). Cross-lingual information retrieval in a hybrid query model for optimality. Journal of Computing & Biomedical Informatics, 5(01), 130-141.

Hasnain, M. A., Ali, Z., Maqbool, M. S., & Aziz, M. (2024). X-ray image analysis for dental disease: A deep learning approach using efficientnets. VFAST Transactions on Software Engineering, 12(3), 147-165.

Rafiqee, M. M., Qaiser, Z. H., Fuzail, M., Aslam, N., & Maqbool, M. S. (2023). Implementation of efficient deep fake detection technique on videos dataset using deep learning method. Journal of Computing & Biomedical Informatics, 5(01), 345-357.

Maqbool, M. S., Fatima, N., Nazeer, R., Aslam, N., Abbas, F., Sumra, U., & Nadeem, M. (2025). A HYBRID DATASET-BASED ENSEMBLE STRATEGY FOR EFFICIENT BREAST CANCER DETECTION. Kashf Journal of Multidisciplinary Research, 2(12), 39-57.

Prajwal K. R., Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C. V. Jawahar. Towards Automatic Face-to-Face Translation. In 27th ACM International Conference on Multimedia (MM '19), pages 1428–1436, New York, NY, USA. Association for Computing Machinery, 2019.

Md. Shohel Rana and Andrew H. Sung. DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection. In 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), pages 70–75, 2020.

Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder. In Computer Vision – ECCV 2020, pages 382–398. Springer International Publishing, 2020.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 2015.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning Lip Sync from Audio. ACM Trans. Graph., 36(4):780–789, 2017.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. Commun. ACM, 62(1):96–104, 2018.

Timesler. Pretrained PyTorch Face Detection (MTCNN) and Recognition (InceptionResnet) Models. Available at https://github.com/timesler/facenet-pytorch.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Information Fusion, 64:131–148, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), pages 6000–6010. Curran Associates Inc., 2017.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic Speech-Driven Facial Animation with GANs. International Journal of Computer Vision, 128:1398–1413, 2020.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), pages 1152–1164. Curran Associates Inc., 2018.

M. Arif Wani, Farooq Ahmad Bhat, Saduf Afzal, and Asif Iqbal Khan. Advances in Deep Learning. Volume 57 of Studies in Big Data. Springer Nature, Singapore, 2020.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. arXiv preprint arXiv:1905.08233v2, 2019.