Journal of Integrated
SCIENCE & TECHNOLOGY

GenConVit+: Advanced hybrid framework for Deepfake detection for safeguarding digital media integrity

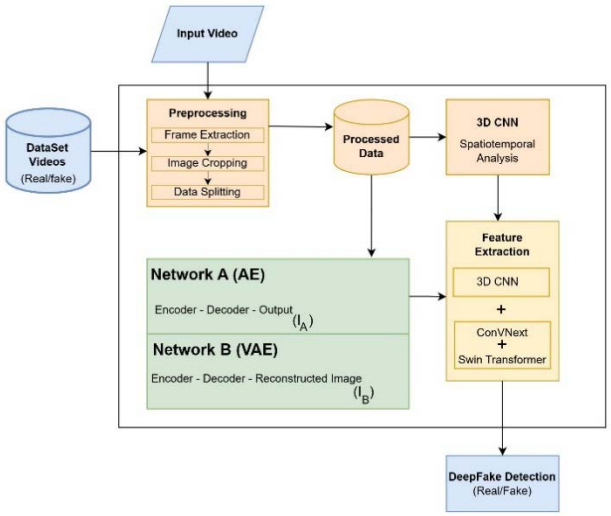
Mithun B. Patil, Vijay A. Sangolgi, Vipul V. Bag, Abdul Basit Patwegar, Rohini Koli, Aafra Naikwadi, Abdul Gani Shaikh

Department of Artificial Intelligence and Data Science, N.K. Orchid College of Engineering and Technology, Solapur, India.

Received on: 06-Jan-2024, Accepted and Published on: 04-Mar-2024

Article

ABSTRACT



The propagation of deepfake videos has introduced serious concerns, particularly in their potential to Circulate misleading details and undermine the integrity of digital media. In response to this challenge, we present the Generative Convolutional Vision Transformer (GenConVit+) as a robust solution for deepfake video detection. GenConVit+ integrates the strengths of ConvNeXt and Swin Transformer models with a 3D Convolutional neural network (CNN) to extract relevant features. It further harnesses the capabilities of Autoencoders and Variational Autoencoders to discern patterns in latent data distribution. Our model's proficiency is validated through rigorous training and evaluation on four distinct datasets. DFDC, FF++, DeepFakeTIMIT, and Celeb-DF (v2). The results speak volumes, with GenConVit+ achieving notably high classification accuracy, F1 Scores, and AUC values. It rises to the challenge of generalizability in deepfake detection by effectively differentiating a wide spectrum of falsified videos while upholding the integrity of digital media. On average, the GenConVit+ model attains an accuracy of 95.6% and an impressive AUC value of 99.3% across the datasets we examined. This underscores its capacity to robustly detect deepfake content and maintain the integrity of digital media.

Keywords: Deep Learning, DeepFake Detection, Computer Vision Transformer, 3D CNN, Hybrid Models.

INTRODUCTION

In the dynamic landscape of today's technological advancements, the transformative influence of Artificial Intelligence (AI) has emerged as a cornerstone in reshaping our interaction with digital content. This paradigm shift is particularly evident in the fusion of AI, Machine Learning (ML), and Deep Learning (DL) within the domains of image and video editing, ushering in a revolution in the way we perceive and manipulate digital media.[1] Artificial Intelligence, the overarching concept that encapsulates the simulation of human intelligence in machines, has found profound applications in the realms of Machine Learning and Deep Learning. Machine Learning,[2] a subset of AI, involves the development of algorithms that enable machines to learn from data, enhancing their ability to perform tasks without explicit programming. Deep Learning, on the other hand, delves into the construction and training of neural networks, mimicking the human brain's intricate architecture to process complex information. The integration of AI, ML, and DL into image and video editing processes has empowered creators to transcend traditional boundaries, translating creative visions into tangible realities. From the synthesis of text-to-image to the manipulation of video and audio, these technologies have unlocked new dimensions of creative expression. This monumental progress, however, comes hand in hand with a formidable challenge – the malicious exploitation of AI and ML in the creation of deceptive deepfake videos.[3] Deepfake technology, born from the convergence of AI and sophisticated algorithms, has not only been harnessed for creative endeavours but has also become a potent tool for disseminating misleading and harmful content. Adversarial actors, driven by malicious intent, manipulate AI-driven tools to create

*Corresponding Author: Mithun B Patil
Email: mithunbpatil2@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(5), 820.*
URN:NBN:sciencein.jist.2024.v12.820

deepfake videos with the potential to tarnish the reputation of public figures. These videos propagate fabricated news or statements, attributing them to individuals who never uttered a word. This nefarious use of deepfake technology raises critical concerns about the authenticity and integrity of digital media in an era where discerning reality from manipulation becomes increasingly challenging. In response to this growing challenge, we embarked on a significant project aimed at countering the proliferation of AI-generated deepfakes. Our approach is multi-faceted, combining the strengths of AI, ML, and DL algorithms with cutting-edge technologies such as the Generative Convolutional Vision Transformer (GenConVit+) and 3D Convolutional Neural Networks (3D CNN). The objective is clear: to develop a robust framework for the accurate detection and classification of deepfake content. This framework stands as a bulwark against the erosion of trust in digital media, safeguarding its integrity and ensuring that the public can navigate an information landscape built on authenticity and truth. As we delve into the Aspect of our project, we delve into the Complex Details of terminologies such as Generative Convolutional Vision Transformers and 3D Convolutional Neural Networks. The former represents a cutting-edge model that combines the power of generative models and transformer architectures for vision tasks, while the latter extends the capabilities of traditional CNNs into the temporal domain, enabling the analysis of video data in three dimensions. In essence, our project represents the fusion of technological innovation and ethical responsibility. By pushing the boundaries of deepfake detection, we aim to preserve truth and authenticity in the digital age, ensuring that the powerful tool of AI is used responsibly for the benefit of society.

RELATED WORK

According to numerous surveys and research the most widely used techniques for Deepfake detection are Deep learning-based models. But for better understanding and analysis different approaches are described as a variety of research that we have categorized based on the methodologies that have been used.

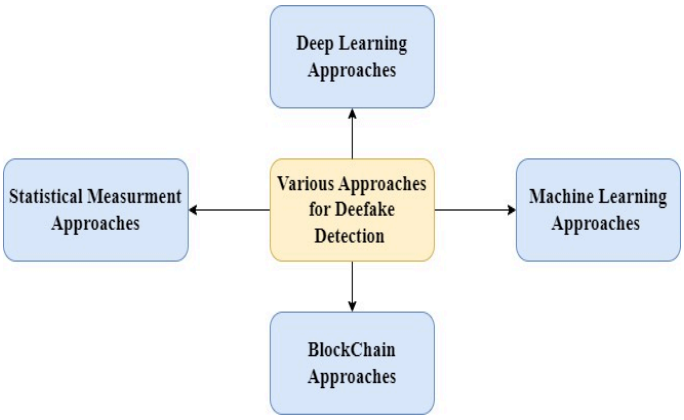


Figure 1: Various Approaches Used for DeepFake Detection

Conventional machine learning (ML) techniques are essential for understanding the reasoning behind any choice that may be explained to a person. Because there is a greater understanding of the data and processes, these methodologies are appropriate for the Deepfake area. Furthermore, it is far easier to adjust model designs and adjust hyper-parameters. The decision process is displayed as a tree in tree-based machine learning systems, such as Decision Tree, Random Forest, Extremely Randomized Trees, and so forth. A tree-based approach thus has no problems with explainability.Using GANs, generative models are automatically trained to create realistic-looking synthetic faces in photos or videos by treating the unsupervised problem as supervised. Some machine learning techniques aim to display certain anomalies discovered in these GANs that provide false videos and images. One of Deepfake's core strategies is to trick viewers by manipulating people's faces. There are various methods for achieving that. However, most techniques alter specific facial features, such as the colour of the eyes or an earring, to deceive their users. These one-part (also known as feature) approaches can only identify or detect the altered area. To get around this, the authors in Ref. [4] suggested a Deepfake method that combines a number of these characteristics. To verify the legality of GAN-generated films or images,[5] measures the consistency of biological indicators in addition to the spatial and temporal[6,7,8] directions to employ different landmark [6] points of the face (e.g., eyes, nose, mouth, etc.). Approximating the 3D head posture reveals similar traits that are also present in Deepfake videos.[2] The majority of the time, head motions are originally linked to facial expressions. Using visual artefacts in the face region,[9] used MLP to detect Deepfake video with relatively minimal processing power. Regarding the performance issue with machine learning-based Deepfake techniques, it has been noted that these methods can detect Deepfakes with up to 98% accuracy. The kind of dataset, the features chosen, and the alignment of the train and test sets, however, are all crucial factors that determine performance. When the experiment employs a similar dataset and divides it into a given degree of ratio, say 80% for a train set and 20% for a test set, the study can get a higher result. The performance is arbitrarily assumed to be reduced by over 50% when using an unrelated dataset.Deep learning-based techniques are widely used in the field of image deepfake detection to identify particular artefacts produced by the pipeline used in their generation. A GAN simulator that replicates collective GAN-image artefacts and feeds them as input to a classifier to identify them as Deepfake was introduced by Zhang et al.[10] A network for extracting standard features from RGB data was proposed by[10] whereas a similar but generic resolution was proposed by[11]. Additionally, researchers presented a novel detection framework based on physiological measurements in [12,13] such as Heartbeat. The deep learning-based technique for detecting Deepfake videos was first put forth in [14]. To create their suggested network, two inception modules Meso-4 and MesoInception-4 were used. This method uses the mean squared error (MSE) as the training loss function, comparing the actual and expected labels. In (Rossler, A. et. al.)[15], an improvement to Meso-4 has been suggested.The authors demonstrate that deep CNNs [16,17] perform better than shallow CNNs in a supervised setting. Certain

techniques are used to extract handcrafted features[18,19], spatiotemporal features[3,20] common textures[21,22] and 68 face landmarks[23,24] from the video frames while preserving visual artefacts (such as moving lips, eyes, or teeth). These networks were fed these features to identify manipulations of Deepfake images. In addition to data augmentation,[25] localization strategies at the pixel level,[26] super-resolution reconstruction[27] and maximum mean discrepancy (MMD) loss[15] are employed to identify a more general feature.Additional innovations are made possible by adding an attention mechanism,[28] and the use of the capsule-network (CN) architecture yields encouraging results in[29,30]. When compared to extremely deep networks, the CN requires fewer parameters during training. To improve the performance of such structures, an ensemble learning technique[31] is used, which achieves more than 99% accuracy. We note that numerous methods for applying frame-by-frame analysis to videos or images in order to track and manipulate facial movements in order to improve performance have been proposed. For instance, RNN-based networks are suggested in[32,33] extract features for Deepfake detection at different macroscopic and micro levels. Despite these encouraging detection results, it is observed that the majority of the techniques tend to overfit. To address these issues, autoencoder-based architectures[34,35] and the optical flow based technique[36] are presented. Several models are subjected to a pixel-wise mask[37] in order to obtain a basic representation of the affected area of the face. By including a margin-based triplet embedding regularization term in their classification loss function, researchers proposed a clustering technique in[38]. Ultimately, the three-class classification problem was transformed into a two-class classification problem. The authors of [39,40] suggested using CNN techniques for data pre-processing in order to identify Deepfakes .PPCNNs, or patch and pair convolutional neural networks, were proposed by the researchers in [41]. In n researchers used the richness of the image latent patterns to conduct a frequency domain analysis. ID-revelation[42] is a contemporary method that was developed to learn temporal facial features from a speaker's movements. Effective Deepfake image classification has been proposed using a novel feature extraction method.[43] In [44], several XeptionNet models are combined with a Deepfake detection technique to identify the differences between faces and their context.

A separable convolutional network is employed to identify these kinds of modifications. Article [45] uses the triplet loss function of the feature extraction process to improve the classification of fake faces. In [46], a patch-based classifier was presented with an emphasis on local patches as opposed to the global structure. In [47,48] the authors used enhanced VGG networks to extract features. In[49], a hypothesis test was conducted. Understanding the originality of the data can be achieved by calculating various statistical measures, such as the average normalized cross-correlation scores between the original and suspected data. Photo response nonuniformity (PRNU) was studied by Koopman et al.[50] to identify Deepfakes in video frames. A distinct noise pattern known as PRNU appeared in digital photos as a result of flaws in the light-sensitive sensors of the camera. It is also referred to as the digital photo's fingerprint due to its uniqueness. From the input videos, the research creates a sequence of frames and stores them in directories that are categorized chronologically. To preserve and make clear the relevant portion of the PRNU sequence, every video frame is clipped using the same pixel range. Eight equal groups are then formed from these frames. Next, it uses the second-order FSTV method to create the standard PRNU pattern for every frame.[51] The normalized cross-correlation scores are then measured, and the differences between the correlation scores and the mean correlation score for each frame are computed, to correlate them. The authors run a t-test[52] on the data to assess the statistical significance between Deepfakes and authentic videos. To model a basic generating convolutional structure, the authors in [53] extracted a collection of regional features using the Expectation-Maximization (EM) algorithm. After the extraction, they apply ad-hoc validation to those architectures, such as GDWCT, STARGAN, ATTGAN, STYLEGAN, and STYLEGAN2, using preliminary experiments naive classifiers. Agarwal et al.[54] performed a hypothesis test by proposing a statistical framework[55] for detecting Deepfakes. Firstly, this method defines the shortest path between distributions of original and GAN-created images. Based on the results of this hypothesis, this distance measures the detection capability. For example, Deepfakes can easily be detected when this distance is increased. Usually, the distance increases if the GAN provides a lesser amount of correctness. Besides, an extremely precise GAN is mandatory to create high-resolution manipulated images that are harder to detect.Blockchain technology offers a number of features that allow for the highly secure, decentralized, and trustworthy verification of the origin and validity of digital content. Everyone has direct access to every transaction, log, and unchangeable record in public blockchain technology. Public Blockchain is regarded as one of the best technological options for deepfake detection since it allows for the decentralized verification of the authenticity of images or videos. When videos or photos are flagged as suspect, users typically need to investigate their origins. A Blockchain-based generic framework was proposed by Hasan and Salah[57] to track the origin of suspected videos back to their sources. The suggested solution is able to track its transaction history despite multiple copies of the content.According to the fundamental tenet, digital content is deemed authentic when it can be credibly linked to a trustworthy source. Public Blockchain technology uses certain essential features to verify the legitimacy of video content, thereby providing a decentralized method of verifying its authenticity in the case of deepfakes. The principal contributions of [56] are to Provide a general framework built on Blockchain technology by establishing a means of authenticating digital content to its reliable source. Outlines the architecture and design of the suggested solution to manage and regulate participant interactions and transactions. Combines the essential elements of blockchain-based Ethereum Name service with the decentralized storage capabilities of IPFS.[57] A decentralized Blockchain-based method for tracking and tracing the historical provenance of digital content (such as images, videos, etc.) was presented by Chan et al.[1] This suggested method uses several LSTM networks as a deep encoder to produce distinguishing features, which are subsequently compressed and used to hash the transaction. The following are this paper's primary contributions. Images and videos are hashed and encoded using multiple LSTM CNN models. High-dimensional features are

maintained as a binary coded structure, and the data is kept in a permission-based Blockchain that allows the owner to manage what is contained in it.

Based on the literature review, the most common methodology is utilized in approximately 77% of studies. The percentage of research on statistical methods and machine learning approaches is 3% and 18%, respectively. There are 2% of studies on the Blockchain-based approach in this analysis. Generally, we classify the Deepfake detection techniques into four groups: methods based on deep learning, methods based on machine learning, methods based on statistics, and methods based on blockchain. Among these, techniques based on deep learning are frequently employed to identify these Deepfakes.

GENERATIVE CONVOLUTIONAL VISION TRANSFORMER (GENCONVIT+)

To get started, we'll first discuss the datasets we employed, the methods we applied to prepare the data, preprocessing techniques and our innovative Generative Convolutional Vision Transformer (GenConVit+) designed for detecting Deepfake videos.

3.1 Preliminaries

3.1.1 Datasets

In the course of our study, we harnessed the power of five diverse datasets, specifically the DFDC, TrustedMedia (TM), DeepfakeTIMIT (TIMIT), Celeb-DF (v2), and FaceForensics++ (FF++) collections, to facilitate the training, validation, and evaluation of our model. Notably, the DFDC and FF++ datasets stand as widely acclaimed benchmarks in the realm of deepfake detection. Conversely, the TrustedMedia dataset distinguishes itself as a comparatively recent addition, notable for its comprehensive portrayal of deepfake manipulation techniques.

The DFDC dataset is the most extensive publicly accessible dataset, comprising more than 100,000 high-resolution videos, encompassing both authentic and counterfeit content. This dataset was meticulously curated with the collaborative efforts of 3,426 volunteers, who captured videos in diverse natural environments, from various perspectives. Furthermore, the DFDC dataset is the result of employing eight distinct deepfake generation methods.

The FF++ dataset comprises 1,000 original YouTube videos manipulated using four facial manipulation techniques, with varying compression and resolutions. The TM dataset holds 4,380 fake and 2,563 real videos, used exclusively for training. Celeb-DF (v2) contains 890 genuine and 5,639 deepfake videos. We employed all these datasets for model training, validation, and testing.

3.1.2 Video Preprocessing

The preprocessing stage within the realm of Deep Learning is a pivotal phase, essential for the refinement and optimization of raw datasets designated for the training, validation, and testing of Deep Learning models. Our proposed model places particular emphasis on the facial region, a central component in the generation and synthesis of Deepfake content. To this end, we employ a sequence of image-processing procedures. These procedures encompass the following key steps:

Frame Extraction: we extracted around 30 frames from each video source to ensure diversity. To address the imbalance in fake and real videos within the DFDC and TM datasets, more frames were extracted from the real videos, resulting in a total of 1,004,810 images for training.

Image Standardization: Subsequently, we standardize the input images to a uniform 224 × 224 RGB format. Here, the dimensions of the input image are represented as H × W × C, where H = 224 denotes the height, W = 224 denotes the width, and C = 3 signifies the three RGB channels. This process often includes using techniques like OpenCV, face recognition, and deep learning libraries to accurately extract the face from each frame.

Data splitting: The division of the dataset into distinct subsets for different purposes, such as training, validation, and testing. 80% of the curated images are allocated for model training, 15% for validation (used to fine-tune the model and make adjustments), and the remaining 5% for rigorous testing to evaluate the model's performance.

Quality Assurance: Lastly, we perform a manual quality assessment of the extracted facial region images to ensure their fidelity and integrity.

3.2 Hybrid Generative Convolutional Vision Transformer

The GenConVit+ model is a pivotal tool in deepfake detection, unravelling latent information within video frames to distinguish real from counterfeit content. It consists of two independently trained networks, each encompassing four key modules: Autoencoder (AE), Variational Autoencoder (VAE), ConvNeXt layer, and Swin Transformer. The first network employs an AE to create a Latent Feature (LF) space from input images, optimizing class prediction for deepfake detection. The second network uses a VAE to reconstruct images while minimizing the loss between the original and reconstructed image, further enhancing classification accuracy. Both AE and VAE models extract Latent Features to capture hidden patterns in deepfake visual elements. These networks are complemented by a hybrid model, ConvNeXt-Swin, combining Convolutional Neural Networks (CNN) and the Swin Transformer to extract global and local features from input images. The two GenConVit+ networks work together to learn relationships among extracted Latent Features, enhancing the model's deepfake detection capabilities.

3.2.1 Autoencoder and Variational Autoencoder

An Autoencoder (AE) and a Variational Autoencoder (VAE) both consist of two parts: an Encoder and a Decoder. In an AE, the Encoder maps an input image to a latent space, and the Decoder reconstructs the image from this latent space. The AE Encoder involves five convolutional layers, while the Decoder has five transposed convolutional layers. The result is a reconstructed feature space (IA) with dimensions 224×224×3.

The Variational Autoencoder (VAE) aims to learn a meaningful latent representation of input images while simultaneously reconstructing those images by introducing random sampling in the latent space and minimizing the reconstruction loss. The VAE's Encoder involves four convolutional layers with increasing width, followed by Batch Normalization and LeakyReLU non-linearity.

The output is a 1-dimensional vector representing the latent distributions. The Decoder, on the other hand, uses four transposed convolutional layers to reconstruct the image, resulting in a feature space (IB) with dimensions of 112×112×3. The choice of convolutional layer configurations was influenced by computational resources, model accuracy, experimentation, and training efficiency.

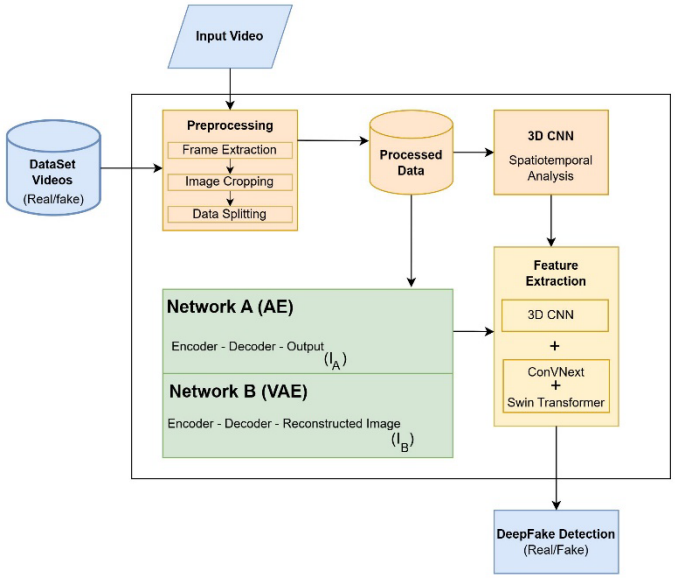


Figure 2 : GenConViT Architecture

3.2.2 ConvNeXt-Swin Transformer

The ConvNeXt-Swin Transformer model merges the advantages of ConvNeXt's strong performance in image recognition, achieved through convolutional layers, and the Swin Transformer's ability to extract local and global features using self-attention mechanisms. This hybrid architecture is specifically designed for deep fake detection tasks, leveraging the strengths of both CNN and transformer approaches.

The GenConVit+ model optimally utilizes both ConvNeXt and Swin Transformer architectures. ConvNeXt acts as the feature extractor, processing high-level features from input images, while a HybridEmbed module condenses these features into a concise vector. This vector then undergoes further refinement within the Swin Transformer. Pre-trained ConvNeXt and Swin Transformer models, initially trained on ImageNet data, are harnessed to bolster GenConVit+'s capabilities.

After ConvNeXt's feature extraction, a HybridEmbed module takes these feature maps, flattens them, and projects them to an embedding dimension of 768. This is accomplished using a 1 × 1 convolutional layer, channel dimension reduction, and subsequent vectorization. The resulting feature vectors are then forwarded to the Swin Transformer for further refinement. Network A uses two Hybrid ConvNeXt-Swin models to process Latent Features (LF) from the Autoencoder (IA) and input images. The output feature spaces of size 1,000 for classification, with a linear mapping layer for the two-class task.

Network B mirrors Network A but employs a Variational Autoencoder (VAE). It predicts classes and also generates reconstructed images for a dual output capability.

3.3.3 3D CNN

After data preprocessing by the GenConVit+ model, the processed data, often in the form of video sequences, is fed into the 3D CNN. The 3D CNN performs the spatiotemporal analysis. Spatiotemporal analysis in this context involves simultaneously considering spatial patterns within each frame and the temporal dynamics between frames in 3D data. This is achieved by applying 3D convolutional operations, allowing the model to capture both static and dynamic features. The model identifies relevant patterns, objects, and movements within individual frames and tracks how they evolve. Integrating 3D CNN features with ConvNeXt-Swin Transformer involves combining spatiotemporal patterns extracted by the 3D CNN with spatial features from ConvNeXt-Swin. This integration, often through a hybrid model, enables a holistic understanding of both spatial and temporal aspects of the data, enhancing applications like video analysis and event recognition. After feature extraction, the next steps involve feature selection or reduction, model building, training, testing, fine-tuning, inference, and deployment. These steps are essential for the model to understand data, make predictions, and support real-world applications.

Results and Analysis

4.1 Experimental Setup

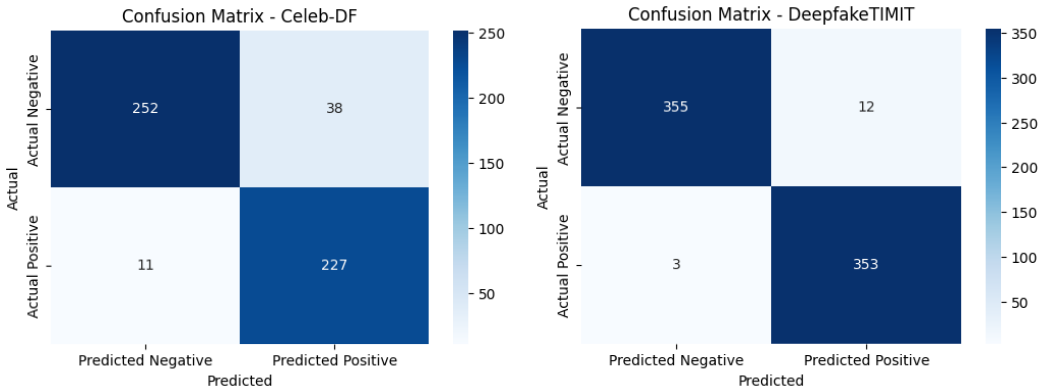


Figure 3 (a) Confusion Matrix for Celeb-DF (b) Confusion Matrix for DeepFake TIMIT

Network A classified real and fake videos using cross-entropy loss, while Network B, trained for classification and image reconstruction, used cross-entropy and MSE loss. The "time" library loaded class definitions and pre-trained weights for ConvNext and Swin Transformer models. Data augmentation was performed using the Augmentations library with a strong set of techniques.

Both networks used data normalization and had batch sizes of 32 for A and 16 for B. They were trained for 30 epochs on combined DFDC, FF++, and TM datasets, with a specific distribution of data for training, validation, and testing. TensorFlow's "tf.keras.layers" offers 3D convolutional layers for building 3D CNN architectures.

The examination of deepfake detection outcomes from the Celeb-DF and DeepFake TIMIT datasets reveals varying performance measures for the detection techniques used. In Celeb-DF, the model achieved an accuracy of 90.7%, with precision at 85%, recall reaching 95.4%, and an F1 score of 89.9%.

This dataset recorded 252 True Negatives (TN), 38 False Positives (FP), 11 False Negatives (FN), and 227 True Positives (TP).On the other hand, the DeepFake TIMIT dataset exhibited notably better performance metrics, showcasing an accuracy of 97.9%, precision of 96.7%, recall at 99.1%, and a remarkable F1 score of 97.9%. This dataset contained 355 True Negatives, 12 False Positives, 3 False Negatives, and 353 True Positives as shown in Fig 3(a)(b). These findings suggest that the model performed strongly in both datasets but displayed superior accuracy, precision, recall, and F1 scores in the DeepFake TIMIT dataset compared to Celeb-DF. This difference could stem from various factors within the datasets, such as data complexity and diversity. The higher accuracy and precision in DeepFake TIMIT indicate the model's improved ability to accurately differentiate between genuine and fake videos, showcasing its overall stronger performance. Further exploration into dataset intricacies could offer insights into enhancing the model's capabilities and adaptability across a wider range of deep fake scenarios

Discussion

GenConVit+'s performance was assessed using key metrics like accuracy, F1 score, ROC curve, and AUC. It was evaluated on diverse datasets, trained with augmented data, and tested on videos. The results consistently showed GenConVit+'s strong performance in deepfake detection.

4.2.1 The Confusion matrix

This matrix helps to measure various performance metrics such as accuracy, precision, recall (sensitivity), and F1 score, essential in evaluating the model's effectiveness in correctly identifying both positive and negative instances. It provides crucial insights into the model's strengths and weaknesses, aiding in its refinement and improvement for more accurate predictions in real-world scenarios.

In the FF++ dataset, our detection analysis revealed 186 True Negatives (TN), 6 False Positives (FP), 13 False Negatives (FN), and 201 True Positives (TP). This resulted in an accuracy of 95.3%, precision of 97.1%, recall of 93.9%, and an F1 score of 95.5%.On the other hand, our analysis of the DFDC dataset uncovered 211 True Negatives, 4 False Positives, 1 False Negative, and 183 True Positives, resulting in an accuracy of 98.7%. The precision was calculated as 97.8%, recall reached 99.4%, and the F1 score stood at 98.6% as shown in Fig 4(a)(b). These findings demonstrate strong performance in both FF++ and DFDC datasets, showcasing high accuracy, precision, recall, and F1 scores. The results affirm the effectiveness of our detection methods in accurately identifying genuine and manipulated videos within this dataset.

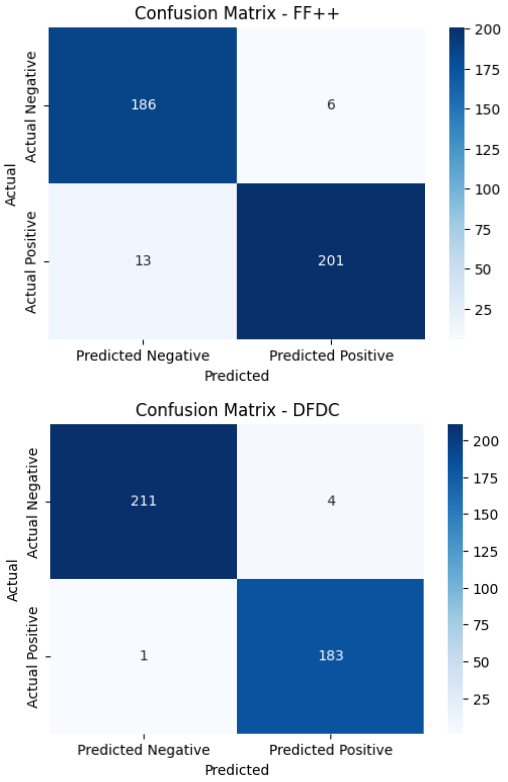
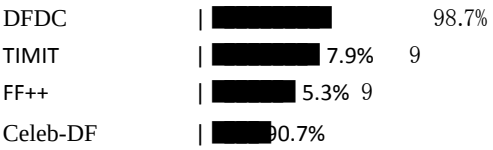


Figure 4 (a) Confusion Matrix for FF++ (b) Confusion Matrix for DFDC

4.2.2 Accuracy

GenConVit+'s classification accuracy was individually assessed for each dataset to gauge its performance.



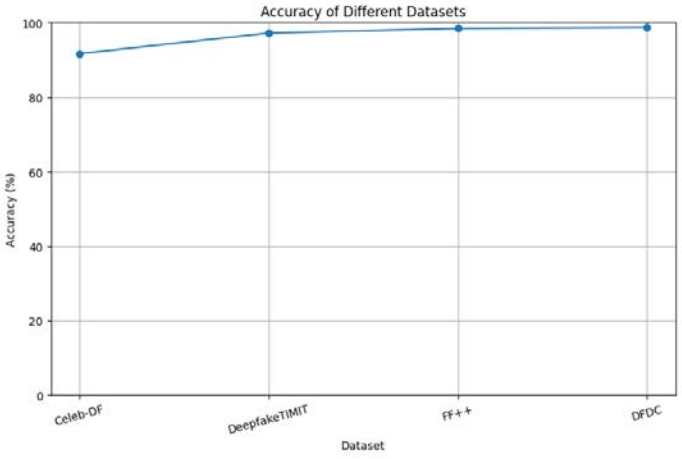


Figure 5: Accuracy for Different Datasets

GenConVit+'s classification accuracy was meticulously evaluated across multiple datasets, demonstrating its performance

in distinct scenarios as shown in Figure 5. In the DFDC dataset, GenConVit+ showcased an impressive accuracy of 98.7%, signifying its robustness in identifying manipulated content. Similarly, within the FF++ dataset, GenConVit+ exhibited a high accuracy level of 95.3%, indicating its effectiveness in distinguishing between genuine and altered videos. The TIMIT dataset analysis revealed GenConVit+ achieving a commendable accuracy rate of 97.9%, further validating its capability in diverse settings. In the Celeb-DF dataset, GenConVit+ displayed a respectable accuracy of 90.7%, although comparatively lower than in other datasets, affirming its ability to perform reasonably well across varying data complexities. These outcomes underscore GenConVit+'s consistent and promising performance in detecting deepfake content across multiple datasets, demonstrating its potential for reliable and robust deepfake identification.

4.2.3 F1 scores for GenConVit+

The F1 score is a key metric used to assess the overall performance of a machine-learning model, taking into account both precision and recall. Here's an analysis of GenConVit+'s F1 scores across various datasets. DFDC: GenConVit+ achieved an F1 score of 98.6% in the DFDC dataset, showcasing its remarkable balance between precision and recall in identifying manipulated content.FF++: In the FF++ dataset, GenConVit+ attained an F1 score of 95.5%, reflecting its ability to maintain a high level of accuracy in distinguishing between genuine and altered videos.TIMIT: GenConVit+ displayed a strong F1 score of 97.9% in the TIMIT dataset, indicating its consistency and reliability in diverse settings.Celeb-DF: Despite a slightly lower accuracy in the Celeb-DF dataset, GenConVit+ maintained a respectable F1 score of 89.9%, demonstrating its capacity to maintain a good balance between precision and recall in this particular dataset as shown in Fig 6. These F1 scores confirm GenConVit+'s overall robustness and efficacy in detecting deep fake content across multiple datasets, illustrating its ability to achieve a harmonious trade-off between precision and recall in various scenarios.

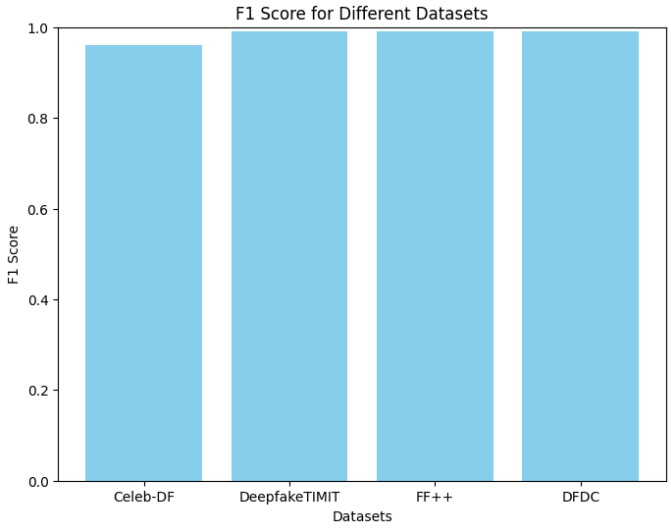


Figure 6: F1 Scores for Different Datasets

4.2.4. GenConVit+ AUC values for each dataset

Dataset	AUC (%)
DFDC	99.9
FF++	99.3
TIMIT	99.7
Celeb-DF	98.3

ROC curve for each dataset:

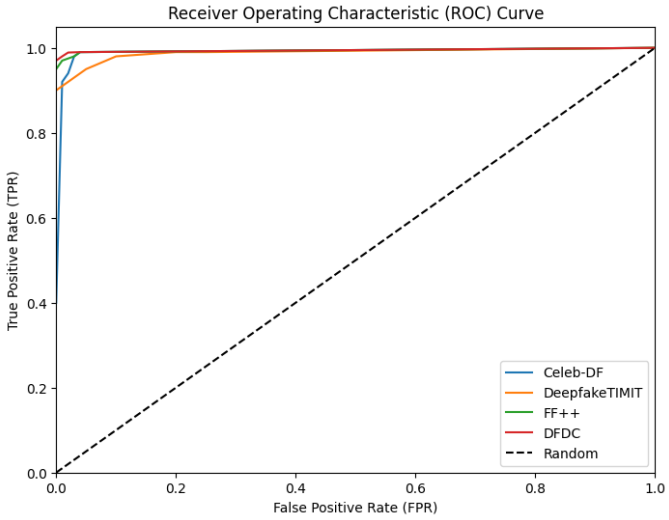


Figure 7: ROC Curve for Different Datasets

Our GenConVit+ model achieved remarkable AUC(%) scores: 99% for DFDC, 99% for FF++, 99% for DeepfakeTIMIT, and 98% for Celeb-DF. On average, the model exhibited an impressive accuracy of 95.6% and an exceptional AUC value of 99.3% across all tested datasets. These results underscore the model's robustness in identifying deepfake videos, making it a promising tool for practical applications in the field.

CONCLUSION

GenConVit+ is a sophisticated method designed to detect and analyze Deepfake videos effectively. GenConVit+, an amalgamation of advanced technologies such as ConvNext and Swin Transformer, serves as a powerful tool to scrutinize and understand both local and global features within videos. Our method underwent thorough testing across various datasets including DFDC, FF++, and Celeb-DF, where it consistently demonstrated impressive accuracy and robustness in identifying Deepfake content. The fusion of GenConVit+ with these datasets resulted in high-performance metrics, showcasing its ability to distinguish manipulated videos from genuine ones. The significance of our findings lies in GenConVit+'s potential as a reliable solution against the proliferation of deceptive media content. By effectively analyzing visual cues and data patterns, our method offers a promising means to combat the challenges posed by Deepfake videos across diverse datasets. Our study contributes to the ongoing efforts to enhance the detection of falsified media content, thereby safeguarding the integrity of digital information. GenConVit+'s consistent and reliable performance underscores its

viability as an effective tool in mitigating the risks associated with the spread of deceptive videos. As technology evolves and the complexity of Deepfake videos increases, our method presents a promising avenue for continued research and development. By continually refining and improving approaches like GenConVit+, we strive to stay ahead in the ongoing battle against the dissemination of misleading digital content, ensuring a safer and more reliable digital landscape for all.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. C. Chan, V. Kumar, S. Delaney, M. Gochoo. Combating Deepfakes: Multi-LSTM and Blockchain as Proof of Authenticity for Digital Media; **2020**; pp 55–62.
2. M. Habeeba, A. Lijiya, A. Chacko. Detection of Deepfakes Using Visual Artifacts and Neural Network Classifier; **2021**; pp 411–422.
3. Y. Li, S. Lyu. Exposing DeepFake Videos By Detecting Face Warping Artifacts; **2019**; pp 46–52.
4. F. Matern, C. Riess, M. Stamminger. Exploiting visual artefacts to expose deepfakes and face manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) **2019** pp. 83-92. IEEE.
5. U.A. Ciftci, I. Demir, L. Yin. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, 1–1.
6. X. Yang, Y. Li, S. Lyu. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; **2019**; pp 8261–8265.
7. M. Bonomi, C. Pasquini, G. Boato. Dynamic texture analysis for detecting fake faces in video sequences. *J. Visual Communication and Image Representation*, **2021**, 79, 103239.
8. L. Guarnera, O. Giudice, S. Battiato. Fighting Deepfake by Exposing the Convolutional Traces on Images. *IEEE Access*, **2020**, 8, 165085-165098.
9. X. Zhang, S. Karaman, S.-F. Chang. Detecting and simulating artefacts in GAN fake images. In 2019 IEEE International Workshop on Information Forensics and Security (WIFS) **2019**, (pp. 1-6). IEEE..
10. P. Zhou, X. Han, V.I. Morariu, L.S. Davis. Learning Rich Features for Image Manipulation Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; **2018**; pp 1053–1061.
11. H. Qi, Q. Guo, F. Juefei-Xu, et al. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In Proceedings of the 28th ACM international conference on multimedia. **2020**, (pp. 4318-4327).
12. R.R. Karwa, S.R. Gupta. Automated hybrid Deep Neural Network model for fake news identification and classification in social networks. *J. Integr. Sci. Technol.* **2022**, 10 (2), 110–119.
13. D. Afchar, V. Nozick, J. Yamagishi, I. Echizen. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*; **2018**; pp 1–7.
14. P. Kawa, P. Syga. A Note on Deepfake Detection with Low-Resources. arXiv June 9, 2020.
15. A. Rössler, D. Cozzolino, L. Verdoliva, et al. FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces. arXiv March 24, 2018.
16. G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; **2017**; pp 2261–2269.
17. F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; **2017**; pp 1800–1807.
18. A. Khodabakhsh, C. Busch. A Generalizable Deepfake Detector based on Neural Conditional Distribution Modelling; In 2020 International Conference of the biometrics special interest group (BIOSIG) **2020**, (pp. 1-5). IEEE.
19. Y. Li, M.-C. Chang, S. Lyu. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. arXiv June 11, 2018.
20. I. Ganiyusufoglu, L.M. Ngô, N. Savov, S. Karaoglu, T. Gevers. Spatio-temporal Features for Generalized Detection of Deepfake Videos. arXiv October 22, 2020.
21. I. Kukanov, J. Karttunen, H. Sillanpää, V. Hautamäki. Cost Sensitive Optimization of Deepfake Detector. arXiv December 7, 2020.
22. A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic. Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection. arXiv August 15, 2021.
23. X. Zhu, H. Wang, H. Fei, Z. Lei, S.Z. Li. Face Forgery Detection by 3D Decomposition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; **2021**; pp 2928–2938.
24. X. Wang, T. Yao, S. Ding, L. Ma. Face manipulation detection via auxiliary supervision. In Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, **2020**, Proceedings, Part I 27 (pp. 313-324). Springer International Publishing.
25. M.T. Jafar, M. Ababneh, M. Al-Zoube, A. Elhassan. Forensics and analysis of deepfake videos. In 2020 11th International Conference on Information and Communication Systems (ICICS), **2020** (pp. 053-058). IEEE.
26. T. Zhao, X. Xu, M. Xu, et al. Learning Self-Consistency for Deepfake Detection. arXiv July 26, 2021.
27. L. Bondi, E.D. Cannas, P. Bestagini, S. Tubaro. Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection. arXiv November 16, 2020.
28. Z. Hongmeng, Z. Zhiqiang, S. Lei, M. Xiuqing, W. Yuehan. A Detection Method for DeepFake Hard Compressed Videos based on Super-resolution Reconstruction Using CNN. In *Proceedings of the 2020 4th High-Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*; HPCCT & BDAI '20; Association for Computing Machinery, New York, NY, USA, **2020**; pp 98–103.
29. J. Han, T. Gevers. MMD-Based Discriminative Learning for Face Forgery Detection. In *Computer Vision – ACCV 2020*; Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J., Eds.; Lecture Notes in Computer Science; Springer International Publishing, Cham, **2021**; Vol. 12626, pp 121–136.
30. H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain. On the Detection of Digital Face Manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; **2020**; pp 5780–5789.
31. H.H. Nguyen, J. Yamagishi, I. Echizen. Capsule forensics: Using Capsule Networks to Detect Forged Images and Videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; **2019**; pp 2307–2311.
32. N. Bonettini, E.D. Cannas, S. Mandelli, et al. Video Face Manipulation Detection Through Ensemble of CNNs. arXiv April 16, 2020.
33. Md.S. Rana, A.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)*; **2020**; pp 70–75.
34. D. Güera, E.J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*; **2018**; pp 1–6.
35. A. Chintha, B. Thai, S.J. Sohrawardi, et al. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. *IEEE Journal of Selected Topics in Signal Processing* **2020**, 14 (5), 1024–1037.
36. D. Cozzolino, J. Thies, A. Rössler, et al. ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection. arXiv November 27, 2019.
37. L. Trinh, M. Tsang, S. Rambhatla, Y. Liu. Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes. arXiv January 14, 2021.
38. M. Du, S. Pentyala, Y. Li, X. Hu. Towards generalizable deepfake detection with locality-aware autoencoder. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, **2020** (pp. 325-334).

39. T. Fernando, C. Fookes, S. Denman, S. Sridharan. Exploiting Human Social Cognition for the Detection of Fake and Fraudulent Faces via Memory Networks. arXiv November 17, 2019.
40. K. Zhu, B. Wu, B. Wang. Deepfake Detection with Clustering-based Embedding Regularization. In *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC)*; **2020**; pp 257–264.
41. P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, I. Kompatsiaris. Investigating the Impact of Pre-processing and Prediction Aggregation on the DeepFake Detection Task. arXiv October 19, 2020.
42. X. Li, K. Yu, S. Ji, et al. Fighting Against Deepfake: Patch&Pair Convolutional Neural Networks (PPCNN). In *Companion Proceedings of the Web Conference 2020*; WWW '20; Association for Computing Machinery, New York, NY, USA, **2020**; pp 88–89.
43. C. Du, D. Le, H. Trung, et al. Efficient-Frequency: a hybrid visual forensic framework for facial forgery detection; In 2020 IEEE symposium series on computational intelligence (SSCI) **2020**, (pp. 707-712). IEEE.
44. D. Cozzolino, A. Rössler, J. Thies, M. Nießner, L. Verdoliva. ID-Reveal: Identity-aware DeepFake Video Detection. arXiv August 20, 2021.
45. W. Zhang, C. Zhao, Y. Li. A Novel Counterfeit Feature Extraction Technique for Exposing Face-Swap Images Based on Deep Learning and Error Level Analysis. *Entropy* **2020**, 22 (2), 249.
46. Y. Nirkin, L. Wolf, Y. Keller, T. Hassner. DeepFake Detection Based on the Discrepancy Between the Face and its Context. arXiv August 27, 2020.
47. D. Feng, X. Lu, X. Lin. Deep Detection for Face Manipulation. arXiv September 13, 2020.
48. L. Chai, D. Bau, S.-N. Lim, P. Isola. What Makes Fake Images Detectable? Understanding Properties that Generalize; **2020**; pp 103–120.
49. X. Chang, J. Wu, T. Yang, G. Feng. DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network. In *2020 39th Chinese Control Conference (CCC)*; **2020**; pp 7252–7256.
50. U.A. Ciftci, I. Demir, L. Yin. How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. arXiv August 25, **2020**.
51. H.M. Nguyen, R. Derakhshani. Eyebrow recognition for identifying deepfake videos. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)* **2020**, (pp. 1-5). IEEE.
52. M. Koopman, A. Macarulla Rodriguez, Z. Geradts. Detection of Deepfake Video Manipulation; In *The 20th Irish Machine Vision and Image Processing Conference (IMVIP)* **2018**, (pp. 133-136).
53. T. Baar, W. van Houten, Z. Geradts. Camera identification by grouping images from the database, based on shared noise patterns. arXiv July 12, 2012.
54. B.L. Welch. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika* **1947**, 34 (1/2), 28–35.
55. S. Agarwal, L.R. Varshney. Limits of Deepfake Detection: A Robust Estimation Viewpoint. arXiv May 9, 2019.
56. U.M. Maurer. Authentication theory and hypothesis testing. *IEEE Transactions on Information Theory* **2000**, 46 (4), 1350–1356.
57. H.R. Hasan, K. Salah. Combating Deepfake Videos Using Blockchain and Smart Contracts. *IEEE Access* **2019**, 7, 41596–41606.