DETECTION OF DEEPFAKE CONTENTS USING DEEP LEARNING TECHNIQUES

K Akhila Krishnan[a] , S Nikilesh[b] , B Kiruba [c] , B Kannapiran[d]

[a,b,c] *Student, Department of Electronics and Communication Engineering, Dr. Mahalingam College of Engineering and Technology, Pollachi-642003*

[d] *Professor, Department of Electronics and Communication Engineering, Dr. Mahalingam College of Engineering and Technology, pollachi-642003*

Abstract*: Detecting Deepfake videos is essential to uphold trust and security in today's digital landscape, where manipulated media is widespread. It is crucial for protecting against the adverse effects of altered media and upholding genuineness, and security in the digital sphere. Deep learning techniques have gained considerable attention for their effectiveness in detecting Deepfakes, leveraging their ability to extract intricate features, handle large datasets, and adapt to evolving threats and various manipulations. This paper specifically targets Deepfake detection in video data using the Viola-Jones Algorithm for face detection. The dataset consists of 600 video samples, partitioned into training and testing sets for model development and evaluation. The proposed model combines Random Forest, face detection, and feature extraction methods to differentiate between authentic and manipulated images. Initially, the faces are identified and isolated from the videos, followed by preprocessing to extract essential features for classification. These facial images undergo feature extraction utilizing ResNet-50, with a focus on the fc1000 layer. The extracted features set is utilized for classification. A Random Forest classifier is then trained on annotated data to distinguish between real and fake videos based on the extracted ResNet-50 features. Performance assessment involves various metrics such as prediction accuracy, precision, and recall. Notably, the Random Forest model achieves a remarkable classification accuracy of 98.60%.*

Keywords: ***Deep learning, Face swapping, Deepfakes, Classification, Feature Extraction, Random Forest.***

1. INTRODUCTION

The rapid evolution of image, video, and audio manipulation technologies has spurred the rise of deepfake techniques. These methods facilitate the creation of highly realistic digital content using minimal resources and easily accessible online tutorials. Deepfakes typically involve replacing atarget individual's face in a video with another person's face, achieved by synthesizing and seamlessly blending facial regions into the original footage. They represent an advanced fusion of deep learning and artificial intelligence, leveraging existing photos and videos to produce misleadingly realistic digital content. Despite their applications in CGI, VR, AR, education, and entertainment, deepfakes pose significant risks, including the spread of false information, malicious hoaxes, and financial scams. Particularly concerning is their exploitation for manipulating public perception of political figures by altering their appearance and speech patterns. Beyond simply generating deceptive content, deepfakes raise serious privacy concerns and can disrupt political landscapes. Their dissemination may sow discord, erode trust, and manipulate public opinion, with potentially far-reaching consequences. Moreover, the proliferation of fabricated media threatens individuals' reputations and subjects them to personal exploitation. Numerous strategies have been put forward in the literature for Deepfakes detection using Random Forest classifier.

Santosh Kolagati, Thenuga Priyadharshini and V. Mary Anita Rajam[1] introduced a novel deep learning approach for fake video detection, integrating Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNN). The model underwent training for 40 epochs employing an Adam learning rate scheduler. Unlike traditional methods relying on feature extraction, this approach directly employs facial data.

Bcombining MLP's ability to discern inconsistencies with CNN's potent feature extraction capabilities, the proposed method achieved notable results. Despite training on a limited subset of data, it attained an accuracy of 84% and an AUC score of 0.87.

Deressa Wodajo and Solomon Atnafu[2] proposed a novel approach for Deepfake detection in their work. The method entails the utilization of a generalized Convolutional Vision Transformer (CViT) architecture, which combines Convolutional Neural Networks with Transformer architecture. Through training on a diverse array of face images encompassing various settings, environments, and orientations, the model effectively learns both local and global image features. This learning process leverages the attention mechanism inherent in the Transformer architecture. Subsequent testing of the model on a dataset comprising 400 DFDC videos yielded results with an accuracy rate of 91.5%, an AUC value of 0.91, and a loss value of 0.32.

Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo [3] proposed a technique leveraging Convolutional Neural Networks (CNNs) trained to identify potential motion discrepancies within the temporal structure of video sequences, utilizing optical flow fields. These disparities in optical flow fields serve as discriminative features for distinguishing between Deepfake videos and authentic ones, employing CNNs. The optical flow fields are integral components of the video sequences, capturing object movements within them, thus bypassing alterations introduced by Deepfake techniques to the visual content of individual frames. The experimental evaluations of the proposed methodology encompassed two distinct operational scenarios: same-forgery and cross-forgery. Notably, the method demonstrated superior performance and efficacy in the cross forgery scenario.

Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun[4] introduced Adaptive Manipulation Traces Extraction Network (AMTEN) as a pre-processing step to reduce image content while highlighting manipulation traces. AMTEN integrates with a Convolutional Neural Network (CNN), using an adaptive convolution layer to anticipate and enhance manipulation artifacts within images. These predicted traces are then utilized in subsequent layers to enhance manipulation artifacts, achieved through weight updates during the backpropagation process. Experimental results shows that AMTENnet achieves 98.52% accuracy in detecting synthetic face images and 95.17% accuracy in identifying faces with unknown post-processing. Notably, on the HFF dataset, AMTENnet improves average detection accuracy by approximately 7.61% compared to MISLnet.

Xu Chang, Jian Wu, Tongfeng Yang, and Guorui Feng[5] introduced the NA-VGG network, a DeepFake image detection system leveraging image noise and augmentation techniques. This approach utilizes the SRM filter within image forensics to extract local noise features from RGB images, subsequently employing this noise distribution data as input for the network. By harnessing the noise features, NA-VGG enhances its ability to classify image authenticity. The experimental evaluations on the Celeb-DF dataset revealed that integrating the SRM filter results in a notable 16.8% increase in image noise compared to the 5 baseline VGG16 network, with a further 12.5% enhancement through image augmentation. These findings underscore the effectiveness of the SRM filter in accentuating crucial image noise features and the significance of image augmentation in enhancing detection accuracy

Priti Yadav, Ishani Jaswal, Jaiprakash Maravi, Vibhash Choudhary, and Gargi Khanna[6] proposed a deepfake detection model employing a pretrained InceptionResNetV2 CNN for feature extraction and vector formation, complemented by RNN and LSTM architectures. InceptionResNetV2 extracts features from each frame during preprocessing, resulting in 2048-dimensional feature vectors post-pooling. The LSTM processes frames sequentially, facilitating comparison of frame features across different time points. The training the LSTM layer with these features yields a confusion matrix used to evaluate validation and testing accuracy. The model achieved accuracies of 84.75% and 91.48% for 20 and 40 epochs, respectively.

Adil Mohammed Parayil, Ameen Masood V, Muhammed Ajas P, Tharun R, and Usha K[7] proposed a deep learning-based approach for detecting deepfakes, combining Convolutional Neural Network (CNN)

and Recurrent Neural Network (RNN) architectures. They employed the Xception network within the CNN component to capture spatial features, while utilizing ytLSTM in the RNN for identifying temporal inconsistencies between frames. The algorithm leverages Xception CNN to extract frame-level features, utilizing 2048-dimensional feature vectors post-pooling as input for sequential LSTM processing. LSTM sequentially analyzes frames to perform temporal video analysis. The models were trained on three standard datasets and validated against them, achieving deepfake predictions with minimal computational time and satisfactory accuracy.

Balaji, Suganthi ST, Mohamed Uvaze Ahamed Ayoobkhan, Krishna Kumar V, Nebojsa Bacanin, Venkatachalam K, Hubálovský Štěpán, and Trojovský Pavel[8] proposed a deep learning-based approach named Fisherface using Local Binary Pattern Histogram (FF-LBPH) for detecting deepfake face images. The method employs the Deep Belief Network (DBN) classifier to distinguish between fake and authentic images. The Fisherface, coupled with LBPH, reduces face dimensionality, aiding in face recognition. The FF-LBPH DBN model achieved an accuracy rate of 98.82% on the CASIA-Web Face image dataset and 97.82% on the DFFD dataset when analyzing deepfake face image manipulations. The technique demonstrates results in accurately detecting manipulated facial images through its innovative fusion of Fisherface and DBN algorithms.

Xinyi Ding, Zohreh Raziei, Eric C. Larson, Eli V. Olinick, Paul Krueger, and Michael Hahsler[9] proposed a deep learning model utilizing transfer learning to detect swapped faces. The approach employs deep transfer learning specifically for detecting face swapping, achieving true positive rates exceeding 96% while maintaining very few false alarms. The method stands out from existing techniques, which typically only offer detection accuracy, by also providing uncertainty estimates for each prediction. The inclusion of uncertainty measures enhances the model's interpretability and aids in understanding the confidence level associated with each detection. The integration of transfer learning techniques has enabled the model in demonstrating robust performance in identifying manipulated faces.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu[10] proposed a method using multi-attention heads to predict spatial attention maps based on deep semantic features. The innovative architecture captures distinct features from different facial regions effectively. The introduction of regional independence loss for network training, aiding in attention-guided data augmentation and adversarial learning. Integration of low-level textural and high-level semantic features is guided by attention maps. These techniques facilitate training of disentangled multiple attentions. Empirical results show significant enhancements across metrics, highlighting the method's efficacy.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo[11] introduced Identity-Driven DeepFake Detection, an approach shifting from artifact-based to identity-based detection. By comparing suspect images/videos with target identities, it reframes detection as identity verification. The Vox-DeepFake dataset includes multiple reference images per suspect content. They proposed the OuterFace algorithm, focusing on outer face regions to learn identity embeddings, achieving robust accuracy without fake sample training. Impressively, a model trained solely on VggFace2 dataset achieved over 96% AUC, demonstrating strong generalization across datasets and resilience to degradations.

Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassne [12] introduced a deep learning method for detecting face swapping and identity alterations. Utilizing separate neural networks, it compares facial region representations with surrounding context to identify disparities. The approach outperforms existing methods on benchmarks like FaceForensics++, Celeb-DF-v2, and DFDC by exploiting the tendency of modern face manipulations to alter internal facial areas while preserving surrounding context. These unique characteristic complements artifact-driven detection, proving effective against previously unseen manipulation techniques. The empirical evaluations confirm its superior performance and detecting fakes even from unfamiliar methods, achieving state-of-the-art results.

2. SOFTWARE DESCRIPTION

The programme used is MATLAB R2023a. MathWorks developed MATLAB, a proprietary multiparadigm programming language and quantitative computation environment. The Viola Jones algorithm does pre-processing. The dataset's features are extracted using the pre-trained Resnet-50 model. The Random Forest model is expressed as coding in a M file in MATLAB 2023. The test function can also be examined using the help command. The Deep Learning Toolbox and the Computer Vision Toolbox were used to create the models.

3. PROPOSED APPROACH FOR DEEPFAKE DETECTION

The proposed system aims to detect face swapping in videos via three phases: face detection/extraction, feature extraction, and classification. Viola-Jones Algorithm detects faces, addressing inconsistencies between faces and backgrounds. Feature extraction employs ResNet-50, converting images to feature vectors via the fc1000 layer. Quality hinges on the chosen extractor. The features are then feed into a Random Forest classifier, trained on annotated data, evaluated for performance metrics like accuracy and F1-score. The classifier distinguishes between authentic and manipulated images, allowing testing for authenticity. The Block diagram for Deepfake detection is shown in Figure 1.

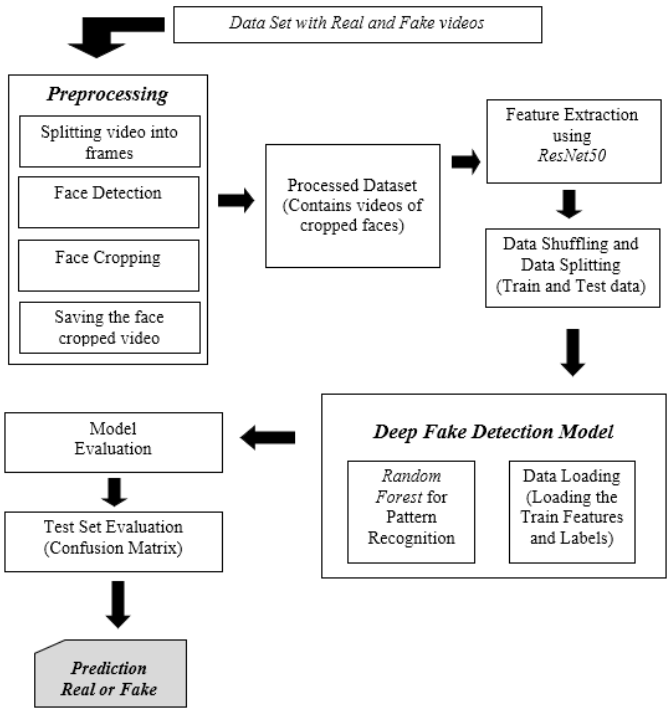


Figure.1 Block Diagram of Deepfake Detection

3.1 Viola jones algorithm

The Viola-Jones algorithm achieves exceptional efficiency through its use of integral image representation and cascade classifier architecture. This representation allows for swift calculation of Haar-like features by pre-computing pixel intensity sums over rectangular regions, reducing computational load during feature assessment. The cascade classifier employs multiple stages, each containing a set of weak classifiers. These simple decision-based classifiers, applied to Haar-like features, create a robust detection system. By

progressively employing more sophisticated classifiers on regions passing earlier stages, the cascade classifier effectively filters out non-object regions, enhancing computational efficacy. Notably, the algorithm's adaptability extends beyond face detection, proving versatile in detecting various objects with precision and speed when trained on relevant datasets. The algorithm extracts images from the video at a rate of 30 frames per second.

3.2 RESNET-50

ResNet-50's adaptable architecture scales to ResNet-101 and ResNet-152, accommodating deeper layers for varying task complexities while maintaining efficiency. It employs residual connections, learning residual functions to prevent gradient vanishing. They are divided into convolutional, identity, convolutional, and fully connected layers. It extracts image features and fully connected layers. It extracts image features, processes them through identity and convolutional blocks, and classifies them. Convolutional layers use batch normalization and ReLU activation, followed by max pooling to preserve crucial features while reducing spatial dimensions. Identity blocks pass input through convolutional layers, aiding residual function learning. Convolutional blocks add 1x1 convolution to reduce filters before 3x3 convolution. Fully connected layers produce final classifications via softmax activation. The architecture of ResNet is shown in figure 2. The convolutional layers extract 200 features per frame from the preprocessed videos.

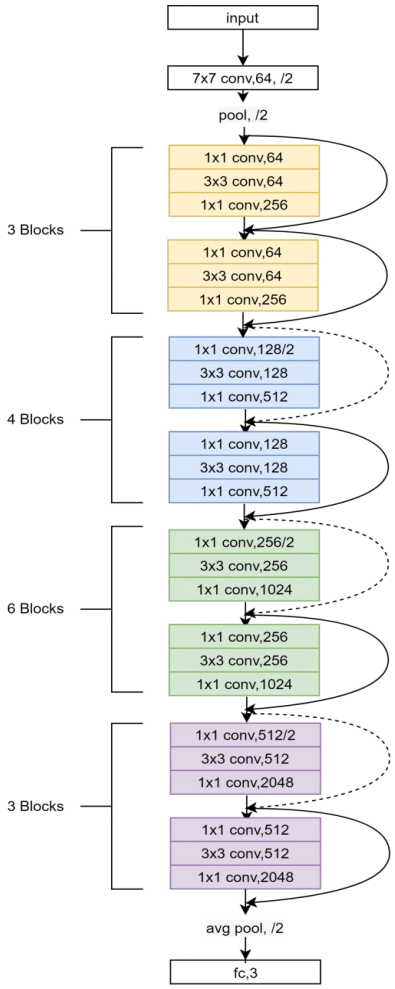


Figure.2 ResNet Architecture

3.3 skip connections

In ResNet-50, skip connections counter vanishing gradients by forming shortcuts between layers, summing input with convolutional output. This enables learning residual mappings, refining details rather than rebuilding from scratch, fostering deeper, accurate representations. Implicit in ResNet-50, skip connections add input to convolutional output in each residual block during feature extraction. Though not explicitly stated in the code, ResNet-50 inherently employs skip connections, vital for addressing training degradation in deep networks, enhancing feature learning. They aid gradient flow, expedite convergence, and boost model performance, pivotal for ResNet-50's effectiveness.

3.4 activation functions

ResNet-50 employs Rectified Linear Unit (ReLU) as its activation function for feature extraction, known for its efficiency and effectiveness in deep learning. It introduces non-linearity, crucial for capturing intricate data patterns.

$$f(x) = \max(0, x) \qquad (1)$$

It outputs positive values directly and zeroes for negatives, promoting sparse activation, reducing overfitting, and enhancing computational efficiency. ReLU's computational efficiency and non-linearity make it ideal for large-scale tasks, mitigate gradient vanishing issues, and encourage sparse activation, improving generalization, crucial for feature extraction and classification in complex datasets.

3.5 random forest

Random Forest Algorithm, popular for Classification and Regression, comprises multiple decision trees enhancing robustness. More trees lead to higher accuracy. It employs ensemble learning, combining classifiers to solve complex problems and enhance performance. Random subsets of features are sampled at each split, reducing overfitting. This randomness ensures each tree captures unique data aspects, improving model robustness and generalization. By decorrelating trees within the ensemble and preventing reliance on specific features, Random Forest achieves better problem-solving ability, making it a favored choice in Machine Learning for its efficiency and effectiveness in handling various tasks.

4. RESULTS AND DISCUSSION

The paper aims to develop a robust system for video-based classification, specifically focusing on distinguishing between real and fake videos. Ultimately, the goal is to contribute to the ongoing efforts in combating misinformation, fake news, and fraudulent content by providing an effective means of detecting and flagging potentially deceptive videos across various online platforms. By leveraging machine learning techniques, particularly the Random Forest classifier trained on features extracted from video frames using deep learning models such as ResNet-50. It aims to create a reliable tool for identifying manipulated or synthetic videos. The training and testing data used in this study was taken from Celeb-DF Dataset. The collected dataset has two folders. One folder with the real videos and the other with the digitally manipulated videos. The separated folders are used to label the features from the videos after feature extraction. The dataset has a total of 800 videos with 400 under each label. After feature extraction using ResNet-50, the extracted features are shuffled and split for training, validation and testing. This includes 70% for training, 15% for validation and 15% for testing. “ReLu” activation function is used by ResNet-50 for feature extraction. Random Forest is used for the prediction. The tree count varies between 50,100 and 150. The minimum leaf size varies between 1,5 and 10.The Random Forest classifier achieved outstanding performance during training, exhibiting a training

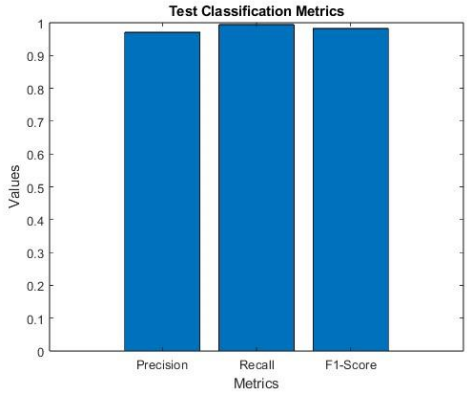
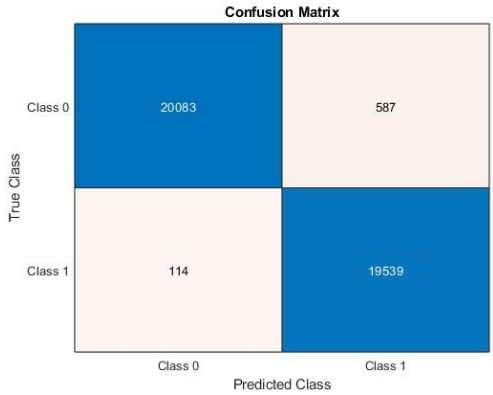


Figure 3 Classification Metrics

accuracy of 100%. This indicates that the model effectively learned the underlying patterns and features present in the training data. Furthermore, the validation accuracy of 98.50% suggests that the model generalized well to unseen data, demonstrating its robustness and ability to make accurate predictions on new samples.

Upon evaluation on the testing set, the Random Forest classifier maintained high precision, recall, and F1-score values, with precision at 0.95, recall at 1.00, and



an F1-score of 0.97. It is shown in figure 3. The high-test accuracy of 98.60% further underscores the efficacy of the developed model in accurately

Figure 4 Confusion Matrix

classifying videos as real or fake. This performance on the testing set confirms the reliability and generalization capabilities of the Random Forest classifier, validating its suitability for real world applications.

The matrix on the figure 4 shows that out of a total of 21,220 instances, 20,083 were correctly classified as real videos (true positives), while 587 were incorrectly classified as fake videos (false positives). Additionally, 114 instances were incorrectly classified as real videos (false negatives), while 19,539 were correctly classified as fake videos (true negatives). This matrix allows for a granular understanding of the model's performance, highlighting its ability to accurately distinguish between real and fake videos, while also shedding light on areas where misclassifications occur.

Table.1 Accuracy Comparison

No of Videos trained	*Training Accuracy*	*Validation Accuracy*	*Testing Accuracy*
150	*100%*	*98.35%*	*98.31%*
300	*100%*	*98.39%*	*98.46%*
450	*100%*	*98.56%*	*98.47%*
600	*100%*	*98.50%*	*98.60%*

As the number of training videos rose to 150, 300, 450, and 600, the training accuracy remained continuously high at 100%. Validation accuracy rose gradually as training data increased, reaching 98.35%, 98.39%, 98.56%, and 98.50% for the corresponding studies. This implies that the model's performance increased when it was exposed to a wider range of training instances. It is shown in Table 1. Overall, the findings show that the suggested technique is excellent at recognising deepfake material, with high training, validation, and testing accuracies over a range of dataset sizes.

5. REFERENCE

1. Santosh Kolagati, Thenuga Priyadharshini, V. Mary Anita Rajam, "Exposing deepfakes using a deep multilayer perceptron – convolutional neural network model", International Journal of Information Management Data Insights, Volume 2, Issue 1,2022,100054,ISSN 2667- 0968,https://doi.org/10.1016/j.jjimei.2021.100054.

2. Deressa,Wodajo , Solomon Atnafu "Deepfake Video Detection Using Convolutional Vision Transformer" arXiv preprint arXiv:2102.11126 (2021).

3. Roberto Caldelli, Leonardo Galteri, Irene Amerini, Alberto Del Bimbo, "Optical Flow based CNN for detection of unlearnt deepfake manipulations", Pattern Recognition Letters, Volume 146,2021,Pages 31-37,ISSN 0167-8655,https://doi.org/10.1016/j.patrec.2021.03.005.

4. Zhiqing Guo, Gaobo Yang, Jiyou Chen, Xingming Sun, "Fake face detection via adaptive manipulation traces extraction network", Computer Vision and Image Understanding, Volume 204,2021,103170,ISSN10773142,https://doi.org/10.1016/j.cviu.2021.103170.

5. X. Chang, J. Wu, T. Yang and G. Feng, "DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network," 2020 39th Chinese Control Conference (CCC), Shenyang, China, 2020, pp. 7252-7256, doi: 10.23919/CCC50068.2020.9189596.

6. Priti Yadav, Ishani Jaswal, Jaiprakash Maravi, Vibhash Choudhary ,Gargi Khanna" DeepFake Detection using InceptionResNetV2 and LSTM", International Conference on Emerging Technologies: s: AI, IoT, and CPS for Science Technology Applications, Volume 3058,2021.

7. Adil Mohammed Parayil, Ameen Masood V, Muhammed Ajas P, Tharun R, Usha K" Deepfake Detection using Xception and LSTM" ,International Research Journal of Modernization in Engineering Technology and Science,2023.

8. Balaji, Suganthi & Ayoobkhan, Mohamed Uvaze Ahamed & Kumar, V. & Bacanin, Nebojsa & Venkatachalam, Kv & Štěpán, Hubálovský & Pavel, Trojovský. (2022)." Deep learning model for deep fake face recognition and detection". PeerJ Computer Science. 8. e881. 10.7717/peerj-cs.881.

9. Ding X, Raziei Z, Larson EC, Olinick EV, Krueger P, Hahsler M. "Swapped face detection using deep learning and subjective assessment". EURASIP Journal on Information Security. 2020 Dec;2020:1-2. 10.

10.H. Zhao, et al., "Multi-attentional Deepfake Detection," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021 pp. 2185-2194. doi: 10.1109/CVPR46437.2021.00222.

11.Dong X, Bao J, Chen D, Zhang W, Yu N, Chen D, Wen F, Guo B." Identity-driven deepfake detection". arXiv preprint arXiv:2012.03930. 2020 Dec 7.

12.Y. Nirkin, L. Wolf, Y. Keller and T. Hassner, "DeepFake Detection Based on Discrepancies Between Faces and Their Context," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, pp. 6111-6121, 1 Oct. 2022, doi: 10.1109/TPAMI.2021.3093446.

13.M. Patel, A. Gupta, S. Tanwar and M. S. Obaidat, "Trans-DF: A Transfer Learningbased end-to-end Deepfake Detector," 2020 IEEE 5th International Conference on Computing Communication and automation (ICCCA), Greater Noida, India, 2020, pp. 796-801, doi: 10.1109/ICCCA49541.2020.9250803.