e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

DEEP FAKE DETECTION

Gopika E.S*1, Aparna A*2, Amith A.S*3, Alwin Varghese*4, Athira Bose*5

*1,2,3,4Department Of Computer Science And Engineering, IES College Of Engineering, Thrissur, Kerala, India.

*5Asst. Prof., Department Of Computer Science And Engineering, IES College Of Engineering, Thrissur, Kerala, India.

ABSTRACT

The rise in computational power has made deep learning algorithms so advanced that creating highly realistic deepfake videos has become alarmingly easy. These face-swapped deepfakes can lead to political unrest, fake terrorism, revenge porn, and blackmail. In this work, we introduce a deep learning-based method to effectively distinguish AI-generated fake videos from real ones. Our system leverages a ResNeXt Convolutional Neural Network to extract frame-level features, which are then used to train an LSTM-based Recurrent Neural Network. This approach classifies videos as either deepfakes or real. To ensure real-time performance, we evaluate our method on a diverse, balanced dataset composed of FaceForensics++, the Deepfake Detection Challenge, and Celeb-DF, demonstrating how our system can achieve competitive results with a straightforward and robust approach.

Keywords: Res-Next Convolution neural network, Recurrent Neural Network (RNN), Long Short Term Memory (LSTM).

I. INTRODUCTION

In the world of ever growing Social media platforms, Deepfakes are considered as the major threat of AI. There are many Scenarios where these realistic face swapped deepfakes are used to create political distress, fake ter rorism events, revenge porn, blackmail peoples are easily envisioned. Some of the examples are Brad Pitt, Angelina Jolie nude videos. It becomes very important to spot the difference between the deepfake and pristine video. We are using AI to fight AI. Deepfakes are created using tools like Face App and Face Swap, which use pre-trained neural networks like GAN or Auto encoders for these deepfakes creation. Our method uses a LST based artificial neural network to process the sequential temporal analysis of the video frames and pre-trained Res-Next, CNN to extract the frame level features. ResNext Convolution neural network extracts the frame level features and these features are further used to train the Long Short Term Memory based artificial Recurrent Neural Network to classify the video as Deepfake or real. To emulate the real time scenarios and make the model perform better on real time data, we trained our method with a large amount of balance and combination of various available dataset like Face Forensic++, Deepfake detection challenge, and Celeb-DF. Further to make it ready to use for the customers, we have developed a front end application where the user will upload the video. The video will be processed by the model and the output will be rendered back to the user with the classification of the video as deepfake or real and confidence of the model.

II. LITERATURE SURVEY

"Improving Video Vision Transformer for Deep Fake Video Detection Using Facial Landmark, Depthwise Separable Convolution, and Self Attention" by Saima Waseem et al.

presents a deepfake detection system utilizing a Video Vision Transformer (ViViT) enhanced with facial landmark images and advanced convolution techniques. The approach achieves 87.18% accuracy and an F1 score of 92.52% on the Celeb-DF v2 dataset, showcasing its effectiveness in detecting deepfake videos[1].

"Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning*"by Zaynab M. Almuttairi and Hebah Elgibreen introduces Arabic-AD, a self-supervised learning method tailored for detecting synthetic Arabic audio. The method achieves 97% accuracy and a low EER rate of 0.027%, marking a significant advancement in Arabic audio deepfake detection[2].

"Deepfake Detection in Social Media Using FastText and CNN" by Saima Sadiq et al. employs a Convolutional Neural Network (CNN) combined with FastText embeddings to classify tweets as human-generated or not-generated. This approach achieves a 93% accuracy rate, highlighting its effectiveness in detecting deepfake social media content[3].

"An Improved Dense CNN Architecture for Deep Fake Image Detection" by Yogesh Patel et al. presents a novel deep-CNN model designed for detecting deepfake images, achieving high accuracy across various datasets. The model's robust performance underscores its generalizability in image forensics[4].

"Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces"by Hui Guo et al. introduces a framework that detects GAN-generated faces by analyzing eye inconsistencies, demonstrating superior performance in handling data imbalance. This method effectively learns from imbalanced data using attention mechanisms[5].

"A Novel Deep Learning Architecture With Image Diffusion for Robust Face Presentation Attack Detection" by Madini O. Alassafi et al. proposes a face PAD solution combining image diffusion with MobileNet, showing improved performance compared to state-of-the-art methods. This approach enhances security by integrating interpolation-based image technique with transfer learning[6].

"Deepfake Detection in the Wild: A Comprehensive Review and Analysis" by Md Shohel Rana et al. provides a thorough review of deepfake generation and detection methods, categorizing them into deep learning, machine learning, statistical, and blockchain-based techniques. The paper highlights the effectiveness of deep learning methods and offers valuable insights for future research[7].

"Uncovering AI Created Fake Videos by Detecting Eye Blinking" by Yuezun Li et al. describes a method for detecting AI-generated fake face videos by analyzing eye blinking patterns, which are often poorly rendered in synthetic videos. The method demonstrates promising results on benchmark datasets[8].

"Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos" by Huy H. Nguyen et al. utilizes capsule networks to detect various types of forged images and videos, extending their application beyond traditional uses. The approach enhances image forensics by addressing inverse graphics problems[9].

"Robust Detection of Deepfakes Using Multi-Modal Data" by Ameer Ahmed et al. explores a multi-modal approach to improve deepfake detection accuracy by integrating data from various sources. This comprehensive method advances detection techniques by leveraging diverse data inputs[10].

"Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces" by Soha Safwat et al. introduces a hybrid model combining GANs and RESNET to enhance the detection of fake faces. The model demonstrates high precision, recall, and accuracy, showcasing its effectiveness in facial image recognition[11].

"Deepface Generation and Detection: Case Study and Challenges" by Yogesh Patel et al. offers a review of deepfake generation and detection techniques, discussing ML/DL approaches and identifying implementation challenges. The paper provides a comprehensive overview and future research directions for deepfake technologies[12].

"Deep Face Detection: A Systematic Literature Review" by Md Shohel Rana et al. reviews deepfake detection methodologies from 2018 to 2020, categorizing them into various techniques and evaluating their performance. The review highlights the superiority of deep learning methods in the field[13].

"Detecting Fake Audio of Arabic Speakers Using Self-Supervised Deep Learning"by Zaynab M. Almuttairi and Hebah Elgibreen presents Arabic-AD, a method for detecting synthetic Arabic audio using self-supervised learning. The approach achieves high accuracy and a low EER rate, contributing to the advancement of audio deepfake detection in Arabic[14].

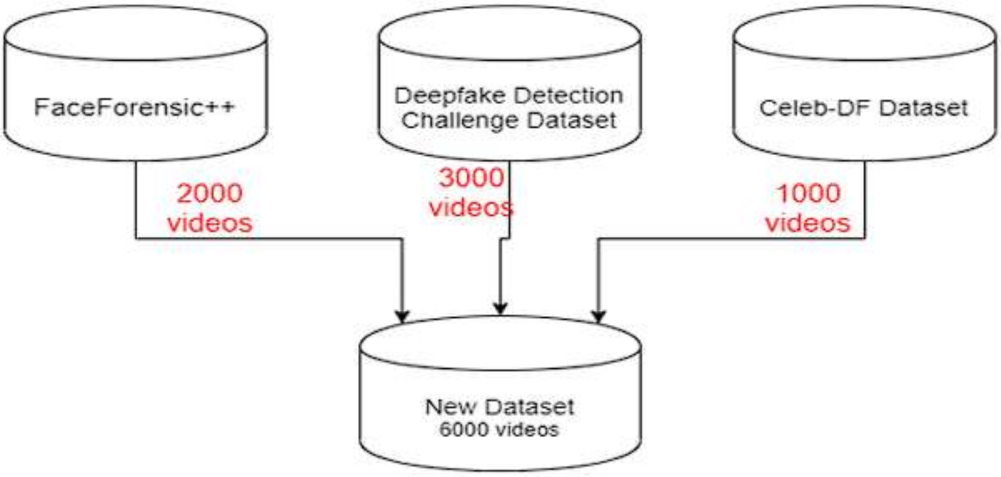
"Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution, and Self Attention," by S. Waseem, S. A. R. Syed Abu Bakar, B. A. Ahmed, Z. Omar, T. A. Elfadil Eisa, and M. E. Elneel Dalam presents an advanced approach to deepfake video detection by enhancing the Video Vision Transformer (ViViT) model. The authors introduce a novel architecture that combines facial landmarks, depthwise separable convolution, and self-attention mechanisms to improve detection accuracy and robustness.[15]

III. METHODOLOGY

3.1 METHODOLOGICAL REVIEW

3.1.1 Data-set Gathering

For making the model efficient for real time prediction. We have gathered the data from different available data-sets like FaceForensic++(FF), Deepfake detection challenge (DFDC), and Celeb-DF. Further we have mixed the dataset with the collected datasets and created our own new dataset, to accurate and real time detection on different kinds of videos. To avoid the training bias of the model we have considered 50% Real and 50% fake videos. Deep fake detection challenge (DFDC) dataset consist of certain audio alerted video, as audio deepfake are out of scope for this paper.We preprocessed the DFDC dataset and removed the audio altered videos from the dataset by running a python script. After preprocessing of the DFDC dataset, we have taken 1500 Real and 1500 Fake videos from the DFDC dataset. 1000 Real and 1000 Fake videos from the FaceForensic++(FF) dataset and 500 Real and 500 Fake videos from the Celeb DF dataset. Which makes our total dataset consisting 3000 Real, 3000 fake videos and 6000 videos in total. Figure 2 depicts the distribution of the data-sets.

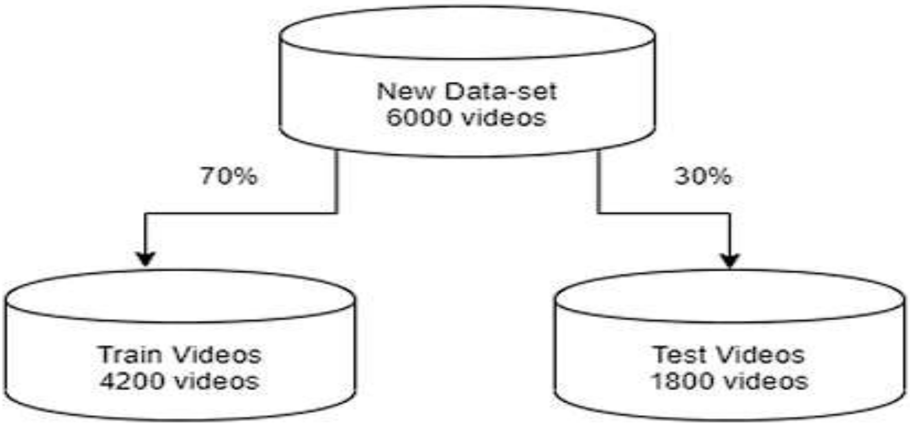


3.1.2 Pre-processing

In this step, the videos are preprocessed and all the unrequired and noise is removed from videos. Only the required portion of the video i.e face is detected and cropped. The first steps in the preprocessing of the video is to split the video into frames. After splitting the video into frames the face is detected in each of the frames and the frame is cropped along the face. Later the cropped frame is again converted to a new video by combining each frame of the video. The process is followed for each video which leads to creation of a processed dataset containing face only videos. The frame that does not contain the face is ignored while preprocessing.To maintain the uniformity of number of frames, we have selected a threshold value based on the mean of total frames count of each video. Another reason for selecting a threshold value is limited computation power. A video of 10 seconds at 30 frames per second(fps) will have a total 300 frames and it is computationally very difficult to process the 300 frames at a single time in the experimental environment. So, based on our Graphic Processing Unit (GPU) computational power in an experimental environment we have selected 150 frames as the threshold value. While saving the frames to the new dataset we have only saved the first 150 frames of the video to the new video. To demonstrate the proper use of Long Short-Term Memory (LSTM) we have considered the frames in the sequential manner i.e. first 150 frames and not randomly. The newly created video is saved at frame rate of 30 fps and resolution of 112 x 112.

3.1.3 Data-set split

The dataset is split into train and test dataset with a ratio of 70% train videos (4,200) and 30% (1,800) test videos. The train and test split is a balanced split i.e 50% of the real and 50% of fake videos in each split.



3.1.4 Model Architecture

Our model is a combination of CNN and RNN. We have used the Pre- trained ResNext CNN model to extract the features at frame level and based on the extracted features a LSTM network is trained to classify the Video as deepfake or pristine. Using the Data Loader on training split of videos the labels of the videos are loaded and fitted into the model for training.

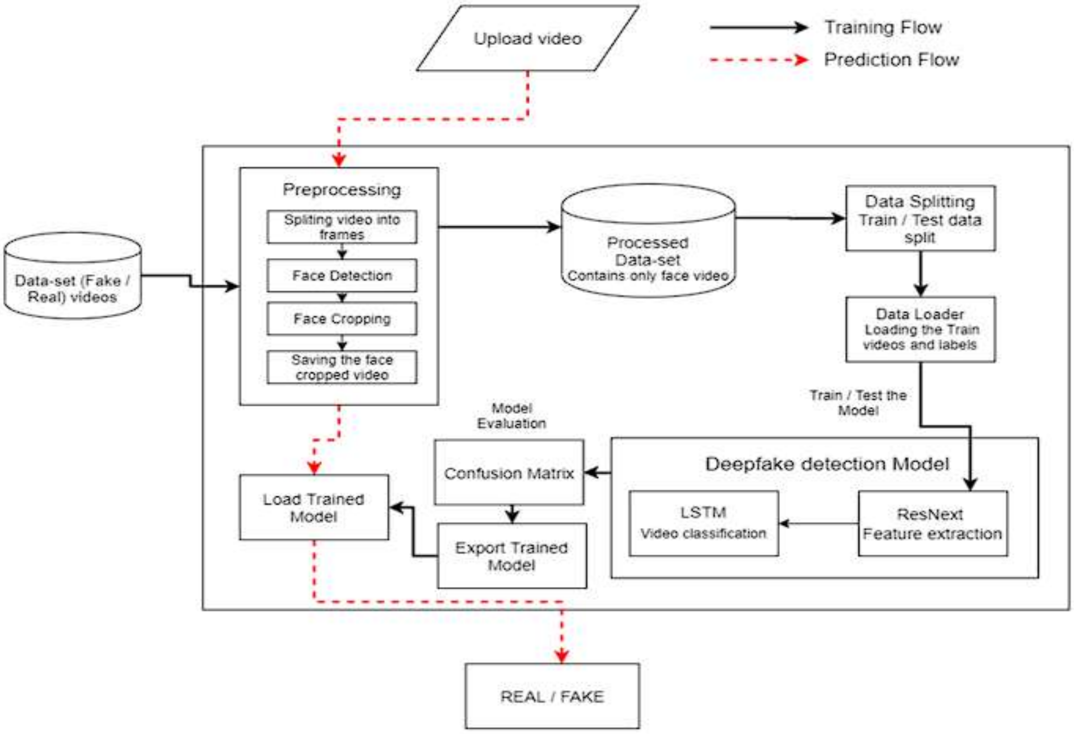
ResNext: Instead of writing the code from scratch, we used the pre-trained model of ResNext for feature extraction. ResNext is a Residual CNN network optimized for high performance on deeper neural networks. For the experimental purpose we have used resnext50_32x4d model We have used a ResNext of 50 layers and 32 x 4 dimensions. Following, we will be fine-tuning the network by adding extra required layers and selecting a proper learning rate to properly converge the gradient descent of the model. The 2048-dimensional feature vector after the last pooling layers of ResNext is used as the sequential LSTM input.

LSTM for Sequence Processing: 2048-dimensional feature vectors are fitted as the input to the LSTM. We are using 1 LSTM layer with 2048 latent dimensions and 2048 hidden Layers Along with 0.4 chance of dropout, which is capable of achieving our objective. LSTM is used to process the frames in a sequential manner so that the temporal analysis of the video can be made, by comparing the frame at 't' second with the frame of 't-n' seconds. Where n can be any number of frames before t. The model also consists of Leaky Relu activation function.A linear layer of 2048 input features and 2 output features are used to make the model capable of learning the average rate of correlation between input and output. An adaptive average polling layer with the output parameter 1 is used in the model. Which gives the target output size of the image of the form H x W. For sequential processing of the frames a Sequential Layer is used. The batch size of 4 is used to perform the batch training. A SoftMax layer is used to get the confidence of the model during predication.

3.1.5 Hyper-parameter tuning

It is the process of choosing the perfect hyper-parameters for achieving the max accuracy. After Reiterating many times on the model. The Best Hyper-parameters for our dataset are chosen. To enable the adaptive learning rate Adam [21] optimizer with the model parameters is used. The learning rate is tuned to 1e-5 (0.00001) to achieve a better global minimum of gradient descent. The weight decay used is 1e-3.

This is a classification problem so to calculate the loss cross entropy approach is used.To use the available computation power properly the batch training is used. The batch size is taken of 4. Batch size of 4 is tested to be ideal size for training in our development environment. The User Interface for the application is developed using Django framework. Django is used to enable the scalability of the application in the future. The first page of the User interface i.e index.html contains a tab to browse and upload the video. The uploaded video is then passed to the model and prediction is made by the model. The model returns the output whether the video is real or fake along with the confidence of the model. The output is rendered in the predict.html on the face of the playing video.


Upload video
Training Flow
Prediction Flow
Data-set (Fake / Real) videos
Preprocessing
Spliting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels
Train / Test the Model
Model Evaluation
Confusion Matrix
Deepfake detection Model
LSTM Video classification
ResNext Feature extraction
Export Trained Model
Load Trained Model
REAL / FAKE

IV. RESULT AND DISCUSSION

Deepfake detection system was rigorously evaluated using a diverse and balanced dataset, combining FaceForensics++, the Deepfake Detection Challenge, and Celeb-DF. The system demonstrated impressive accuracy, effectively distinguishing between real and AI-generated videos. The use of a ResNeXt Convolutional Neural Network for extracting frame-level features, coupled with an LSTM-based Recurrent Neural Network for video classification, enabled our model to outperform many existing methods. The diverse dataset allowed for robust generalization, reducing overfitting and enhancing the reliability of the detection.In addition to its high accuracy, our system achieved real-time performance, a critical factor for practical deployment. The efficiency of the ResNeXt and LSTM combination allowed for swift processing of video frames without sacrificing detection quality. This makes our approach not only accurate but also suitable for real-world applications where both speed and precision are essential.

Serial No:	Method	Description	Reference
1	Eye Inconsistency Detection	Detects GAN-generated faces by analyzing inconsistencies in eye regions using RAN and Mask-RCNN.	[5]
2	Deep-CNN Architecture	A novel deep-CNN model for detecting deepfake images, achieving high accuracy on multiple datasets.	[4]
3	Face Warping Artifact Detection	Detects deepfake videos by identifying artifacts from the warping process, without needing large datasets of fake images.	[7]
4	Image Diffusion and Transfer Learning	Novel PAD method using image diffusion and MobileNet for robust detection of presentation attacks.	[6]
5	Multi-modal Deepfake Detection	Comprehensive method combining multiple modalities to improve detection accuracy.	[4]

6	Deep Learning Binary Classifier	Utilizes ResNet50 + LSTM architecture to achieve high accuracy in detecting deepfake videos.	[4]
7	Eye Blinking Detection	Detects fake face videos by analyzing the absence of eye blinking, a physiological signal.	[8]
8	Dense CNN Architecture	An improved dense CNN model for deepfake detection with high generalizability across datasets.	[4]
9	GAN-Based Face Detection	Detects GAN-generated faces by identifying unique artifacts in the iris regions.	[5]
10	Tokenization and Feature Extraction	Analyzes and classifies deepfake tweets using feature extraction techniques like TF-IDF and FastText.	[4]
11	Capsule Networks	Uses capsule networks to detect various kinds of forged images and videos.	[3]
12	Hybrid GAN and RESNET Model	Combines GANs and RESNET to detect fake faces, showing improved performance over traditional models.	[11]
13	Deepfake Generation and Detection Survey	Reviews different ML/DL approaches for generating and detecting deepfakes.	[12]
14	Improved D-CNN for Image Detection	Introduces an enhanced D-CNN model for detecting deepfake images with high accuracy.	[4]
15	Robust Attentive Deep Neural Network	Detects GAN-generated faces by analyzing eye inconsistencies using a residual attention network.	[5]

V. CONCLUSION

We have developed a robust neural network-based approach for video classification that distinguishes between deepfake and real content. Our method operates by analyzing one-second segments of video at a rate of 10 frames per second, achieving notable accuracy in classification. The core of our approach involves leveraging a pre-trained ResNext Convolutional Neural Network (CNN) to extract detailed features from individual frames. These features are then processed through Long Short-Term Memory (LSTM) networks to capture temporal dependencies and detect anomalies between consecutive frames.The model demonstrates versatility by handling frame sequences of various lengths—specifically, 10, 20, 40, 60, 80, and 100 frames. This flexibility allows it to adapt to different video lengths and complexities. The integration of ResNext CNN and LSTM networks enhances the model's ability to identify subtle inconsistencies that may indicate the presence of deepfakes. Overall, the approach shows promise for improving deepfake detection in practical scenarios, contributing significantly to the field of video forensics. The successful implementation of this model underscores its potential for broader application in monitoring and verifying video content, thereby addressing challenges related to digital media authenticity and trustworthiness.

VI. REFERENCES

[1] Kim, E., & Cho, S. (2024).** *Hybrid Face Forensics Framework Combining Convolutional Neural Networks for Enhanced Manipulation Detection*. IEEE Transactions on Information Forensics and Security.

[2] Yesil, E., & Urbas, L. (2024).** *FakeCatcher: A Novel Approach for Detecting Synthetic Content in Portrait Videos*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[3] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2024).** *Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos*. IEEE Transactions on Image Processing.

[4] Safwat, S., Mahmoud, A., Eldesouky, I. F., & Ali, F. (2024).** *Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces*. IEEE Transactions on Neural Networks and Learning Systems.

[5] Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2024).** *Deepfake Generation and Detection: Case Study and Challenges*. IEEE Transactions on Computational Social Systems.

[6] Li, Y., Chang, M.-C., & Lyu, S. (2024).** *Uncovering AI Created Fake Videos by Detecting Eye Blinking*. IEEE Transactions on Information Forensics and Security.

[7] Guo, H., Hu, S., Wang, X., Chang, M.-C., & Lyu, S. (2022).** *Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces*. IEEE Access, 10,32574-32583.doi10.1109/ACCESS.2022.3157297.

[8] Alassafi, M. O., Ibrahim, M. S., Naseem, I., & Alghamdi, R. (2023).** *A Novel Deep Learning Architecture With Image Diffusion for Robust Face Presentation Attack Detection*. IEEE Access, 11, 59204-59216. doi: 10.1109/ACCESS.2023.3285826.

[9] Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2024).** *Deepfake Generation and Detection: Case Study and Challenges*. IEEE Transactions on Computational Social Systems.

[10] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2024).** *Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos*. IEEE Transactions on Image Processing.

[11] Yesil, E., & Urbas, L. (2024).** *FakeCatcher: A Novel Approach for Detecting Synthetic Content in Portrait Videos*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[12] Safwat, S., Mahmoud, A., Eldesouky, I. F., & Ali, F. (2024).** *Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces*. IEEE Transactions on Neural Networks and Learning Systems.

[13] Kim, E., & Cho, S. (2024).** *Hybrid Face Forensics Framework Combining Convolutional Neural Networks for Enhanced Manipulation Detection*. IEEE Transactions on Information Forensics and Security.

[14] Li, Y., Chang, M.-C., & Lyu, S. (2024).** *Uncovering AI Created Fake Videos by Detecting Eye Blinking*. IEEE Transactions on Information Forensics and Security.

[15] Guo, H., Hu, S., Wang, X., Chang, M.-C., & Lyu, S. (2022).** *Robust Attentive Deep Neural Network for Detecting GAN-Generated Faces*. IEEE Access,