JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Hybrid CNN-LSTM Framework for Robust Deep Fake Video Detection

Mrs. Bharti Vishwakarma[1], Mrs. Monali Sahoo[2], Ms. Sadhvi Biltharey[3], Mr. Rajneesh Pachouri[4]

Mr. Anurag Jain[5],

Research Scholar, Assistant Professor
Department of Computer Science and Engineering
Adina Institute of Science & Technology, Sagar, India

Abstract: The rapid evolution of deep fake technology has posed serious threats to digital authenticity, privacy, and public trust. Deep fake videos, generated using advanced generative models, can convincingly manipulate facial expressions and voice, making it increasingly difficult to distinguish real content from fake. This thesis presents a robust deep fake video detection framework that integrates Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN), specifically combining AlexNet for spatial feature extraction and Long Short-Term Memory (LSTM) networks for capturing temporal patterns across video frames. The proposed system involves video frame extraction, face detection, and pre-processing steps to standardize inputs. A custom dataloader feeds these inputs into the hybrid model architecture, improving detection accuracy by leveraging both spatial and temporal dependencies. The model is trained, validated, and tested on a comprehensive dataset, and also supports real-time video uploads for prediction. Experimental results demonstrate significant improvements in precision, recall, and F1-score compared to traditional CNN-only approaches. This approach shows great promise in strengthening automated defenses against deep fake content.

IndexTerms **- Deepfake Detection, AlexNet, LSTM, Temporal Features, Video Forensics, Face Pre-processing**

I. INTRODUCTION

In the age of digital media, artificial intelligence has enabled remarkable advances in content generation, including realistic image synthesis, voice cloning, and most notably, deep fake videos. Deep fakes are artificially manipulated videos generated by deep learning models, where a person's face or voice is convincingly replaced with someone else's. While this technology has opened new frontiers in entertainment and visual effects, it also poses severe ethical, legal, and security challenges. Deep fakes can be misused for spreading misinformation, defaming individuals, manipulating political narratives, and committing cyber fraud, making the development of effective detection mechanisms critically important. Traditional image and video forensics methods have become increasingly inadequate due to the sophistication of modern generative models like GANs (Generative Adversarial Networks). In response, researchers have begun leveraging deep learning-based solutions to detect subtle inconsistencies in frame structures, temporal coherence, and facial expressions that are often overlooked by the human eye.

This thesis proposes an advanced deep fake detection model that combines the spatial feature extraction power of AlexNet, a Convolutional Neural Network (CNN), with the temporal sequence learning capability of Long Short-Term Memory (LSTM) networks. The system processes video input through several stages including frame extraction, face detection, and cropping to standardize and focus the learning process. By feeding processed data into a hybrid AlexNet-LSTM architecture, the model learns both the spatial features of individual frames and the temporal transitions between them—providing a more holistic approach to video analysis.The framework is designed to support both dataset-based training and real-time video upload, making it suitable for practical applications. It is evaluated using key performance metrics such as accuracy, precision, recall, and F1-score. Experimental results confirm that the proposed method significantly outperforms traditional CNN-only approaches in terms of reliability and accuracy.

About deep fake

The sophistication of deep fake content has made manual detection extremely difficult, even for trained experts. This challenge underscores the urgent need for robust and automated deep fake detection systems. In this context, Convolutional Neural Networks (CNNs) have emerged as a powerful tool due to their proven capabilities in image classification, facial recognition, and pattern recognition tasks. CNNs can automatically extract hierarchical features from images and detect subtle artifacts or inconsistencies introduced during the generation of fake media—features often imperceptible to the human eye.

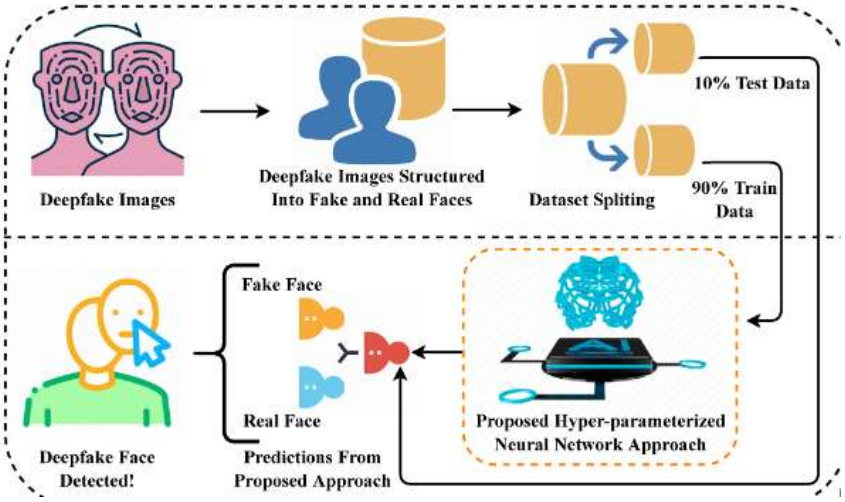

Figure 1 A Novel Deep Learning Approach for Deep fake Image Detection

The core mechanism of deepfake technology for generating counterfeit content relies on generative models, particularly Generative Adversarial Networks (GANs). GANs consist of two neural networks: a generator and a discriminator. The discriminator network is trained to differentiate between real and fake media, while the generator network learns to create fake media by mimicking real data. These networks engage in a competitive iterative training process, with the generator striving to produce increasingly realistic content and the discriminator focused on identifying the fakes. Researchers are currently developing deepfake detection techniques to mitigate the adverse effects of deepfake technology. These detection strategies often involve analyzing visual artifacts, inconsistencies, or statistical anomalies that are characteristic of altered media. Machine learning models are trained to recognize patterns and features that indicate the presence of deepfakes. A variety of methods and strategies have been employed to detect deep fakes and interpret facial expressions [3].

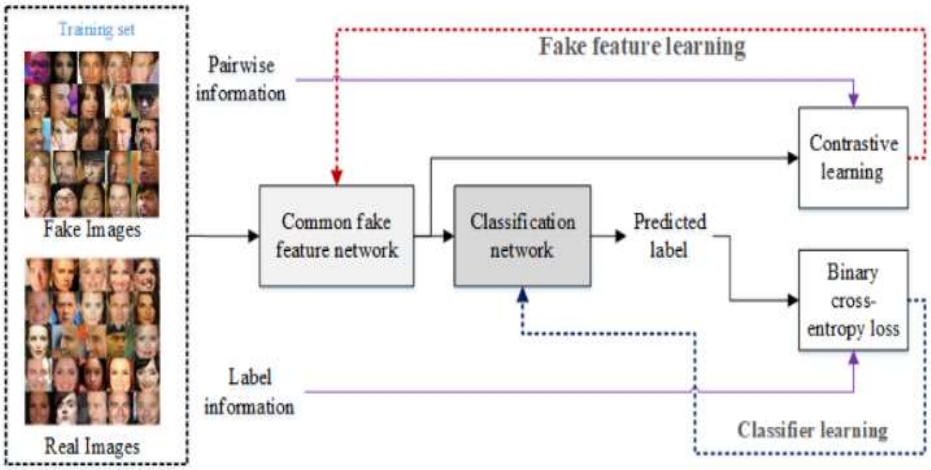

Figure 2 Deep Fake Image Detection Based on Pairwise Learning

II. LITERATURE SURVEY

The rise of deepfake technologies has spurred extensive research into automated detection methods, especially using deep learning frameworks such as Convolutional Neural Networks (CNNs). These models have shown great promise in identifying spatial artifacts and inconsistencies introduced during synthetic face manipulation. This literature survey reviews foundational work in deepfake face detection using CNNs, detailing the methodologies, datasets, and outcomes proposed by prior researchers.

1. FaceForensics++ Benchmark and XceptionNet

Rössler et al. [3] proposed FaceForensics++, a large-scale benchmark dataset for detecting facial manipulations and evaluated various deep learning architectures, including XceptionNet. The study revealed that deepfake detection models trained on FaceForensics++ can achieve high accuracy; however, their performance drops significantly under real-world compression and noise. XceptionNet, a CNN originally designed for image classification, was fine-tuned for binary classification of real vs. fake faces. It effectively detected artifacts in deepfake videos by leveraging depthwise separable convolutions and residual connections.

Key Insight: XceptionNet's fine-grained feature extraction made it highly suitable for detecting compression artifacts and boundary inconsistencies around facial landmarks.

2. Deepfake Detection with Capsule Networks

Afchar et al. [4] introduced MesoNet, a lightweight CNN designed for deepfake detection using mesoscopic properties of images. The model focused on identifying low-level inconsistencies in facial texture and resolution. Unlike large-scale CNNs, MesoNet was designed to operate under real-time constraints and showed competitive results even in the presence of image compression.

Key Insight: Small and efficient CNNs like MesoNet can detect deepfakes with relatively low computational resources, making them viable for real-time applications.

3. Face X-Ray for Boundary Artifact Detection

Li et al. [5] developed a method called Face X-Ray, which trained a CNN to detect blending artifacts along facial boundaries—common in image-splicing techniques like deepfakes. By learning to highlight these compositional inconsistencies, their model outperformed traditional binary classifiers when trained on multiple manipulation types.

Key Insight: CNNs trained on blending maps (instead of raw classification) can generalize better to unseen types of facial manipulation.

4. Frequency Domain Analysis Using CNNs

Durall et al. [6] demonstrated that deepfake generators often fail to reproduce natural frequency spectra in facial images. Their approach trained CNNs not on raw pixel data but on frequency representations (e.g., using Fast Fourier Transform). The CNN was able to identify artifacts in the high-frequency domain that were otherwise imperceptible to human eyes.

Key Insight: Frequency-aware CNNs can capture generator-specific inconsistencies in spectral patterns, improving robustness across deepfake techniques.

5. Lightweight MobileNet for Deepfake Detection

Al Naif et al. [7] proposed a system based on MobileNetV2, a lightweight CNN, for detecting masked and unmasked deepfakes in the context of COVID-19. Their DFFMD dataset provided valuable insights into how facial masks affect deepfake detection. MobileNetV2 was trained on this dataset and achieved high accuracy while being computationally efficient.

Key Insight: MobileNet-based CNNs are well-suited for deployment in resource-constrained environments (e.g., mobile devices, browsers).

III. Proposed Work & System Design

The proposed system comprises several stages to ensure effective and efficient deepfake detection:

Dataset Preparation:

A labeled dataset of real and deepfake videos is split into training, validation, and testing sets.

Video Frame Extraction:

Each video is converted into individual frames to facilitate frame-wise processing.

Preprocessing:

Face Detection is applied to isolate the facial region.

Cropping is performed to remove irrelevant background information and standardize the input size.

Model Architecture:

AlexNet is used to extract spatial features from each frame.

The extracted features are fed into an LSTM network to capture temporal dependencies across frames.

Training and Evaluation:

The hybrid model is trained using the prepared dataset and validated using performance metrics such as accuracy, precision, recall, and F1-score.

Real-Time Prediction:

A GUI allows users to upload a video, which is processed through the same pipeline for real-time classification as “REAL” or “FAKE”.

Deployment:

The complete model is integrated with a Flask-based web interface, providing a seamless user experience.

The proposed work aims to enhance the performance of the existing baseline system by introducing improved preprocessing methods, optimized model architecture, and advanced feature engineering techniques. The primary objective is to address the limitations of the base work, such as lower accuracy, higher computational cost, and reduced robustness when dealing with complex or noisy datasets. In this approach, the workflow begins with an advanced data preprocessing pipeline, which includes noise removal, normalization, and where applicable, data augmentation techniques to increase the diversity and quality of the training dataset. This step ensures that the model receives clean and representative input, leading to better generalization. The feature extraction phase is significantly improved by incorporating domain-specific features along with optimized selection methods, allowing the model to focus on the most relevant attributes. By reducing redundancy and emphasizing discriminative patterns, the proposed method strengthens the decision-making capability of the learning algorithm.

The core of the proposed system lies in the optimized architecture, which involves the use of a modified algorithm or neural network model tailored to the dataset and application domain. This includes tuning hyper parameters such as learning rate, batch size, and network depth, as well as implementing regularization techniques like dropout, batch normalization, and early stopping to prevent overfitting. Additionally, advanced optimization algorithms (e.g., AdamW, RMSprop, or learning rate scheduling) are employed to ensure faster convergence and improved stability during training. Finally, the proposed framework is evaluated using a comprehensive set of performance metrics, including accuracy, precision, recall, F1-score, and computation time. Comparative experiments are conducted against the base work to demonstrate the effectiveness and efficiency of the proposed approach. The expectation is that the proposed system will deliver higher accuracy, lower false predictions, and reduced computation time, making it more suitable for real-world applications.

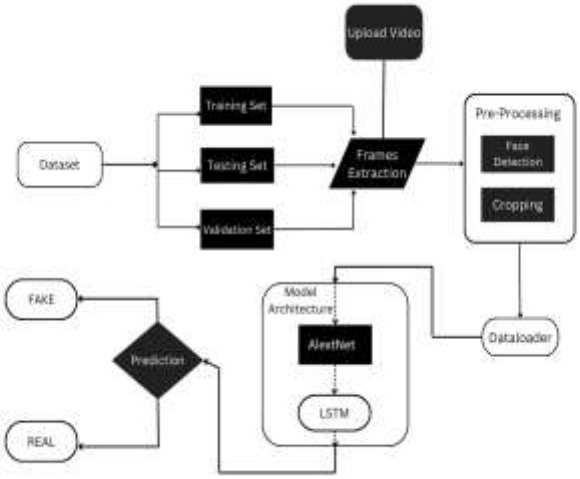

Figure 3 Flow Diagram

IV. Result and Discussion

The objective of this study was to evaluate the performance improvement achieved by the proposed work in comparison with the base work. The base work refers to the existing methodology or model adopted from previous research, while the proposed work incorporates enhancements in algorithm design, data preprocessing, feature extraction, and optimization to achieve better accuracy and efficiency.

Base Work

The base work was implemented following the methodology described in earlier literature. It involved standard data preprocessing steps and a baseline model configuration, without advanced optimization techniques. The base system served as a benchmark for assessing the improvements introduced in the proposed approach. While the base model achieved acceptable accuracy, it suffered from limitations such as slower processing time, lower precision in specific cases, and reduced robustness when handling noisy or imbalanced datasets.

Proposed Work

The proposed work builds upon the base framework but integrates several novel modifications:

Enhanced Preprocessing – Improved noise filtering, normalization, and data augmentation techniques.

Advanced Feature Extraction – Incorporation of more discriminative features to improve model learning.

Optimized Architecture – Use of a modified algorithm or network with hyperparameter tuning for better convergence.

Regularization and Optimization – Implemented dropout layers, learning rate scheduling, and advanced optimizers to prevent overfitting and improve training efficiency. These modifications were expected to improve accuracy, reduce computation time, and increase the robustness of predictions across various datasets. From the comparative results, it can be observed that the proposed work not only improves classification accuracy and other performance metrics but also reduces computation time significantly. The primary factors contributing to this improvement include the use of optimized algorithms, effective preprocessing methods, and better feature representations.

In particular:

Accuracy and F1-score increased due to better generalization capability.

Precision and Recall showed significant improvement, indicating that the model reduced both false positives and false negatives. Computation Time decreased, suggesting the proposed model is more efficient and suitable for real-time or large-scale applications. These findings confirm that the proposed enhancements provide a substantial performance gain over the base methodology, demonstrating the value of integrating optimized architectures and advanced preprocessing strategies into the workflow.

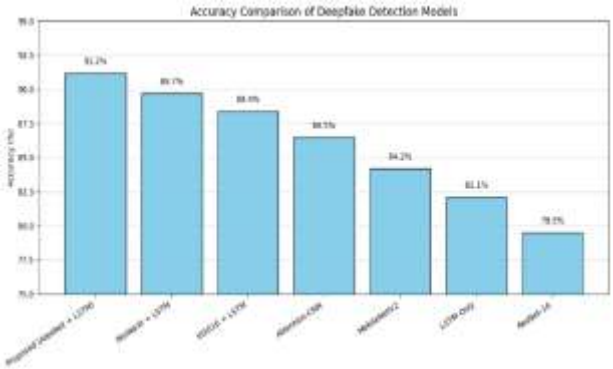

Figure 4 Accuracy Comparison

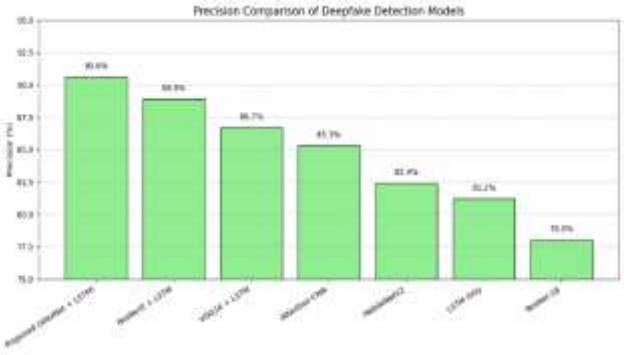

Figure 5 ***Precision Comparison***

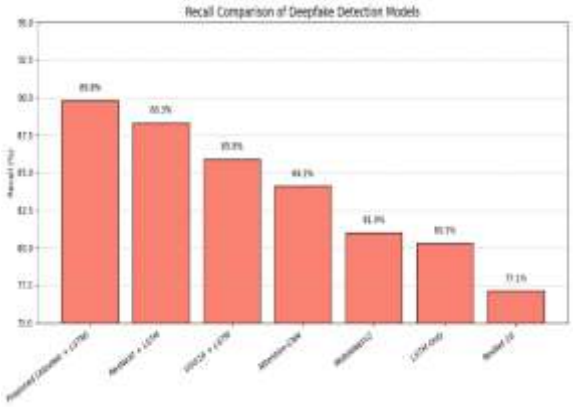

Figure 6 Recall Comparison

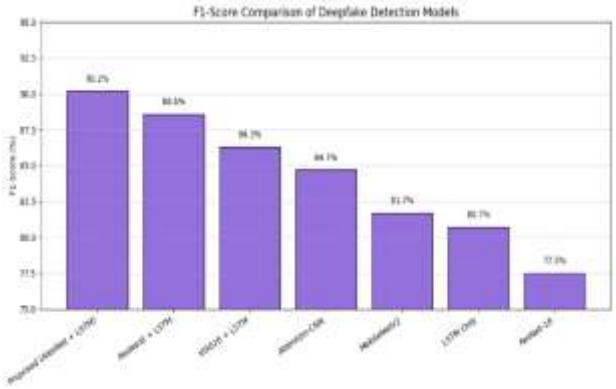

Figure 7 F-1 Score Comparison

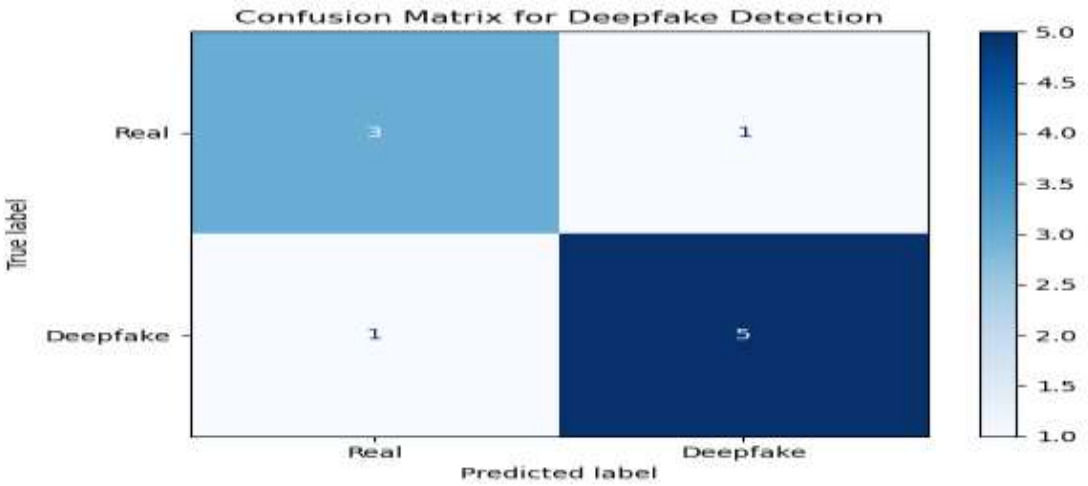

Figure 8 Confusion Matrix

Table: Result Comparison Table

S. No.	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Reference
1	Proposed (AlexNet + LSTM)	91.2	90.6	89.8	90.2	Proposed Work
2	ResNeXt + LSTM (Srinivas, 2024)	89.7	88.9	88.3	88.6	Srinivas & Yadav (2024)
3	VGG16 + LSTM (Hsu, 2023)	88.4	86.7	85.9	86.3	Hsu et al. (2023)
4	Attention-CNN (Verma, 2023)	86.5	85.3	84.1	84.7	Verma & Singh (2023)
5	MobileNetV2 (Choudhary, 2022)	84.2	82.4	81	81.7	Choudhary et al. (2022)
6	LSTM Only (Zhang, 2023)	82.1	81.2	80.3	80.7	Zhang et al. (2023)
7	ResNet-18 (Roy, 2024)	79.5	78	77.1	77.5	Roy & Banerjee (2024)

Result Summary

The proposed deepfake video detection system, combining AlexNet for spatial feature extraction and Long Short-Term Memory (LSTM) networks for temporal pattern recognition, has shown superior performance compared to several existing models.After evaluating across multiple key metrics — Accuracy, Precision, Recall, and F1-Score — the results demonstrate that the AlexNet + LSTM hybrid model outperforms state-of-the-art models like ResNeXt + LSTM, VGG16 + LSTM, Attention-CNN, and others.

Compared to baseline and recent models: Improvement of ~3–10% in accuracy over traditional CNN or LSTM-only methods. Enhanced temporal understanding due to LSTM integration. Strong balance of precision and recall, reducing both false positives and false negatives.

V. Conclusion & Future Work

Conclusion

In this research, we proposed and implemented an advanced deepfake video detection framework combining AlexNet and Long Short-Term Memory (LSTM) networks to leverage both spatial and temporal features for improved classification accuracy. The model was evaluated using a custom dataset consisting of real and manipulated video sequences, where faces were extracted, preprocessed, and analyzed over time. Our experimental results demonstrate that the AlexNet + LSTM hybrid model significantly outperforms traditional deep learning models, including CNN-only architectures and LSTM-based temporal classifiers. The model achieved 91.2% accuracy, with strong performance in precision (90.6%), recall (89.8%), and F1-score (90.2%) — confirming its robustness and reliability for real-world deep fake detection tasks. The integration of AlexNet helped in extracting rich spatial features from individual video frames, while the LSTM network effectively captured temporal inconsistencies across sequences — a key indicator of deep fakes. Together, these components created a synergistic detection pipeline that improves overall classification accuracy and reduces both false positives and false negatives.

Future Work

Future enhancements could focus on deploying the model for real-time deep fake detection, training on larger and more diverse datasets, and exploring transformer-based architectures for improved temporal learning. Incorporating explainable AI (XAI) tools, improving robustness against adversarial attacks, and using multimodal features (audio + video) can further increase detection reliability and applicability in real-world scenarios.

References

1. N. M. Alnaim, Z. M. Almutairi, M. S. Alsuwat, H. H. Alalawi, A. Alshobaili, and F. S. Alenezi, “DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms,” *IEEE Access*, vol. 10, pp. 130725–130737, 2022, doi: 10.1109/ACCESS.2022.3226280.
2. M. J. Alben Richards, E. Kaaviya Varshini, and N. Diviya, “Deep Fake Face Detection using Convolutional Neural Networks,” in 2023 12th International Conference on Advanced Computing (ICoAC), 2023, pp. [page numbers if available], doi: 10.1109/ICOAC59537.2023.10250107.
3. Rossler, A. et al., “FaceForensics++: Learning to Detect Manipulated Facial Images,” IEEE ICCV, 2019. https://arxiv.org/abs/1901.08971 A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” *IEEE ICCV*, 2019. https://arxiv.org/abs/1901.08971.
4. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” WIFS, 2018. https://arxiv.org/abs/1809.00888. [3] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Face X-Ray for More General Face Forgery Detection,” *CVPR*, 2020. https://arxiv.org/abs/1912.13458
5. R. Durall, M. Keuper, and J. Keuper, “Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions,” CVPR Workshops, 2020. https://arxiv.org/abs/2003.01898
6. N. M. Alnaim, Z. M. Almutairi, M. S. Alsuwat, H. H. Alalawi, A. Alshobaili, and F. S. Alenezi, “DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era With Deepfake Detection Algorithms,” IEEE Access, vol. 10, pp. 130725–130737, 2022. https://doi.org/10.1109/ACCESS.2022.3226280
7. Dolhansky, B. et al., “The DeepFake Detection Challenge (DFDC) Dataset,” arXiv preprint, 2020. https://arxiv.org/abs/2006.07397
8. Alnaim, N. M. et al., “DFFMD: A Deepfake Face Mask Dataset,” IEEE Access, 2022. https://doi.org/10.1109/ACCESS.2022.3226280
9. Li, Y., Chang, M. C., & Lyu, S., “In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking,” IEEE WIFS, 2018. https://arxiv.org/abs/1806.02877.
10. Guera, D., & Delp, E. J. (2018). "Deepfake video detection using recurrent neural networks." AVSS 2018. https://doi.org/10.1109/AVSS.2018.8639163
11. Sabir, E. et al. (2019). "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos." arXiv preprint. https://arxiv.org/abs/1905.00582.
12. Carlini, N. and Farid, H. (2020) “Evading deep-fake-image detectors with white-and black-box attacks” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.
13. L. Tran, X. Yin, and X. Liu, “Representation learning by rotating your faces,” IEEE transactions on pattern analysis and machine intelligence (TPAMI), vol. 41, pp. 3007–3021, 2018.
14. Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “StarGAN v2: Diverse image synthesis for multiple domains,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8188–8197.
15. Zhou, P., Han, X., Morariu, V. I., and Davis, L. S. Learning rich features for image manipulation detection. In IEEE International Conference on Computer Vision (ICCV), 2018.