Safeguarding Public Trust Through DeepFake Detection

Authors: Gotam Raj[1], Ritik Lal[2] and Muhammad Moiz Mirza[3]

DHA Suffa University

Abstract:

In today's digital age, the increase of hyper-realistic fake content poses unprecedented threats to information integrity and personal identity security. Advanced Deep-Learning techniques, in particular the use of Auto Encoders (AEs) and Generative Adversarial Networks (GANs) to create deepfakes, are contributing to the failure of public trust in digital media. The impact of deepfakes extends far beyond mere technological concerns, with potential implications for widespread misinformation increased vulnerability to identity theft and fraud.

To mitigate these risks, it is crucial to invest in research and develop advanced Deep-learning approaches focused on accurate identification of deepfakes. Which again comes with different obstacles, including the limited availability of diverse training data, the rapid evolution of deepfake technology, and resource-intensive model development. Additionally, achieving generalization across different content types are important.

In this paper, we explore the complexities of detecting deepfakes, highlighting the importance of a complete strategy that combines new technology with ethical considerations. We'll do this by using the latest research and working together across different fields. This paper aims to provide better understanding of (1) how deepfakes are created and detected, (2) the recent advancement in this domain, (3) limitations of existing detection approaches, and (4) need of further research in required area.

Keywords:- Deep Learning (DL), Auto Encoders (AEs), Generative Adversarial Networks (GANs), Audio Deepfakes (AD), Singing Voice Deepfake Detection (SVDD), Vision Transformer (ViT), Interpretable Spatial-Temporal Video Transformer (ISTVT).

I-Introduction:

With the advent of social media, there has been a rise in the demand for altering multimedia data, such as photos, videos and audios to attract a larger audience [1]. Earlier, the task of manipulating multimedia data was highly time-consuming editing process. The realistic modification of facial features in digital images and videos using these tools like Adobe and GIMP has traditionally faced limitations due to factors such as the requirement for domain expertise, complexity of the process [1]. However, recent advancements in technology have significantly simplified the multimedia data manipulation process, yielding more realistic outputs.

The last few years have been significant change the landscape which reduced efforts required to manipulate content with the rise in deepfake technology. The combination of “deep learning” and “fake,” specifically refers to manipulated content created using artificial neural networks are what the research community commonly refers to as “deepfakes” [2]. While Deepfake techniques rely on advanced deep learning models like Auto Encoders (AEs), which first emerged in 2017 [1], and Generative Adversarial Networks (GANs) to analyze a person’s features and behaviors, enabling the synthesis of manipulated features that mimic similar gestures and movements [1]. These advanced techniques enable users to effortlessly create genuine content with identities that do not exist or produce highly realistic content manipulations without the need for manual editing [3].

While deepfake technology has demonstrated effective applications, such as voice mimicking without reshooting film scenes [4], digital try-ons of clothes, and improved traditional teaching methods to engage students, the negative consequences of deepfake outweigh benefits of this because it raises serious concerns for public security enabling unauthorized manipulation of content specially person's facial expression and voice to spread disinformation in political videos [5], to influence elections, and abusing it on social media platforms [2]. A number of content manipulation applications, such as FaceApp, FaceSwap and ResembleAI, have emerged [1] [6]. The release of another smart undressing app DeepNude in 2019 raised tensions in world made it even worse and It has become increasingly challenging for regular users to filter out manipulated content [1].

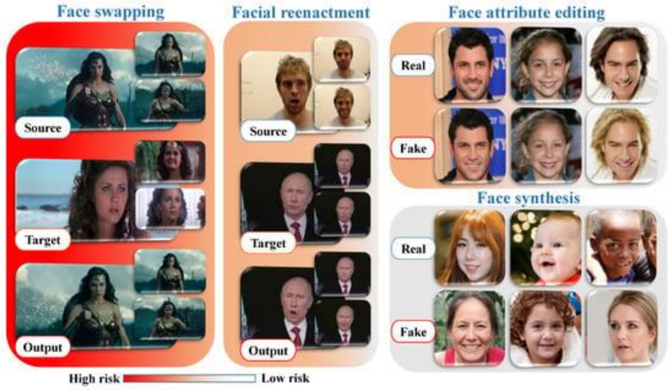

Fig 1: Shows different categories of face manipulation [1].

The research community has put significant efforts to introduce a number of competitions to aid in the development of efficient deepfake detection solutions in response to the growing threat of increasingly sophisticated and realistically manipulated facial images being used to detect deepfake content. Understanding the advancements and limitations of deepfake generation and detection techniques, as well as the current detection methods and obstacles to overcome for implementing effective detection systems, is crucial. These include the following: the Deepfake Detection Challenge (DFDC) sponsored by Facebook [7], the Media Forensics Challenge (MFC2018) sponsored by the National Institute for Standards and Technology (NIST), and the Deeper Forensics Challenge 2020 hosted on the CodaLab platform. Convolutional Neural Networks (CNNs) are often used, however their effectiveness in identifying deepfake content is questionable [1].

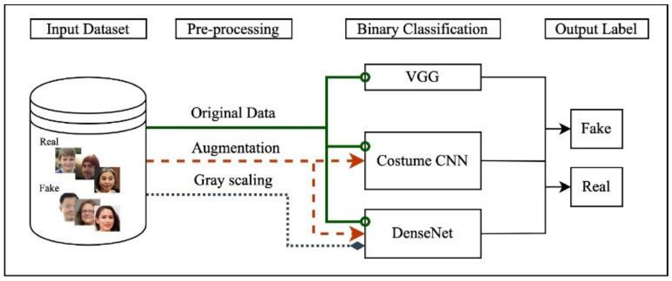

Fig 2: General overview of approach to detect deepfake content [8].

To confront these challenges, our research focuses on developing an advanced detection system that robustly identifies and detects deepfake videos. We employ a novel integration of ResNeXt50 Convolutional Neural Networks (CNNs) for their exceptional spatial analysis capabilities, coupled with Long Short-Term Memory (LSTM) networks that excel in temporal data processing. This hybrid approach harnesses the strengths of both architectures to enhance the accuracy and efficiency of deepfake detection. This paper elaborates on the technical development and rigorous evaluation of our detection system, discusses its broader implications for maintaining digital media integrity, and considers future research directions aimed at bolstering the security and trustworthiness of digital content.

2-LITERATURE REVIEW

Deepfake technology has evolved significantly since the early 1990s, becoming a major phenomenon with the introduction of terms like "deep learning" and "fake" around 2017 [3]. Initially used for entertainment and academic purposes, the real potential for creating hyper-realistic deepfakes was realized with the advent of Generative Adversarial Networks (GANs) in 2014 [1]. By 2017, deepfakes became widely recognized through manipulated celebrity images on public forums, highlighting the risks and the need for effective detection methods. Early detection focused on spotting anomalies like unnatural blinking or inconsistent lighting, but as deepfakes improved, these methods faltered [9]. Recent efforts have concentrated on developing more adaptable and generalizable detection models to keep pace with the rapid advancements in deepfake technology, using comprehensive datasets like DFDC [7], FaceForensics++ [10] and Celeb-DF to enhance detection capabilities across various platforms.

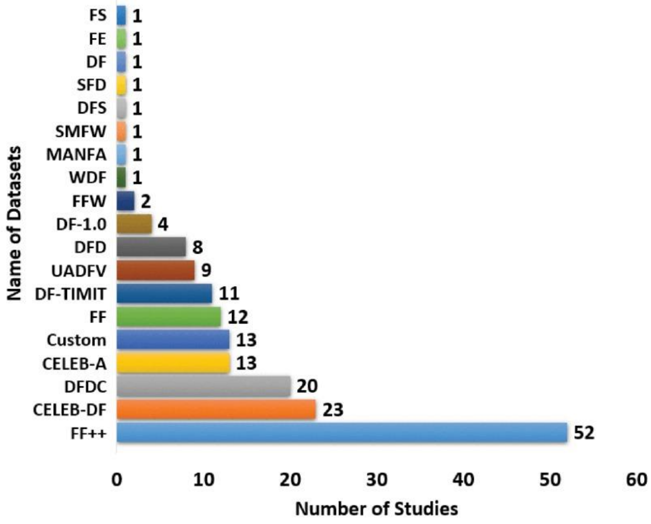

Table 1: Lists no of published researches on several datasets of deepfake [11].

Deepfake detection has become a critical field of study as manipulated media grows ever more sophisticated, thanks to advancements in machine learning technologies and the availability of diverse datasets. The complexity of these deepfakes has prompted researchers to develop a variety of detection techniques that strive for accuracy, efficiency, and adaptability across various platforms and media types.

One effective method employs SVM classifiers enhanced with PCA, transitioning images from RGB to YCbCr color space, which has shown promise in accurately identifying manipulated facial images [12]. Despite its effectiveness, this approach struggles with scalability and adaptability to diverse and dynamic real-time environments. Conversely, time-distributed CNN models integrated with LSTM networks handle sequential video frames effectively, yet they underscore the need for future enhancements in computational efficiency to better manage the demanding nature of video data analysis [13].

The phoneme-viseme mismatch detection technique represents another novel approach, focusing on the discrepancies between audio and visual signals to pinpoint manipulations. This method has proven effective in controlled scenarios but requires further development to enhance its applicability and generalizability across different contexts and languages [14]. Wavelet-packet decomposition techniques, combined with SVM classifiers, also show potential in deepfake detection by analyzing textural and frequency-based features of images, though they need improvements in robustness against various types of deepfake attacks [15].

In the realm of audio deepfake detection, engineers have explored specific feature engineering to enhance the training of machine learning models. Utilizing Mel-Frequency Cepstral Coefficients (MFCC), researchers have developed methods that offer notable accuracy in detecting deepfake audio, suggesting that expanding the diversity of audio features and datasets could further strengthen classifier robustness [16].

Predictive representation models, proposed by researchers such as Shiming Ge and colleagues, analyze video frames for anomalies [3], indicating high accuracy but also highlighting the resource-intensive nature of such approaches. These models demand significant computational power, presenting challenges in deploying them effectively in real-time applications [17].

The battle against deepfakes on social media has led to the development of methods that combine various machine learning models to detect manipulations effectively, though these techniques often remain confined to specific platforms [18]. This limitation underscores the necessity for detection models that boast a broader generalizability across different media types.

Multi-attentional networks that focus on capturing local discriminative features have achieved state-of-the-art performance in detecting deepfakes. However, these networks require substantial computational resources and risk overfitting to specific datasets, indicating a need for balanced approaches that maintain adaptability without compromising performance [19].

Combining audio and visual cues through integrated CNN and RNN architectures has also shown promising results in detecting inconsistencies between modalities, offering high accuracy but at a high computational cost [20]. This complexity necessitates further optimization to make such methods viable for broader applications.

Furthermore, novel data augmentation-based deep learning methods have been explored to enhance detection accuracy and robustness [21]. These methods improve model generalization across different deepfake generation techniques but introduce increased training time and complexity.

High-performance CNN architectures that focus on spatial features within video frames have been developed to detect subtle artifacts indicative of deepfakes. While these models achieve high performance, they face challenges related to dataset variability and require methods that enhance their generalizability [22].

Hybrid models combining CNNs for spatial feature extraction and RNNs for temporal dynamics, often optimized with techniques like particle swarm optimization, advocate for efficient architectures that enhance practical applications, though they remain computationally expensive [23].

Identity-aware deepfake detection techniques use facial recognition technologies to ensure consistency across video frames, achieving high accuracy in specific identity-related forgery scenarios but limited outside these contexts [24].

The ongoing development in deepfake detection illustrates a dynamic field that leverages diverse methodologies to combat the evolving threats posed by digital content manipulation. As researchers continue to push the boundaries of technology, the integration of innovative approaches such as transfer learning and domain adaptation plays a crucial role in enhancing the generalizability and efficiency of deepfake detection systems across various datasets and real-world applications.

3) Proposed Methodology

3.1 System Overview

The methodology for "Safeguarding Public Trust through Deep Fake Detection" integrates recent research insights to develop and evaluate advanced deepfake detection models. Utilizing ResNeXt for feature extraction and LSTM for video classification, the approach aims to enhance detection accuracy and adaptability. The system is designed to efficiently process and analyze video data, structured into key modules for development, testing, and validation against datasets like DFDC, FaceForensics++, and Celeb-DF. The architecture supports both Training

Flow and Prediction Flow, enabling the system to learn from historical data and apply patterns to new, unseen videos.

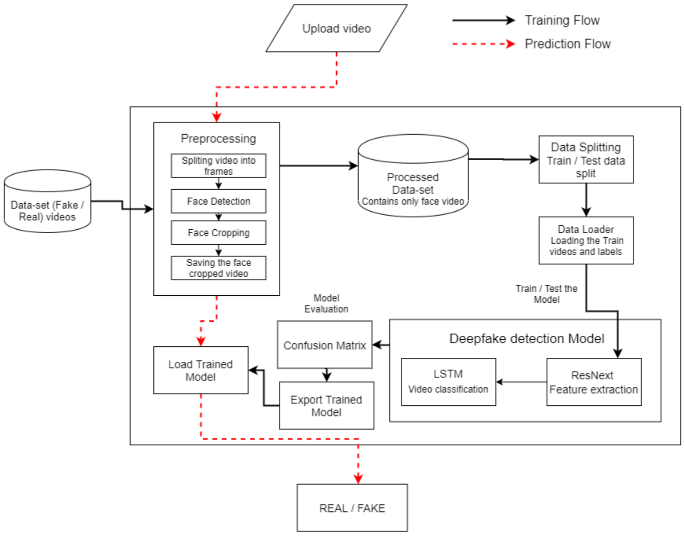
Upload video
Training Flow
Prediction Flow
Data-set (Fake / Real) videos
Preprocessing
Spliting video into frames
Face Detection
Face Cropping
Saving the face cropped video
Processed Data-set Contains only face video
Data Splitting Train / Test data split
Data Loader Loading the Train videos and labels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

Figure 4 shows overall System Architecture

3.2 Dataset Description

The effectiveness of the deepfake detection system depends on the diversity and quality of the datasets used for training and testing. Our system utilizes the following primary datasets, each known for its robustness and relevance in the field:

- **FaceForensics++ (FF++)**: A dataset renowned for its diverse collection of video manipulations created using various deepfake generation methods. It serves as a benchmark for assessing the effectiveness of detection models, containing 2000 videos [10].
- **Deepfake Detection Challenge (DFDC)**: Provided by Facebook, this dataset includes a wide variety of video conditions to simulate real-world applications, making it ideal for testing generalization capabilities, with a total of 3000 videos [7].
- **Celeb-DF**: This dataset comprises high-quality celebrity video deepfakes, which pose a significant challenge due to their realistic manipulations, containing 1000 videos.

These datasets are augmented to create a balanced compilation, consisting of an equal proportion of real and fake videos, enhancing the detection model's accuracy and generalizability across different scenarios. The combined dataset totals 6000 videos, with 50% real and 50% fake videos.

3.3 Network Architecture

The network architecture is designed to efficiently process and classify video data as real or manipulated. The system architecture incorporates two main components:

1. **Feature Extraction with ResNeXt-50**:

The ResNeXt-50 model, known for its deep feature extraction capabilities, is employed to analyze the spatial attributes of each video frame. The architecture of ResNeXt-50, which includes a series of residual blocks with grouped convolutions, allows for efficient learning of intricate patterns in image data.

$$F_l = \mathrm{ReLU}(W_l * F_{l-1} + B_l)$$

Where *Fl denotes* the feature map at the l-th layer, *Wl* and *Bl* are the weights and biases, and $*$ denotes the convolution operation.

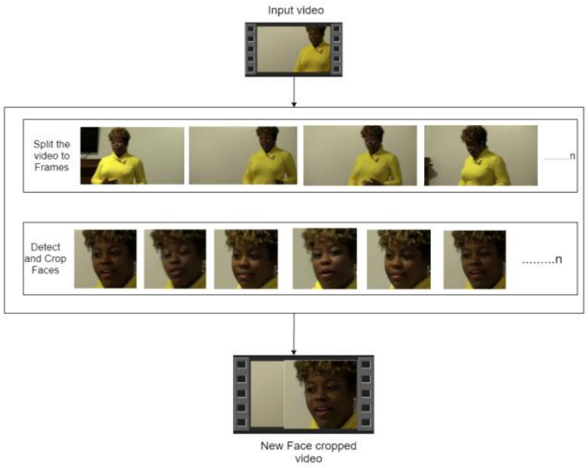

Figure 3 show frame splitting and face cropping

2. **Temporal Analysis with LSTM**:

Following feature extraction, the sequence of feature vectors is processed by an LSTM network designed to capture temporal dynamics and inconsistencies across frames indicative of deepfakes.

$$(h_t,ct_t)=\text{LSTM}(h_{t-1},c_{t-1},x_t)$$

Where h_t and c_t are the hidden and cell states at time t_1, and x_1 is the input at time t.

3. **Classification**:

The combined spatial and temporal features are fed into a fully connected layer followed by a softmax classifier that categorizes the video into 'real' or 'fake'.

$$P(y|x)=\text{softmax}(W_c h_T+b_c)$$

Where P(y|x) is the probability distribution over classes, W_c and b_c are the weights and biases of the classifier, and h_T is the final hidden state of the LSTM.

4. **Optimization**:

The network is optimized using the Adam optimizer, a popular choice for deep learning tasks due to its adaptive learning rate capabilities.

The loss function employed is binary cross-entropy, suitable for the binary classification task:

$$L(y,\hat{y})=-y\log(\hat{y})-(1-y)\log(1-\hat{y})$$

Where y is the true label and y^ is the predicted probability of the video being real.

This comprehensive architecture ensures that the system not only identifies deepfakes with high accuracy but also adapts to new and evolving manipulation techniques, making it robust against various forms of digital content tampering. The modular design allows for scalability and flexibility, accommodating future advancements in deepfake generation and detection technologies.

4) Results and Discussions

The testing and validation phase of our project aimed to ensure the model's effectiveness and reliability under various conditions. The methodologies employed to evaluate the model, the testing procedures, and the results are discussed below.

4.1 Test Setup and Methodology

To validate the accuracy and generalizability of our detection model, the test environment was configured to mirror the training conditions. The testing involved both quantitative and qualitative assessments, using a reserved portion of the dataset that was not utilized during the training phase. This reserved dataset included a diverse mix of deepfake and real videos to thoroughly assess the model's detection capabilities across different scenarios and techniques.

4.2 Performance Metrics

The model's performance was measured using several standard metrics:

- **Accuracy**: The percentage of total correct predictions (both real and fake).

$$\text{Accuracy}=\frac{\text{True Positives (TP)}+\text{True Negatives (TN)}}{\text{Total Number of Samples}}$$

- **Precision and Recall**: Precision measures the accuracy of positive predictions, and recall measures the model's ability to identify all relevant instances.

$$\text{Precision}=\frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Positives (FP)}}$$

$$\text{Recall}=\frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Negatives (FN)}}$$

- **F1-Score**: The harmonic mean of precision and recall, providing a balanced metric.

- **ROC Curve and AUC Score**: The Receiver Operating Characteristic (ROC) curve and the Area under the Curve (AUC) offer insights into the model’s diagnostic ability at various threshold settings.

$$\text{AUC} = \int_0^1 \text{TPR}(t)\,dt$$

4.3 Results

FaceForensics++ Dataset

The model demonstrated high accuracy on the FaceForensics++ dataset, with performance improving as the sequence length increased. The highest accuracy achieved was 97.76% with 100 frames per video.

Model Name	No. of Videos	Sequence Length	Accuracy
model_90_acc_FF_data	2000	20	90.95%
model_95_acc_FF_data	2000	40	95.23%
model_97_acc_FF_data	2000	60	97.49%
model_97_acc_FF_data	2000	80	97.73%
model_97_acc_FF_data	2000	100	97.76%

Celeb-DF Dataset

On the Celeb-DF dataset, the model achieved an accuracy of 93.98%, indicating its robustness in detecting high-quality deepfakes.

Model Name	No. of Videos	Sequence Length	Accuracy
model _Celeb-DF_data	3000	100	93.98%

Combined Dataset

The model's performance on the combined dataset shows good accuracy, with a peak of 89.35% for 40 frames per video, demonstrating its capability to generalize across different types of deepfakes.

Model Name	No. of Videos	Sequence Length	Accuracy
model_87_acc_final_data	6000	20	87.79%
model_84_acc_final_data	6000	10	84.21%
model_89_acc_final_data	6000	40	89.35%

4.4 Discussion

The results suggest that integrating spatial and temporal analytical techniques significantly enhances the detection capability, addressing both subtle and overt deepfake manipulations. The high accuracy achieved across different datasets, including FaceForensics++ and Celeb-DF, indicates the model's strong detection capabilities. However, challenges such as high computational demands and the need for extensive training data were evident.

Key Findings

Advanced Feature Utilization: Through innovative feature extraction techniques, the model effectively distinguished between real and manipulated content.

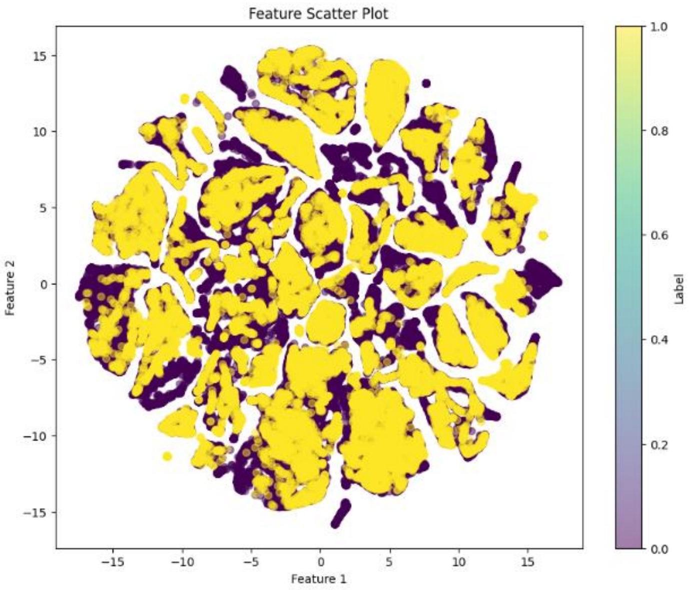

Figure 4 highlights separation between two classes

High Detection Accuracy: The model showed excellent performance on datasets like FaceForensics++ and Celeb-DF.

Effective Generalization: The model maintained robust performance across diverse datasets, indicating its ability to generalize well to various types of deepfakes.

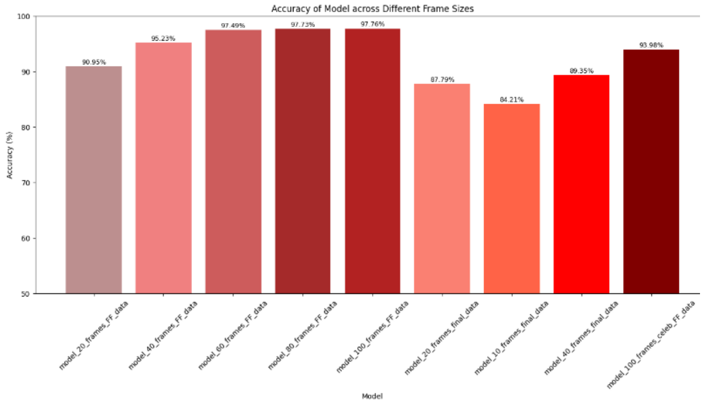

Figure 5 shows Accuracy of model across various frame sizes

5) Conclusion

The project "Safeguarding Public Trust through Deep Fake Detection" addresses the growing threat of deepfake technology by developing an advanced detection system. Our model combines ResNeXt for analyzing spatial features and LSTM networks for processing temporal data, achieving impressive accuracy. It performed notably well on the FaceForensics++, Celeb-DF and our Combined datasets, with accuracy rates of up to 97.76%, 93.98%, and 89.35% respectively.

Comprehensive testing showed the model's robustness and ability to generalize well across different types of deepfakes. Key performance metrics, including accuracy, precision, recall, F1-score, ROC curve, and AUC score, demonstrated the model's effectiveness.

However, we faced challenges such as high computational demands and the need for extensive training data. The model's adaptability to new and evolving deepfake techniques without extensive retraining remains a critical area for future research.

In summary, this project makes significant strides in deepfake detection, providing a robust and scalable solution. By effectively integrating spatial and temporal analysis, our system enhances the detection of manipulated content, contributing to maintaining the integrity of digital media and public trust. Continued innovation is essential to keep up with the evolving nature of deepfake technology and ensure a safer digital environment.

6) Future Work

This study highlights the ongoing need for research in deepfake detection to keep up with rapidly advancing AI-generated content. Future efforts should focus on several key areas. Enhanced model architectures should be developed to reduce computational overhead while improving detection accuracy. Advanced data augmentation techniques are essential to create more diverse training datasets, improving model robustness against various deepfake attacks [21]. Real-world application and testing are necessary to understand the practical effectiveness and limitations of models in diverse environments and on different devices.

Integrating audio analysis with visual data will enhance detection capabilities as deepfakes become more sophisticated across multiple sensory dimensions [4]. Ethical considerations must be addressed to ensure responsible use and prevent misuse, maintaining public trust. Developing more resource-efficient models for deployment on mobile and edge devices without compromising performance is crucial. Reflecting on emerging trends, future research should leverage synthetic data for training robust models, explore hybrid models combining multiple detection modalities, and focus on the temporal dynamics of videos [23]. By pursuing these directions, deepfake detection can advance, ensuring technologies remain robust, efficient, and ethically sound amidst evolving AI capabilities.

References

[1] M. Dang and T. Nguyen, "Digital Face Manipulation Creation and Detection: A Systematic Review.," *Electronics,* vol. 12, no. 16, p. 3407, 2023.

[2] Waseem, Saima and Abu Bakar, Syed Abdul Rahman Syed and Ahmed, Bilal Ashfaq and Omar, Zaid and Eisa, Taiseer Abdalla Elfadil and Dalam and Mhassen Elnour Elneel, "DeepFake on Face and Expression Swap: A Review," *IEEE Access,* vol. 11, pp. 117865-117906, 2023.

[3] Heidari, J. N. Arash, D. Nima, U. Hasan and Mehmet, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *WIREs Data Mining and Knowledge Discovery,* p. e1520, 2023.

[4] Z. M. Almutairi and H. ElGibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," *Algorithms,* 2022.

[5] S. L. C. a. M. K. Barari, "Political deepfake videos misinform the public," *but no more than other fake media,* vol. 13, 2021.

[6] Y. Zang, Y. Zhang, M. Heydari and Z. Duan, "SingFake: Singing Voice Deepfake Detection," *ICASSP,* 2024.

[7] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., &. Ferrer and C. C., "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv,* p. 2006.07397, 2020.

[8] Taeb, Maryam, a. Hongmei and Chi., "Comparison of deepfake detection techniques through deep learning," *Journal of Cybersecurity and Privacy,* vol. 2, no. 1, pp. 89-106, 2022.

[9] Xiangyu Zhu, H. W. . H. F. Z. L. and S. L. , "Face Forgery Detection by 3D Decomposition," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* p. 2928–2938, 2021.

[10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *IEEE/CVF International Conference on Computer,* p. 1–11, 2019.

[11] M. S. Rana, a. Nobi, M. N. Murali, B. Sung and A. H., "Deepfake Detection: A Systematic Literature Review," *IEEE Access,* vol. 10, pp. 25494-25513, 2022.

[12] M. S. M. Altaei, "Detection of Deep Fake in Face Images Based Machine Learning," *Al-Salam Journal for Engineering and Technology,* vol. 2, pp. 1-12, 2023.

[13] A. Saimbhi, N. Singh, A. Singh and M. Mittal, "DeepFake video detection: a time-distributed approach," *N Comput Sci,* p. 212, 2020.

[14] S. Agarwal, H. Farid, O. Fried and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops,* pp. 660-661, 2020.

[15] M. Wolter, F. Blanke, R. Heese and J. Garcke, "Wavelet-packets for deepfake image analysis and detection," *Machine Learning,* vol. 11, pp. 4295-4327, 2022.

[16] A. Hamza, A. Javed, F. Iqbal, N. Kryvinska, A. Almadhor, Z. Jalil and R. Borghol, "Deepfake audio detection via MFCC features using machine learning.," *IEEE Access,* vol. 10, pp. 134018-134028, 2022.

[17] S. Ge, F. Lin, C. Li, D. Zhang, W. Wang and D. Zeng, "Deepfake video detection via predictive representation learning," *ACM Transactions on*

Multimedia Computing, Communications, and Applications (TOMM), vol. 2, pp. 1-21, 2022.

[18] A. Mitra, S. Mohanty, P. Corcoran and E. Kougianos, "A machine learning based approach for deepfake detection in social media through key video frame extraction," *SN Computer Science,* vol. 2, p. 98, 2021.

[19] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang and N. Yu, "Multi-attentional deepfake detection," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 2185-2194, 2021.

[20] D. Cozzolino, A. Pianese, M. Nießner and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 943-952, 2023.

[21] F. Iqbal, A. Abbasi, A. Javed, A. Almadhor, Z. Jalil, S. Anwar and I. Rida, "Data augmentation-based novel deep learning method for deepfaked images detection," *ACM Transactions on Multimedia Computing, Communications and Applications,* 2023.

[22] V. Tran, S. Lee, H. Le and K. Kwon, "High performance deepfake video detection on cnn-based with attention target-specific regions and manual distillation extraction," *Applied Sciences,* vol. 16, p. 7678, 2021.

[23] A. Al-Adwan, H. Alazzam, N. Al-Anbaki and E. Alduweib, "Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm.," *Computers,* vol. 4, p. 99, 2024.

[24] D. Cozzolino, A. Rössler, J. Thies, M. Nießner and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," *In Proceedings of the IEEE/CVF international conference on computer vision,* pp. 15108-15117, 2021.