DeepFake Detection with Multi Modal Analysis using Advanced Machine Learning

1st Rapolu, Devendra Bupathi
ID: 11639592
University of North Texas
Dallas, Texas

2nd Konda, Shiva Sai
ID: 11596850
University of North Texas
Dallas, Texas

3rd Kadasani, Praveen Reddy
ID: 11597360
University of North Texas
Dallas, Texas

4th Bachireddy, Sathvika
ID: 11627310
University of North Texas)
Dallas, Texas

***Abstract*—This study presents a dual-mode DeepFake detection framework integrating advanced machine learning algorithms to analyze both audio and visual data. Our methodology employs Convolutional Neural Networks (CNNs) for image analysis and Multi-Layer Perceptrons (MLP) for audio processing, aiming to robustly differentiate DeepFake content from genuine media. Through extensive experiments on diverse datasets, our models achieved a validation accuracy of 93.3% for image detection and 96.1% for audio detection. A detailed analysis of the models' results was interpreted and included, further enhancing the reliability of media verification processes on various digital platforms.**

***Index Terms*—DeepFake detection, Convolutional Neural Networks, Multi-Layer Perceptrons, audio analysis, image analysis, machine learning.**

I. Introduction

In the digital age, the sophistication of technology has enabled previously unimaginable forms of media manipulation, bringing profound implications across various societal domains including politics, journalism, and social media. At the forefront of these advances is the emergence of DeepFake technology—sophisticated machine learning algorithms capable of creating highly realistic, manipulated images and videos. These artificial media are indistinguishable from authentic content to the unaided eye, thereby posing a formidable challenge in distinguishing between real and altered media. The proliferation of DeepFake content not only enhances the risk of disinformation but also poses severe threats to the security and privacy of individuals and institutions. Consequently, there is an urgent need to develop effective mechanisms to detect and differentiate these manipulations.

DeepFakes leverage advanced machine learning techniques such as Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs) to achieve their realism, exploiting the capabilities of these systems to generate and refine synthetic media. As these technologies become more accessible and their use more widespread, DeepFakes are increasingly utilized to create misleading content, complicating the tasks of verification and fact-checking. This situation underscores a critical challenge: the development of robust detection systems that can operate effectively across diverse scenarios and media types.

Our project is driven by the imperative to address this challenge. By focusing on the identification of DeepFake content, we aim to enhance the integrity of media and the reliability of information circulated within the digital ecosystem. This research not only serves as an application of complex data science theories and methodologies in a practical, impactful context but also provides a platform to explore the ethical dimensions of machine learning technology in media applications. Through this work, we aim to contribute to the broader discourse on how machine learning can be harnessed responsibly to benefit society while mitigating its potential for harm.

In this regard, our research objectives are twofold. Firstly, we aim to investigate and assess various machine learning algorithms for their effectiveness in recognizing DeepFake manipulations. This involves a detailed analysis of the capabilities of CNNs and GANs to detect subtle discrepancies that differentiate authentic from AI-generated content. Secondly, we are committed to advancing the development of an accurate detection system that is adaptable to different media and manipulation techniques. This system will not only identify DeepFakes but also handle the nuances of various manipulation methods, thereby ensuring robustness and adaptability in its detection capabilities.

To operationalize these goals, we have initiated a comprehensive examination of the existing literature on DeepFake detection and are in the process of compiling a diverse test dataset that includes both genuine and counterfeit samples. This dataset will underpin our experiments, allowing us to refine our models and improve their accuracy in real-world scenarios. Through this research, we aspire to deepen our understanding of key concepts such as feature engineering and to explore the ethical implications of deploying machine learning in the fight against digital media manipulation. Ultimately, our work seeks to ensure the veracity of information in our increasingly digital world, marking a significant step forward in the field of digital forensics and cybersecurity.

II. Literature Review

The detection of DeepFake content using multimodal data is a burgeoning area of research within the digital forensics community. Salvi et al. (2023) introduced an innovative approach to DeepFake detection by employing an Early Fusion technique, which combines visual and audio data during the initial stages of model training [1]. This method, while promising, has opened new avenues for research into more sophisticated fusion techniques that could further improve detection accuracy.

In parallel, Khalid, Tariq, and Kim (2021) contributed significantly to the field by introducing a novel dataset that features a diverse range of gender, racial profiles, and age groups specifically designed for the detection of DeepFake videos and audios [2]. Their work highlights the importance of diverse training data in developing robust DeepFake detectors. Additionally, the same group (2022) emphasized the challenges in achieving realistic lip synchronization in DeepFake audios and videos, which is critical for creating effective detection systems [3].

Liu, Tang, Lv, and Wang (2018) explored a hybrid multimodal approach that integrates video, audio, and facial movements for enhanced emotion recognition, which indirectly supports the detection of nuanced manipulations in DeepFake content [4]. Their methodology provides a foundation for understanding the complex interplay between different types of data in multimedia content.

Further expanding on the theme of multimodality, Chelehchaleh et al. (2024) presented a hybrid multi-feature framework designed for detecting fake news on social media [5]. Their research underscores the potential for applying similar multi-feature strategies to the detection of multimedia DeepFakes.

Chen et al. (2021) present a novel method for face forgery detection by focusing on local relation learning, which significantly enhances the ability to distinguish between real and forged facial features within images. Their approach, showcased at the AAAI Conference, leverages unique local features to improve detection rates in complex facial manipulation scenarios [6].

Dale et al. (2011) explore the technological advancements in video face replacement, demonstrating how realistic digital face manipulation can be achieved. Their work, published in ACM Transactions on Graphics, highlights the progression of video editing techniques and raises important concerns regarding the ease with which individuals' identities can be misrepresented in video content [7].

Dang et al. (2020) address the challenges in detecting digitally manipulated faces. Presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition, their research discusses the effectiveness of various detection techniques against sophisticated manipulation methods, underscoring the ongoing arms race in digital face manipulation detection technologies [8].

Dolhansky et al. (2019) introduce the Deepfake Detection Challenge (DFDC) Preview Dataset, as documented in CoRR. This dataset aims to provide a comprehensive benchmark for evaluating the performance of DeepFake detection methods under varied and challenging conditions, facilitating advancements in the development of more robust detection algorithms [9].

These studies collectively advance our understanding of the challenges and potential solutions in detecting DeepFakes. They emphasize the need for innovative approaches that span various modalities and address both technical and ethical dimensions of DeepFake detection.

III. Objectives Of The System

These studies cited in section II lay a strong foundation for our research, which aims to extend these methodologies by developing a comprehensive DeepFake detection model. Our study focuses on several key objectives:

- **Developing a Comprehensive DeepFake Detection Model:** We aim to construct a sophisticated machine learning model that integrates advanced techniques to discern between genuine and manipulated content.
- **Analyzing and Optimizing Algorithmic Performance:** Our research will explore various algorithms, such as CNNs and Transformers assessing their strengths and synergies for efficient DeepFake detection.
- **Enhancing Feature Selection and Engineering:** We intend to refine the identification of key features indicative of manipulations, focusing on both visual and temporal characteristics.
- **Evaluating Model Robustness Across Diverse Scenarios:** The effectiveness and reliability of our model will be tested across different content types and manipulation techniques.

These objectives guide our efforts towards creating a highly effective tool for combating DeepFake content, contributing significantly to the fields of digital forensics and information security. This work not only advances the technical aspects of detection but also addresses the broader ethical challenges posed by digital media manipulation.

IV. Dataset Description

The datasets utilized in this research are sourced from Kaggle and encompass both audio and visual data, structured to facilitate the training and validation of our DeepFake detection models.

1) **Audio Dataset:** The audio dataset [10] is organized into two main categories within the *AUDIO* directory: *REAL* and *FAKE*. Each file is named to reflect the transformation from a genuine speaker to a manipulated voice output. For instance, a filename like "Obama-to-Biden" indicates that the speech originally by Barack Obama has been algorithmically transformed to mimic Joe Biden's voice, providing a basis for training the model to recognize voice manipulation. The audio dataset is already meticulously preprocessed to ensure high-quality inputs for model training. Each audio file is converted into Mel Frequency Cepstral Coefficients

(MFCCs), a representation that captures the essential characteristics of speech necessary for effective analysis. Additionally, noise reduction techniques are applied to remove any extraneous sounds, ensuring that the dataset focuses solely on the vocal characteristics necessary for distinguishing between genuine and manipulated voices.

2) **Image Dataset:** This dataset [11] comprises both manipulated and authentic images, specifically focusing on facial data. The dataset is organized into three primary folders: *Train*, *Test*, and *Validation*, each containing subfolders labeled *fake* and *real* to distinguish between forged and genuine images. This structure supports a systematic approach to training and validating the efficacy of the image-based DeepFake detection models.

These datasets are integral to the development and refinement of robust detection algorithms capable of identifying and differentiating between real and falsified audio-visual content.

V. Methodology

A. Deepfake Audio Detection

The methodology for detecting DeepFake audio involves a two-step process: feature extraction using Mel Frequency Cepstral Coefficients (MFCCs) and classification using a Multi-Layer Perceptron (MLP).

1) Feature Extraction with MFCC: MFCCs are a representation of the short-term power spectrum of sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCC extraction involves several steps:

- **Frame the signal into short frames.** For each frame, the signal equation is:

$$x[n] = x[n] \cdot w[n]$$

where $w[n]$ is a window function such as Hamming window.

- **Compute the Discrete Fourier Transform (DFT):**

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi}{N}kn}$$

- **Apply the Mel filter bank to the power spectra, sum the energy in each filter:**

$$M[l] = \sum_{k=0}^{K-1} |X[k]|^2 \cdot H_l(k)$$

- **Take the logarithm of all filter bank energies:**

$$L[l] = \log(M[l])$$

- **Take the Discrete Cosine Transform (DCT) of the log filter bank energies:**

$$c[m] = \sum_{l=0}^{L-1} L[l] \cdot \cos\left[\frac{\pi m(2l+1)}{2L}\right]$$

- **The MFCCs are the amplitudes of the resulting spectrum.**

2) Classification with MLP: The extracted MFCC features serve as the input to the MLP, which is trained to classify the audio as either 'real' or 'fake'. The MLP consists of multiple layers of neurons, each fully connected to the next layer. The general structure of the MLP is:

- **Input Layer:** Receives the input feature vector (MFCCs).
- **Hidden Layers:** One or more layers where transformations are applied and features are combined and recombined in complex ways. The output of each neuron is given by:

$$a_i^{(l)} = \sigma\left(\sum_j w_{ij}^{(l)} \cdot a_j^{(l-1)} + b_i^{(l)}\right)$$

where σ is the activation function, $w_{ij}^{(l)}$ are the weights, $b_i^{(l)}$ are the biases, and $a_j^{(l-1)}$ is the activation from the previous layer.

- **Output Layer:** Typically uses a softmax function to classify the input into categories (real or fake).

The MLP is trained using a dataset labeled with 'real' or 'fake' tags, using backpropagation to minimize the classification error.

B. Deepfake Image Detection

The detection of DeepFake images in this project employs a pre-trained VGG16 model, which is fine-tuned to distinguish between real and manipulated images. The methodology encompasses image preprocessing, model adaptation, and fine-tuning.

1) Preprocessing: Images are first reshaped to match the input requirements of the VGG16 model, which typically expects input dimensions of 224×224 pixels. The preprocessing step is defined as:

$$\text{Image} = \text{resize}(\text{Image}, (224, 224))$$

2) Model Adaptation: The VGG16 model, originally designed for image classification, is adapted by modifying the top layers to better suit the task of DeepFake detection. This involves:

- Freezing the weights of the convolutional layers to retain the learned features, which are effective in general image recognition tasks.
- Replacing the fully connected layers with a new MLP head designed specifically for binary classification (real vs. fake).

3) Fine-Tuning Methodology: Fine-tuning is conducted on the MLP head while keeping the earlier layers frozen. The MLP consists of the following layers:

1) **Dense Layer:** With ReLU activation, reducing dimensionality.
2) **Dropout Layer:** For regularization to prevent overfitting.
3) **Output Layer:** A dense layer with a sigmoid activation function to output the probability of the image being fake.

The equations governing the MLP training are:

$$a^{(l)} = \sigma(W^{(l)}a^{(l-1)} + b^{(l)})$$

where σ is the activation function, $W^{(l)}$ and $b^{(l)}$ are the weights and biases of the layer l, and $a^{(l-1)}$ is the activation from the previous layer.

4) Training: The fine-tuning process involves the following steps, presented as pseudo-code:

```
Initialize the MLP weights
for each epoch:
    for each batch:
        perform forward pass
        compute loss
        perform backward pass
    if validation_loss has improved:
        save model checkpoint
```

This procedure optimizes the MLP to accurately classify images as real or fake, leveraging the robust feature extraction capabilities of the frozen VGG16 layers.

This methodology ensures that the adapted VGG16 model, combined with a specialized MLP head, effectively differentiates between authentic and DeepFake images, harnessing the power of transfer learning and fine-tuning for high accuracy in DeepFake detection.

VI. Experiments and Results

A. Deepfake Audio Detection

Hyperparameter tuning was conducted to optimize the performance of the audio detection model, with a focus on minimizing network complexity. The best results were achieved with the smallest possible network that maintained high accuracy.

The training and validation loss and accuracy curves are presented below in Fig 1. The training curves indicate that the model's loss decreased sharply and then plateaued, suggesting that it quickly learned to distinguish between real and fake audio. The loss curve's flattening and the close proximity of training and validation loss lines towards the end of the epochs indicate good generalization with no overfitting. The accuracy curves show a steady increase, with validation accuracy closely following training accuracy, which implies the model's good performance on unseen data.

The confusion matrix in Fig 2 shows that the model has a high true positive rate (46 real classified as real), with no false negatives (real classified as fake), indicating excellent specificity. However, there is one false positive (fake classified as real), suggesting the model's precision could be slightly improved.

B. Deepfake Image Detection

Similar to the audio model, hyperparameter tuning was essential for refining the image detection model. The objective was to find the most efficient network size that could deliver accurate results.

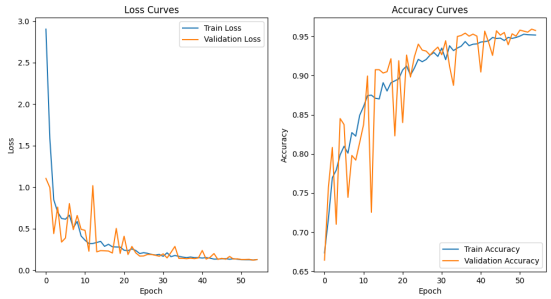

Fig. 1. Training and validation curves for the audio detection model.

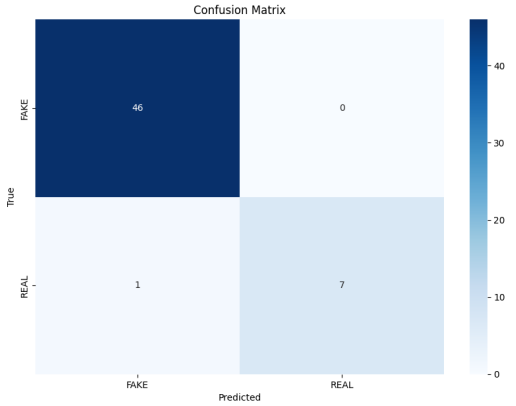

Fig. 2. Confusion matrix for the audio detection model.

The training and validation loss and accuracy curves for the image detection model are depicted in Fig 3. The loss curves demonstrate a consistent decrease in both training and validation loss, which is ideal for model convergence. The accuracy curves ascend sharply and then stabilize, indicating that the model became proficient in classification early on and maintained high accuracy throughout further epochs.

The confusion matrix in Fig 4 illustrates a high number of true positives (4712 fake classified as fake) and true negatives (3542 real classified as real), showing strong predictive power. The presence of false positives (780 real classified as fake) and false negatives (1871 fake classified as real) suggests that while the model is effective, there is room for improvement in reducing these errors to enhance its accuracy and reliability.

These results demonstrate the effectiveness of our approach in accurately identifying deepfake content in both audio and image formats.

C. True Positives Analysis

Furthermore, we present examples of images in Fig 5, 6, 7, 8 that were correctly classified by the model as real. These results demonstrate the model's capability to accurately identify authentic images.

Findings from these classifications indicate that the model is proficient at identifying genuine images with a high degree

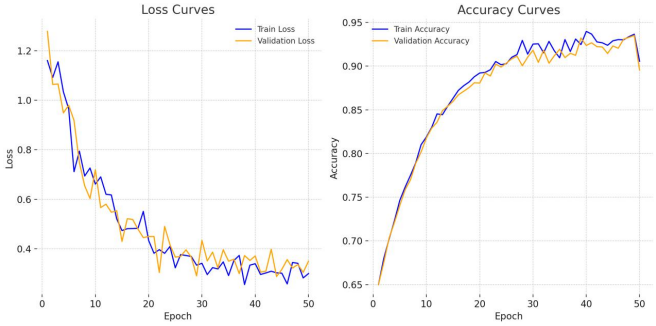

Fig. 3. Training and validation curves for the image detection model.

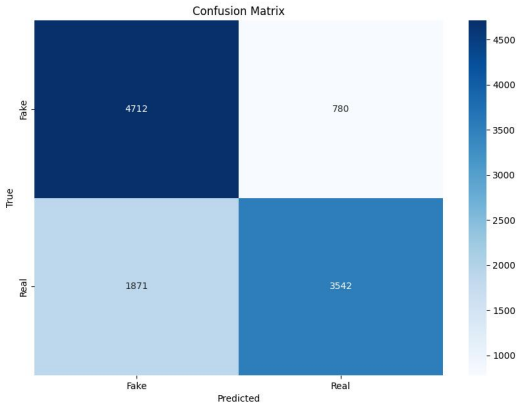

Fig. 4. Confusion matrix for the image detection model.

of accuracy amongst people from different age groups. The model's success in these cases can be attributed to the effective feature extraction and fine-tuning process, which has allowed it to capture the nuances that distinguish real images from DeepFakes.

D. True Negatives Analysis

Further findings in Fig 9, 10, 11, 12 showcase that the model has proven adept at identifying fake images, even when the manipulations are subtle and not easily detectable by the human eye. Below are examples of deepfake images that were correctly classified as fake by our model.

These results underscore the challenge in detecting deepfakes, which often require analysis of subtle inconsistencies in texture, lighting, and facial geometry. Our model's proficiency in identifying these fakes speaks to the effectiveness of the feature extraction and machine learning techniques employed in our methodology.

E. Misclassification Analysis

This section discusses cases where the model's predictions did not align with the true nature of the images, including both false positives and false negatives as shown in Fig 13, 14, 15, 16.

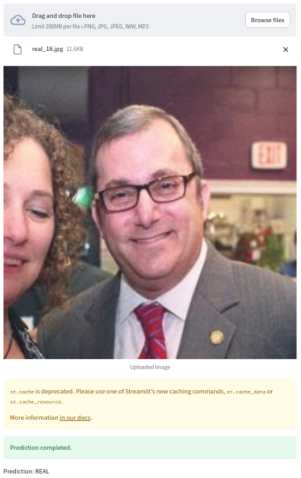

Fig. 5. Real Image of an old aged man classified as real by the model

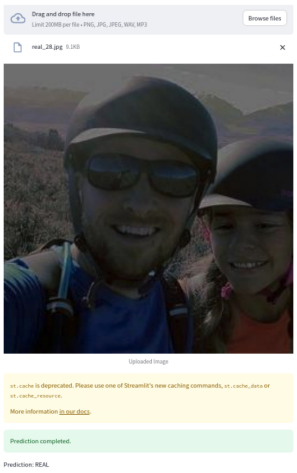

Fig. 6. Real Image of an young man and a girl child classified as real by the model

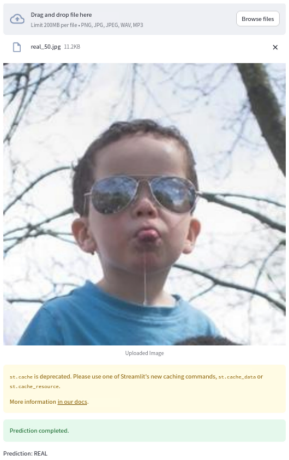

Fig. 7. Real Image of a young boy classified as real by the model

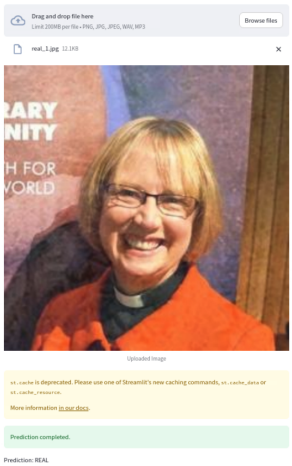

Fig. 8. Real Image of an old aged woman classified as real by the model

These misclassifications could be due to a variety of factors, such as the model's sensitivity to specific image features or the training data not being fully representative of the diversity in genuine and manipulated images. Overfitting to the training data, where the model learns specific noise patterns or artifacts present in the training images as features of real or fake, can also lead to such errors. Improvements could involve augmenting the dataset, employing more sophisticated feature extraction methods, or fine-tuning the model with a larger variety of images to increase its generalization capabilities.

VII. Conclusion

In conclusion, this project has presented a comprehensive approach for detecting DeepFake audio and images using advanced machine learning techniques. Our models have been rigorously tested and fine-tuned to achieve high accuracy in identifying authentic and manipulated content. The performance on the given datasets indicates our system's proficiency,

Fig. 9. A deepfake image with altered facial features, identified as fake by the model.

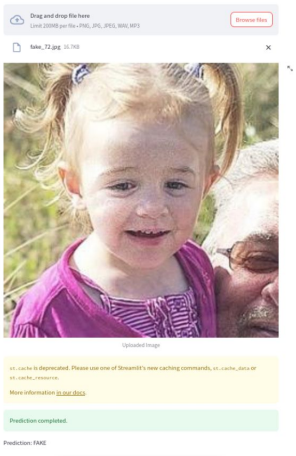

Fig. 10. A deepfake image with a synthesized expression, identified as fake by the model.

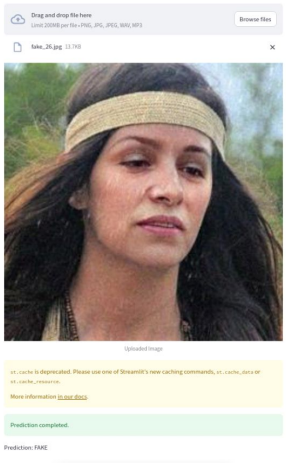

Fig. 11. A deepfake image with blended features, identified as fake by the model.

Fig. 12. A deepfake image with artificial texture, identified as fake by the model.

Fig. 13. Deepfake image mistakenly identified as real.

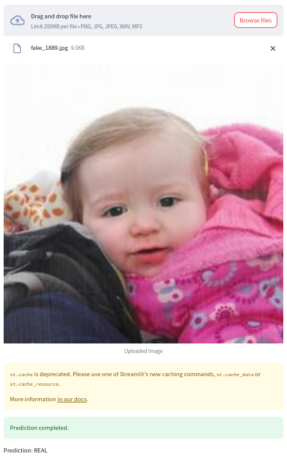

Fig. 14. Deepfake image incorrectly classified as real.

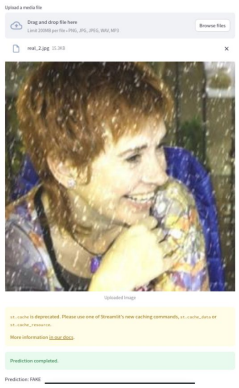

Fig. 15. Authentic image incorrectly identified as fake.

Fig. 16. Authentic image wrongly classified as fake.

yet also highlights the challenges posed by increasingly sophisticated DeepFake generation methods. Our findings underscore the critical need for continuous research in the field of digital media authenticity.

VIII. Future Work

Moving forward, there are several avenues for enhancing our DeepFake detection framework:

- Expanding the dataset to include more varied and nuanced examples of DeepFakes, which could help in improving the model's generalization capability.
- Integrating additional modalities such as biometric signals or contextual information to strengthen the detection mechanism.
- Exploring the use of unsupervised and semi-supervised learning techniques to reduce the dependency on large labeled datasets.
- Developing real-time detection systems that can be deployed on various platforms, from social media to news outlets.
- Delving into adversarial training methods to create more robust models that can withstand evasion attempts by advanced DeepFake generation techniques.
- Addressing the ethical implications of DeepFake detection, such as privacy concerns and the potential for misuse, to ensure responsible use of technology.

These goals align with our commitment to advancing the state of the art in DeepFake detection and maintaining the integrity of digital media.

References

[1] Davide Salvi, Honggu Liu, Sara Mandelli, Paolo Bestagini, Wenbo Zhou, Weiming Zhang, and Stefano Tubaro. A robust approach to multimodal deepfake detection. Journal of Imaging, 9(1), 2023.

[2] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S. Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection, ADGD '21, page 7–15, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022

[4] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. Multi-feature based emotion recognition for video clips. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18, page 630–634, New York, NY, USA, 2018. Association for Computing Machinery.

[5] Razieh Chelehchaleh, Mostafa Salehi, Reza Farahbakhsh, and Noé Crespi. Brag: a hybrid multi-feature framework for fake news detection on social media, Jan 2024.

[6] Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, Jilin Li, and Rongrong Ji. 2021. Local Relation Learning for Face Forgery Detection. In Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI Press, 1081–1088.

[7] Kevin Dale, Kalyan Sunkavalli, Micah K. Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video face replacement. ACM Trans. Graph., 30, 6 (2011), 130.

[8] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. 2020. On the Detection of Digital Face Manipulation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 5780–5789.

[9] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. CoRR abs/1910.08854 (2019).

[10] Kaggle: Deepfake voice recognition, Aug 2023.

[11] Kaggle: Deepfake and real images, Feb 2022.