Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19-21, 2024
Paper No. MVML 115
DOI: 10.11159/mvml24.115

Development of a Multimodal Framework for Deepfake Detection: Combining Visual and Audio Analysis

Ahmed Ashraf Bekheet[1], Ghada Khoriba[2], Amr S. Ghoneim[2]
[1]Computer Science Department, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt
Giza, Egypt
ahmed_ashraaf@fci.helwan.edu.eg; ghadakhoriba@nu.edu.eg
2Center for Informatics Science (CIS), School of Information Technology and Computer Science (ITCS), Nile University
Giza, Egypt
amr.ghoneim@fci.helwan.edu.eg

Abstract **-** Machine learning and social media advancements enable the rapid spread of realistic fake content, encompassing images, videos, and audio. Initially, fake content generation primarily focused on manipulating either audio or video streams. However, recent advancements in deep learning have enabled more sophisticated alterations, commonly called "deepfakes." While existing research predominantly concentrates on detecting fake videos by exploiting either visual or audio modalities, few approaches address audio-visual deep-fake detection. Nevertheless, these methods often need more accuracy when evaluated on a multimodal dataset with deepfake videos and manipulations in both streams. Due to neglecting facial features in preprocessing and using traditional training models. In response to this challenge, we propose a robust audio-visual deepfake detection (MAVDD) approach that analyzes audio and visual streams to enhance detection capabilities. Effectively utilizing pretrained models in image classification tasks for detecting visual deepfakes, alongside advanced preprocessing techniques for optimal facial and audio features extraction. Our experiments conducted on the multimodal audio-visual deepfake dataset "FakeAVCeleb" demonstrate that our proposed approach surpasses both unimodal (audio-only and visual-only) and multimodal (audio-visual) deepfake detection approaches in terms of accuracy and AUC (Area Under The Curve) as dedicated to tables I, II, and III. The implementation of our research work and the dataset are publicly available at the following link: https://github.com/mutlimodalDeepfakeDetection/AV-detector.

Keywords**:** Audio-visual Deepfake Detection - Multimodal - Unimodal - Pre-trained Model - Features Extraction - Data Augmentation

1. Introduction

In recent years, the proliferation of image and video capture has significantly contributed to the wealth of information available on social media platforms. Concurrently, technological advancements have led to the development of manipulation tools, allowing users to easily alter images, audio, and videos through basic adjustments like cropping and color manipulation, often relying on conventional techniques.

Neural architectures have furthered these capabilities, creating more complex "deep-fake techniques" utilizing deep learning. These techniques involve intricate manipulations of digital media content. Traditional and deepfake techniques have played a significant role in spreading fake news.

Generative Adversarial Networks (GANs) [1] are the primary methodology for generating synthesized videos. Synthesized videos, known as video deepfakes, involve the fabrication of fake content by replacing a person's face with another individual (Face Swap), altering their expressions (Expression Swap), or synchronizing lip movements with external sound (Puppet Mastery). Audio deepfakes, on the other hand, entail the generation of cloned voices, making it appear as if a person is saying things they never uttered. Text-to-speech synthesis (TTS) and Voice Conversion (VC) are key techniques for creating audio deepfakes. In TTS synthesis, the person's authentic voice is synthesized based on the provided input text. Conversely, VC is a technology that modifies a source person's audio to resemble a target person's voice. Hence, the realistic quality achieved by deepfake videos and audio produced by the previously mentioned deep learning algorithms has reached a point where distinguishing them from genuine content poses a considerable challenge for human perception.

Videos pose more significant challenges than images or audio, given that video manipulation can occur in both visual and audio domains. Consequently, our paper presents a robust multi-modal deepfake detection approach that operates on audio and visual streams.

In visual deep fake detection, our approach utilizes frame extraction and subsequent facial region cropping for preprocessing. The preprocessed data is then fed into a deep-learning model trained for facial feature extraction and authenticity prediction.

In audio Deepfake detection, Preprocessing involves extracting Mel Frequency Cepstral Coefficients.
(MFCC) or Mel-spectrogram features from the audio signal.

These features capture the audio's time-frequency representation and essential characteristics, aligning with human perception. Subsequently, the preprocessed audio data is fed into a separate deep-learning model that extracts these features and predicts the authenticity of the audio content.

Our approach incorporates the multimodal deepfakes dataset FakeAVCeleb [2], encompassing videos susceptible to manipulation in either audio, video, or both modalities.

In summary, the main contributions of this paper are as follows:

1) An enhanced multimodal audio-visual deepfake detector, named MAVDD, is crafted by optimizing audio and Visual preprocessing procedures are based on efficient pre-trained models.
2) A comprehensive comparison of our proposed approach with existing research in the domain of unimodal audio-only, visual-only, and multimodal audio-visual deepfake detectors.

The paper is outlined as follows: Section 2 reviews existing literature on detecting deepfakes in audio-only, visual-only, and audio-visual contexts. The proposed approach is presented in Section 3. Section 4 discusses details regarding the FakeAVCeleb dataset, the detailed implementation of our approach, and comprehensive results, including a thorough comparison with related approaches in visual-only, audio-only, and audio-visual deepfake detection. The paper concludes with Section 5, covering the summary and future directions for research.

2. Related Work

Significant research efforts have been directed towards detecting deepfakes in multimedia content. Based on the modalities involved, techniques for detecting deepfakes can be grouped into three categories.

Visual deepfake detection extensively employs deep learning techniques such as CNNs, RNNs, and Transformers. CNNs, such as ELA-based forged face detection, are used for feature extraction and classification [3]. RNNs, including LSTM, process sequential data as exemplified in a hybrid CNN-LSTM model designed for detecting deepfakes in video with optical flow features [4]. Transformers used in visual deepfake detection, such as Convolutional Vision Transformers (CViT) [5], combine CNNs with Transformer-based processing for complex data relationships.

Audio deepfake detection methods are categorized into feature-based, image-based, and waveform-based approaches. Feature-based methods extract features from short-term window transforms, such as Mel Frequency Cepstral Coefficients (MFCC), while image-based methods analyze spectrograms as images. For instance, Bartusiak et al. employed normalized grayscale spectrograms and waveform-based methods with deep neural networks for synthetic speech detection [6].

Audio-visual deepfake detection methods, such as AVFakeNet [7], JointAV [8], and VFD [9], utilize multimodal models that integrate both audio and visual data. However, none have explored the optimal rate for extracting frames from videos. This is crucial as an increased frame extraction rate may result in redundant frames without added benefits, while a decreased rate may lead to the omission of essential frames. Furthermore, none of these methods have addressed the performance of their multimodal detectors in situations involving missing modalities, specifically audio.

3. The Proposed Method

Our multimodal audio-visual deepfake detector (MAVDD) consists of two components: the visual module, named M_{visual}, and the audio module, named M_{audio}. M_{visual} focuses on detecting visual deepfake video content, providing a binary classification (C_{visual}) of real or fake. M_{audio} concentrates on detecting audio-based deepfake content, creating a binary classification (C_{audio}). A video is considered real (C_{video} = real) only when both audio and visual modalities yield a real

classification (C_{audio} = C_{visual} = real); otherwise, it is classified as fake (C_{video} = fake) if either or both modalities result in a fake classification (C_{audio} = fake or C_{visual} = fake or both). Figure 1 illustrates the MAVDD, which analyzes both audio and visual data streams.

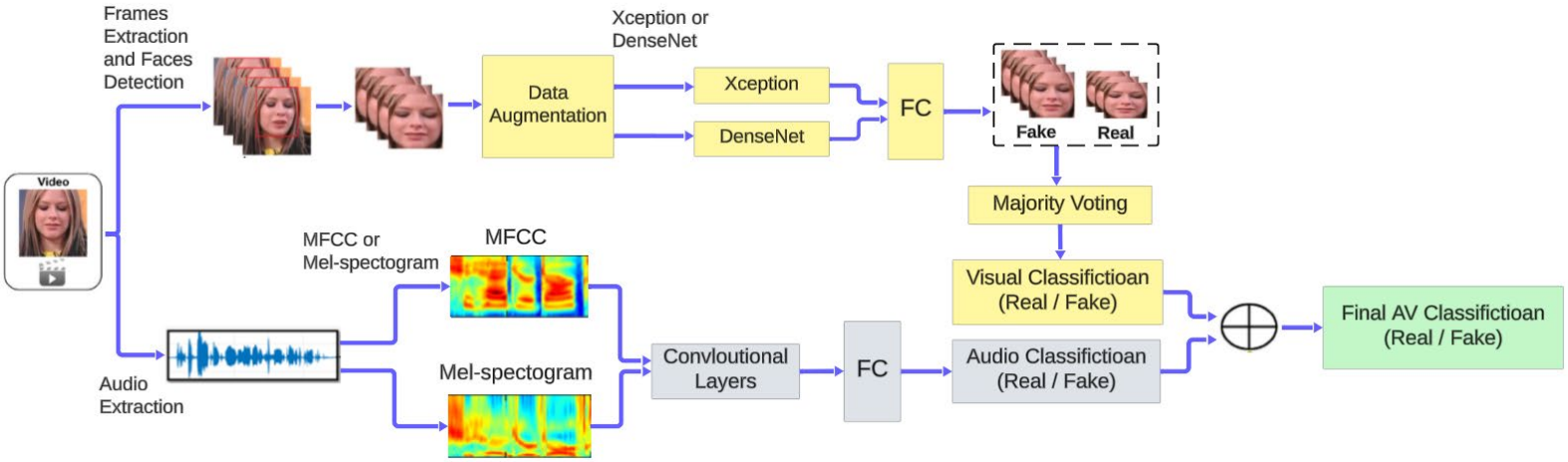

Fig. 1. Overview of our proposed Multimodal Audio-Visual Deepfake Detector (MAVDD).

3.1. Visual Deepfake Detection Modality

Unlike conventional methods that blindly sample frames at a fixed rate, our approach breaks the mold by dynamically selecting frames based on facial changes within the video, focusing on capturing alterations in facial expressions and features. Recognizing the imbalance in FakeAVCeleb dataset, where real visual data constitutes the minority class, we adopted data augmentation instead of oversampling. Augmentation techniques such as rotations, transformations, and adjustments were applied to enrich the dataset's diversity, enhancing model generalization. After that, we employ pre-trained models such as Xception [10] or DenseNet [11] for our deep learning model, leveraging their extensive training on general image classification tasks. This groundwork aids in generalizing to diverse datasets, improving performance on new data, and reducing computational complexity. By fine-tuning these models for our task, we exploit their learned features while customizing top layers for task-specific requirements. Our approach utilizes the pre-trained model as a feature extractor, then classification using fully connected and output layers. To assess video authenticity, individual frame classifications undergo a majority voting rule to determine the overall video classification. The predominant classification among all frames within the video is considered, with the class having the higher count assigned to the entire video.; for instance, if a video comprises 55 frames, with 30 classified as fake and 25 as real, the video is categorized as fake based on the majority voting outcome.

3.2. Audio Deepfake Detection Modality

In audio deepfake detection, the initial step involves extracting the audio component from the video source, followed by crucial preprocessing. This involves extracting either Mel Frequency Cepstral Coefficients (MFCC) or Mel-spectrogram features from the audio data, chosen for their effectiveness in capturing intrinsic audio characteristics. Both MFCC and Mel-spectrogram provide compact representations of sound signals, with MFCC capturing spectral features and Mel-spectrogram offering a visual depiction of frequency components over time. These methods detect subtle variations in audio content, essential for detecting anomalies or manipulations. Following preprocessing, the data undergoes analysis through convolutional layers, and the data is adept at capturing hierarchical features. It proceeds through fully connected and output layers, creating a binary classification indicating authenticity or manipulation.

4. Implementation Details and Results

This section provides comprehensive details regarding the utilized dataset, implementation, and comprehensive results, including comparisons with related approaches.

4.1. Dataset

The FakeAVCeleb dataset comprises 20,000 videos featuring audio and visual manipulations of celebrities, derived from the VoxCeleb2 dataset [12] with content from 500 celebrities. It offers balanced representation across various demographics and manipulation types, making it ideal for training deep learning models with strong generalization abilities. The dataset includes 500 authentic videos and 19,500 fake videos categorized into four classes:

- RARV (Real Audio Real Video): Authentic videos with genuine audio and visual content.
- FARV (Fake Audio Real Video): Videos with manipulated audio while retaining genuine visual content.
- RAFV (Real Audio Fake Video): Videos featuring authentic audio but manipulated visual content.
- FAFV (Fake Audio Fake Video): Videos exhibiting fabrication in both audio and visual aspects.

4.2. Implementation Details

We present a detailed technical walkthrough of MAVDD, encompassing an overview of the utilized dataset and a meticulous breakdown of each implementation stage involved in training the visual and audio streams.

We partition the FakeAVCeleb dataset into training, validation, and test sets, with proportions of 60%, 20%, and 20%, respectively. The 20% allocated for the test set remains consistent and is reserved for comprehensively evaluating the model's performance on audio and visual streams using previously unseen data. In contrast, the training and validation sets undergo random partitioning during each runtime.

Implementation of Visual Modality

The initial stage systematically extracts frames from each video across the four classes. Unlike other methods that uniformly sample frames, our approach dynamically selects frames based on facial changes within the video. This process, facilitated by OpenCV, extracts frames at a default rate of 25 per second and applies face detection to capture facial emotions. Frames with varying facial expressions are selectively saved, resulting in images representing diverse emotional expressions. Facial emotion classification is conducted using the pre-trained deep learning model FER, with detected faces resized to 128 x 128 dimensions to reduce computational cost.

As mentioned before, in the FakeAVCeleb dataset, there exists an imbalance in the visual data, with the minority class being real visual data and the majority class being fake. To address this imbalance, data augmentation will be employed to oversample the minority class. Our model employs data augmentation techniques using ImageDataGenerator, a utility class integrated into the Keras library within TensorFlow. This class facilitates real-time data augmentation during the training of deep learning models. Following data augmentation, the preprocessed data is introduced into our deep learning model, incorporating various pre-trained models. Upon evaluating several pre-trained models, including Xception, VGG-19, InceptionNet, DenseNet, and EfficientNet, a comprehensive analysis revealed that Xception and DenseNet exhibit optimal compatibility with our data, particularly suited for facial deepfake detection. Consequently, each pre-trained model will be individually employed in our subsequent analyses.

The Xception model is initialized with weights from the ImageNet dataset [13], excluding the top classification layer to facilitate transfer learning. A fine-tuning strategy is applied, freezing 50 layers of the model while enabling subsequent layers to adapt during training. For optimization, the Adam optimizer is utilized with a learning rate of 1e-6, and the binary cross-entropy loss function is employed, with accuracy as the primary evaluation metric. The training spans 50 epochs, with a batch size of 32, and model performance is assessed on a validation set. Early stopping, with a patience parameter of 15 epochs, is implemented to mitigate overfitting. During training, the model achieves a training set accuracy of 99.9%, while the validation set accuracy reaches 97.5%.

The identical implementation was executed using DenseNet instead of Xception, involving a modification solely in the

patience parameter, set at 20 epochs, and an adjustment of the learning rate to 1e-7 in the Adam optimizer. The model achieves a training set accuracy of 99.9% throughout the training process, while the validation set accuracy reaches 98.6%.

Implementation of Audio Modality

Audio extraction from each video is conducted, saving the audio content in .mp3 format files. Subsequently, preprocessing entails extracting MFCC or Mel-spectrogram features from these audio files, generating corresponding images or representations. The resulting images are stored as NumPy arrays. The deep learning model is trained twice, employing each preprocessed data independently, to identify the most suitable representation, either Mel-spectrogram or MFCC, for the audio deepfake detection task. The model architecture in the first experiment, utilizing Mel-spectrogram representations, comprises convolutional layers, max-pooling and densely connected layers. Specifically, two convolutional layers with 64 and 32 filters are followed by max-pooling layers for spatial downsampling. The flattened representation undergoes processing through densely connected layers, incorporating rectified linear unit (ReLU) activation functions. Dropout regularization with a rate of 0.5 is applied to mitigate overfitting. Utilizing a sigmoid activation function, the output layer yields binary classification results. A learning rate scheduler, governed by a decay factor, dynamically adjusts the learning rate during training. The model is compiled using the Adam optimizer with binary cross-entropy as the loss function, and accuracy is the primary metric. During training, the model achieves a training set accuracy of 95%, while the validation set accuracy reaches 99.9%. The identical model implementation is applied to MFCC input, including a reshape layer to adapt to the specific input shape of MFCCs. During training, the model achieves a training set accuracy of 99.9%, and the validation set accuracy reaches 99.9%.

4.3. Results and Analysis

MAVDD analyzes both visual and auditory cues for video classification. As previously defined, a video is considered real only if visual and audio models classify it as real ($C_{visual} = C_{audio} = real$). Recognizing that some videos might lack an audio track (e.g., silent clips), MAVDD seamlessly adjusts its classification process, relying exclusively on available visual cues in such cases. The absence of audio ($C_{audio} = missing$) triggers a mechanism that utilizes solely the visual model's output (C_{visual}) for the final video classification ($C_{video} = C_{visual}$). This robust handling of missing audio ensures reliable classification performance, even for videos with incomplete modalities.

The evaluation involves processing the entire test set through our visual and audio models. However, before joint analysis, we independently assess each modality.

- **Visual-only deepfake detection:** We employ majority voting on individual frame classifications. The Xception-based model achieves 97% accuracy and 94.8% AUC, demonstrating robust training and generalization, while the DenseNet-based model achieves 93.5% accuracy but has a lower AUC (85.6%).

The Xception-based model's superior AUC and minority class discrimination highlight its stronger generalizability.

- **Audio-only deepfake detection:** Mel-spectrogram and MFCC representations achieve near-perfect results (99.9% accuracy and AUC), showcasing robust training and generalization. Notably, MFCC representations show slight performance improvements.

In **Multimodal Audio-Visual Deepfake Detection,** as previously noted, our visual approach utilizes either Xception or DenseNet models. Simultaneously, the audio approach leverages either MFCC or Mel-spectrogram features. This combination results in four distinct model configurations:

Xception-MFCC: Achieves 98.6% accuracy and 96.8% AUC, demonstrating the most effective combination.
Xception-Mel-spectrogram: Yields 94.1% accuracy but lower AUC (88.2%).
DenseNet-MFCC: Achieves 93.5% accuracy and 84% AUC, performing well but not as effectively as Xception-MFCC.
DenseNet-Mel-spectrogram: Shows the lowest performance (92.7% accuracy, 83.5% AUC).

These outcomes indicate that the most effective technique for video deepfake detection is Xception with MFCC, achieving an accuracy of 98.6% and an AUC of 96.8%. This superiority aligns logically with Xception's previously noted advantages in visual deepfake detection, particularly its heightened AUC, reflecting enhanced generalization capabilities.

Meanwhile, MFCC demonstrates specific advantages in audio deepfake detection. Hence, we will opt for the Xception-based model in conjunction with the MFCC-utilized model and proceed to undertake a thorough comparison across the three modalities of deep fake detection: visual deep fake detection, audio deep fake detection, and Multi-modal deep fake detection.

This comprehensive evaluation will encompass contrasting MAVDD with techniques assessed on the FakeAVCeleb dataset within each modality.

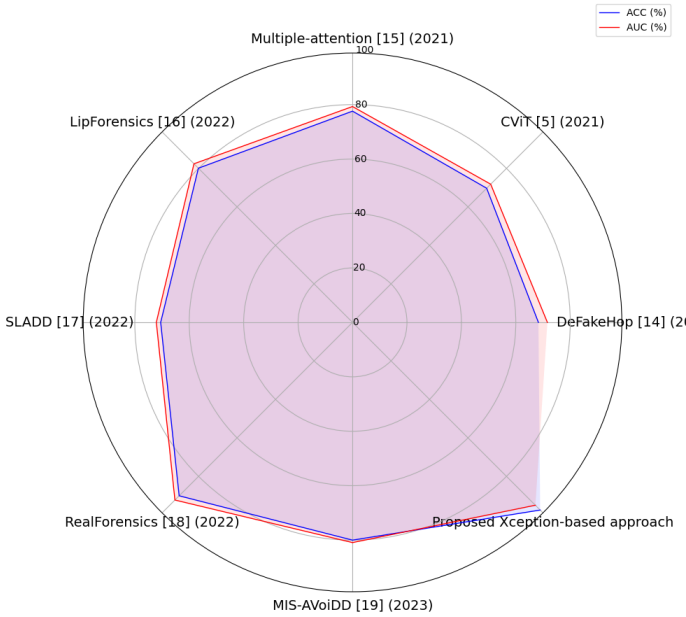

Figure 2: Radar Chart Illustrating the Comparative Analysis of Audio-Only Deepfake Detection on FakeAVCeleb Dataset.

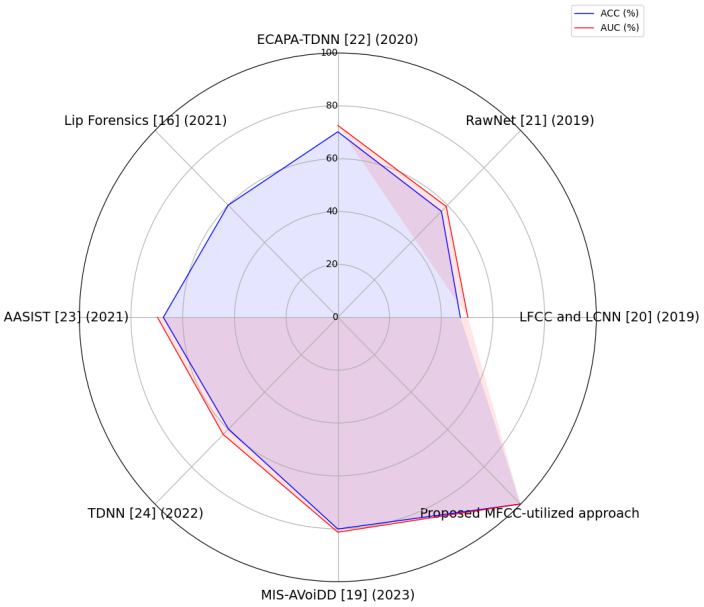

Figure 3: Radar Chart Illustrating the Comparative Analysis of Audio-Only Deepfake Detection on FakeAVCeleb Dataset.

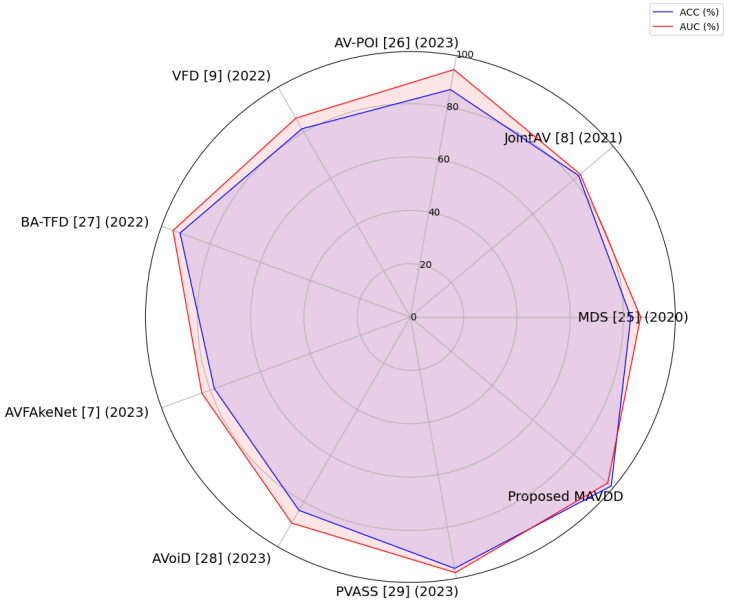

Figure 4: Radar Chart Illustrating the Comparative Analysis of Audio-visual Deepfake Detection on FakeAVCeleb Dataset.

In Figure 2, the radar graph illustrates that our Xception-based approach exhibits superior performance compared to other methods, achieving the highest values for both ACC and AUC metrics. Similarly, the radar graph depicted in Figure 3 demonstrates that our MFCC-utilized approach achieves optimal performance for audio deepfake detection regarding ACC and AUC metrics.

Figure 4 presents a radar graph showcasing the competitive performance of our framework. Notably, our AUC scores align closely with recently published methods like PVASS. Thus, improving AUC on the imbalanced FakeAVCeleb dataset remains challenging due to the data distribution. In the test set, 4,187 videos belong to the majority (fake) class, with 4,130 correctly classified and 57 misclassified. Conversely, the minority (real) class consists of only 100 videos, with 95 correctly classified and five misclassified. This highlights the need for further investigation to enhance the model's ability to generalize and reliably discriminate the real (minority) class. Additionally, regarding ACC scores, our method surpasses all comparative approaches.

This accomplishment extends to both unimodal (visual-only and audio-only) and multimodal (audio-visual) deepfake detection scenarios, establishing MAVDD as a state-of-the-art performance.

4. Conclusion

The growing sophistication of video deepfakes, characterized by manipulated videos accompanied by synchronized synthetic visual and audio elements, poses a growing threat. This prompts research into developing advanced multimodal audio-visual deepfake detectors capable of collectively detecting audio and visual manipulations. Current detectors often rely on the fusion of audio and visual streams; however, due to the heterogeneous nature of these streams, there is a demand for a more sophisticated mechanism for detecting multimodal manipulations. In this study, we introduce an enhanced multimodel audio-visual deepfake detector that leverages optimal preprocessing techniques and deep learning components tailored for both audio and video domains, resulting in enhanced performance compared to existing unimodal and multimodal deepfake detectors, as demonstrated in our comprehensive comparisons. Recognizing the limitations of our current approach, which focuses on video manipulation's visual and audio aspects, we intend to explore emerging facets of video manipulation, such as text, motion, and context, to enhance detection capabilities. Furthermore, we are committed to developing innovative techniques to enhance deepfake detection performance**.**

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139-144, Oct. 2020.

[2] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," arXiv preprint arXiv:2108.05080, Aug. 2021.

[3] N. Nida, A. Irtaza, and N. Ilyas, "Forged face detection using ELA and deep learning techniques," in 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), Islamabad, Pakistan, Jan. 2021, pp. 271-275.

[4] Saikia, P., Dholaria, D., Yadav, P., Patel, V., & Roy, M. (2022). A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

[5] Wodajo, D., & Atnafu, S. (2021, February 22). Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126.

[6] Hao, H., Bartusiak, E. R., Güera, D., Mas Montserrat, D., Baireddy, S., Xiang, Z., Yarlagadda, S. K., Shao, R., Horváth, J., Yang, J., & Zhu, F. (2022). Deepfake detection using multiple data modalities. In Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks (pp. 235-254). Cham: Springer International Publishing.

[7] H. Ilyas, A. Javed, and K. M. Malik, "AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection," Applied Soft Computing, vol. 136, p. 110124, Mar. 2023.

[8] Y. Zhou and S. N. Lim, "Joint audio-visual deepfake detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14800-14809.

[9] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, "Voice-face homogeneity tells deepfake," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 20, no. 3, pp. 1-22, Nov. 11, 2023.

[10] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251-1258.
[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700-4708.
[12] K. S. Tikhonov and M. V. Feigel'man, "Strange metal state near quantum superconductor-metal transition in thin films," Annals of Physics, vol. 417, pp. 168138, Jun. 1, 2020.
[13] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, Jun. 20, 2009, pp. 248-255. IEEE.
[14] H. S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, and C. C. Kuo, "Defakehop: A light-weight high-performance deepfake detector," in 2021 IEEE International Conference on Multimedia and Expo (ICME), Jul. 5, 2021, pp. 1-6. IEEE.
[15] 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, 2021, pp. 2185-2194.
[16] S. A. Shahzad, A. Hashmi, S. Khan, Y. T. Peng, Y. Tsao, and H. M. Wang, "Lip sync matters: A novel multimodal forgery detector," in 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Nov. 7, 2022, pp. 1885-1892. IEEE.
[17] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18710-18719.
[18] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14950-14962.
[19] V. S. Katamneni and A. Rattani, "MIS-AVoiDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection," arXiv e-prints, Oct. 2023, arXiv:2310.
[20] J. Monteiro, J. Alam, and T. H. Falk, "End-to-end detection of attacks to automatic speaker recognizers with time-attentive light convolutional neural networks," in 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), Oct. 13, 2019, pp. 1-6. IEEE.
[21] J. W. Jung, H. S. Heo, J. H. Kim, H. J. Shim, and H. J. Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," arXiv preprint arXiv:1904.08104, Apr. 17, 2019.
[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," arXiv preprint arXiv:2005.07143, May 14, 2020.
[23] J. W. Jung, H. S. Heo, H. Tak, H. J. Shim, J. S. Chung, B. J. Lee, H. J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 23, 2022, pp. 6367-6371. IEEE.
[24] A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva, "Deepfake audio detection by speaker verification," in 2022 IEEE International Workshop on Information Forensics and Security (WIFS), Dec. 12, 2022, pp. 1-6. IEEE.
[25] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in Proceedings of the 28th ACM International Conference on Multimedia, Oct. 12, 2020, pp. 439-447.
[26] 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, 2023, pp. 943-952.
[27] Z. Cai, K. Stefanov, A. Dhall, and M. Hayat, "Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," in 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Nov. 30, 2022, pp. 1-10. IEEE.
[28] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2015-2029, Mar. 27, 2023.
[29] Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, "Pvass-mdd: Predictive visual-audio alignment self-supervision for multimodal deepfake detection," IEEE Transactions on Circuits and Systems for Video Technology, Aug. 29, 2023.