CMDD: A novel multimodal two-stream CNN deepfakes detector

Leonardo Mongelli[1], Luca Maiano[1] and Irene Amerini[1]

[1]Sapienza University of Rome, Department of Computer, Control and Management Engineering A. Ruberti, Via Ariosto 25, Rome, 00185, Italy

Abstract

Researchers commonly model deepfake detection as a binary classification problem, using an unimodal network for each type of manipulated modality (such as auditory and visual) and a final ensemble of their predictions. In this paper, we focus our attention on the simultaneous detection of relationships between audio and visual cues, leading to the extraction of more comprehensive information to expose deepfakes. We propose the *Convolutional Multimodal deepfake detection model (CMDD)*, a novel multimodal model that relies on the power of two Convolution Neural Networks (CNNs) to concurrently extract and process spatial and temporal features. We compare it with two baseline models: DeepFakeCVT, which uses two CNNs and a final Vision Transformer, and DeepMerge, which employs a score fusion of each unimodal CNN model. The multimodal FakeAVCeleb dataset was used to train and test our model, resulting in a model accuracy of 98.9% that places our model in the top 3 ranking of models evaluated on FakeAVCeleb.

Keywords

Deepfake detection, Multimodal deepfake, Misinformation, Multimedia Forensics

1. Introduction

The generation of synthetic data offers several benefits in the industry. Movie companies utilize manipulation tools to alter the appearance of characters during hazardous scenes and to manipulate their voices, while e-commerce companies use them to boost customer purchase speed [1]. Unfortunately, the advancements in AI-generated data also result in strong security issues. The significant use of social media makes it effortless to gather personal data such as photos, videos, and information regarding one's habits and preferences. This data can be modified with AI technologies in a malicious way, leading to trustworthiness problems. Malicious individuals can utilize a person's face and retina to bypass biometric authentications and their personal habits and preferences to pass security questions. Stolen or false identities can be used to open or access bank accounts and commit financial fraud. A recent example dates back to 2020 when scammers stole $35 million from a Japanese company using a sophisticated voice clone created that tricked the branch manager into thinking he was talking to the company

ROMCIR 2024: The 4th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2024: the 46th European Conference on Information Retrieval), March 24, 2024, Glasgow, UK.

mongelli.2020024@studenti.uniroma1.it (L. Mongelli); maiano@diag.uniroma1.it (L. Maiano); amerini@diag.uniroma1.it (I. Amerini)

CEUR Workshop Proceedings (CEUR-WS.org)

director and convinced him to initiate several fund transfers to new accounts[1]. Similarly AI tools may potentially influence social opinion through harmful misinformation campaigns. Examples can be found in several deepfake videos of the Ukrainian President Zelensky[2], or the one of Mark Zuckerberg used to dispute politicians' unwillingness to impose restrictions on significant technology firms such as Facebook[3].

Most deepfake detection methods are typically posed as binary classification problems, where the input is either "real" or "fake" [2]. The most common approach in deepfake detection is the unimodal approach, which refers to techniques that concentrate solely on one modality (i.e., audio or video). This work focuses on the audio and visual modalities, which are the most informative when dealing with deepfake recordings. Unimodal detectors are subjected to relevant issues; for instance, a detector that only processes video data cannot identify acoustic manipulation, while a detector that processes audio data can be readily tricked by manipulating images [3]. Many unimodal models can be combined using an ensemble method. However, in such cases, the model does not analyze the relationship between audio and video cues but considers them independently. Due to the various forms of manipulation used in audio-visual deepfakes, learning features from a single modality can lead to an inaccurate assessment of media authenticity [2]. Introducing a multimodal approach for the simultaneous detection of manipulated modalities allows extracting more comprehensive information to expose deepfakes. However, this approach does not consistently guarantee higher accuracy compared to using a single modality [2], for instance, in cases where one of the two modalities is unavailable. In such situations, two distinct unimodal approaches are generally more practical than using multimodal detection methods [4].

This paper presents the *CMDD* model, a novel multimodal architecture for deepfake detection based on the concurrent use of two Convolutional Neural Networks for audio-visual features extraction and a last shared convolutional block used for further processing and a final multi-class classification. We utilize an effective strategy to extract relevant features from the input data using the Mel-Spectrogram information for the audio part and spatiotemporal characteristics between frames in the visual component. We also develop two multimodal baselines named *DeepFakeCVT* and *DeepMerge*: the former is similar to the CMDD model, but it utilizes a shared Vision Transformer block instead of another convolution step for multimodal classification, while the latter is an ensemble method of two unimodal audio and video models. We evaluate the proposed model using a perfectly balanced version of the FakeAVCeleb deepfake dataset such that the number of videos, gender, age, ethnic backgrounds, and manipulation techniques are balanced in each class. In our experiments, the proposed solution achieves good performance, reaching an accuracy value of 98.9%, which places our model in first position in the state-of-the-art (SOTA) ranking if we consider methods trained and tested on FakeAVCeleb using our balancing technique and in third position when it comes to all possible versions of the dataset using any techniques. Our main contributions are summarized below.

- We propose a novel multimodal deepfake architecture named the *CMDD* model, which

[1] https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

[2] https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/

[3] https://www.independent.co.uk/tech/mark-zuckerberg-deepfake-ai-meta-b2236388.html

comprises two branches for audio and visual feature extraction and a shared convolutional block for further processing and final multi-class classification.

- We train and test our model using a balanced version of the FakeAVCeleb dataset, a benchmark dataset that contains both auditory and visual manipulation;
- To the best of our knowledge, our CMDD model achieves a SOTA performance with an accuracy value of 98.9% that places our model in the top 3 ranking of models evaluated on the FakeAVCeleb dataset.

The rest of this paper is organized as follows. Section 2 proposes a detailed analysis of the existing unimodal and multimodal state-of-the-art models. Section 3 delves into our proposed model and the pre-processing stage. Section 4 reports our experimental setup, describing our proposed baselines and comparative results between our models and state-of-the-art solutions. Finally, Section 6 draws the conclusions of this work and gives an overview of possible future improvements to our work.

2. Related work

This section presents various state-of-the-art deepfake detection techniques based on two approaches: unimodal and multimodal detectors.

2.1. Unimodal deepfake detection

A first approach to classifying audio deepfakes is using machine learning (ML) models. Singh et al. [5] use a Quadratic Support Vector Machine (Q-SVM) model to differentiate artificial speech from humans, taking advantage of the Cepstral and Bi-spectral analysis. The Mel Cepstral analysis of the audio allows the detection of significant power components in natural speech absent from the AI-synthesised speech, while the Bi-spectral analysis is used for the opposite. A remarkable comparison between ML and deep learning (DL) models for audio deepfake detection is done in Liu et al. [6]. Specifically, the authors utilize an SVM with Mel-frequency cepstral coefficient (MFCC) characteristics and a standard Convolutional Neural Network (CNN) with five convolutional layers. This analysis shows that CNN is more reliable than SVM despite their similar results. Therefore researchers thus turn their attention to the use of CNN architectures. Wani et al. [7] propose a CNN with four convolutional blocks and two fully connected layers, and apply transfer learning on two pre-trained models (i.e., VGG16 and MobileNet), to detect inconsistencies in the frequency domain of mel-spectrogram input images. Instead in Wijethunga et al. [8], the authors combine the CNN architecture with a Recurrent Neural Network (RNN) to leverage their specific qualities: identification of long-term dependencies in temporal variations with the RNN while the potential of feature extraction capabilities of the CNN. Arif et al. [9] introduce the ELTP-LFCC, a novel audio feature descriptor, by combining the linear frequency cepstral coefficients (LFCC) with the local ternary pattern (ELTP). Differently from the previous unimodal audio approaches we implement the C4N model, a CNN-based architecture with 4 convolutional layers able to extract the audio track from the video, convert it into a digital signal, generate its correspondent mel-spectrogram image and analyze possible inconsistencies in the frequency spectrum. The architecture is inspired by Wani et al. [7] with strong modification at the layer level.

For what concerns video deepfake detection methods, a large group of works focuses on detecting local texture inconsistencies caused by applying manipulation techniques. For instance, Hu et al. [10] propose a novel frame inference framework named FInfer that uses an encoder to acquire facial representations for current and future frames and an auto-regressive model to predict future faces using the encoder knowledge. Instead of analyzing facial inconsistencies, Nirkin et al. [11] also focus on the region around the face. Specifically, the authors notice that the deepfake generator modifies the face region during manipulation, leaving the context untouched. Therefore, they compare the resulting two identity vectors to detect discrepancies using context recognition and face identification networks. A different approach is taken in Maiano et al. [12], where the authors present a DepthFake method that uses depth maps to classify deepfake videos. They use the FaceDepth detector to estimate the depth of human faces and a pre-trained Xception network to classify each frame of the recordings. A method named ID-Reveal is presented in Cozzolino et al. [13]. The proposed model uses metric learning in conjunction with an adversarial training strategy to learn temporal facial features peculiar to a person's movements during speech. They utilize a CNN architecture composed of a facial feature extractor, a temporal ID network to detect bio-metric anomalies, and a generative adversarial network to predict each peculiar person's motions. The use of Visual Transformers in video deepfake detection is proposed by Khan et al. [14], where the authors leverage incremental learning. Their model employs an XceptionNet as an image feature extractor, a Single Stage Detector (SSD) to extract and crop faces frame after frame, and a 3D Dense Face Alignment (3DDFA) model to generate UV texture maps from face images. They use face pictures and UV texture maps to extract the image characteristics since they present complementary information. Instead, in our video detection approach we utilize a ResNet18-3D model that can remove the frames from the input video, concatenate them into a unique tensor, and extract the necessary information to detect inconsistencies between spatiotemporal features.

2.2. Multimodal deepfake detection

Numerous studies demonstrate how merging several modalities can yield to better conclusions and complementary information. Many modalities, such as facial signals, speech cues, background context, hand gestures, and body position and orientation, can be extracted from a video, even to detect deepfake content and can be combined to determine the authenticity of a particular video [15]. In Asha et al. [16], the authors propose the "D-Fence" layer, which is an ensemble deepfake detection method of two unimodal networks: one detects artificial face information using a VGG16 model and an optical flow estimator while the other one extracts the audio information using the Mel Frequency Cepstral Coefficients (MFCC) temporal feature vector which is then fed into a VGG16 model. The dissonance audio-video cues are efficiently detected by the two cross-modal networks. Zhang et al. [17] also follow the ensemble approach. The authors separately extract the audio and visual information from the deepfake videos and then compare the results. They employ an RNN network for the audio portion and a 3D-ResNet [18] for the visual portion, while the classification is actuated using an adaptive modality dissonance score (MDS) [15] criterion. Instead, authors in [19] extend the applicability of [17] to cases where one modality is missing.

Instead of presenting an ensemble approach based on Convolutional Neural Networks, Hashmi et al. [3] propose an Audio-Visual Transformer-based Ensemble Network (AVTENet) architecture that takes into account both visual and audio manipulations. Specifically, AVTENet combines three exclusively transformer-based networks, integrated with pre-trained models through supervised and self-supervised learning, to identify significant cues in audio, video, and audio-visual modalities. In Cozzolino et al. [20] authors present a Person of Interest (POI) deepfake detector which relies on biometric features. This results in an extremely accurate detector because each individual possesses unique biometric features unlikely to be replicated by a synthetic generator.

Instead of focusing on biometric characteristics, authors in Mittal et al. [15] rely on emotional behaviors and features. They suggest a deep learning network inspired by triplet loss and Siamese network architecture. To determine whether an input video is genuine, they extract and compare affective signs from the auditory and visual modalities within the video that correspond to the perceived emotion. Another study focuses on the lip movements of people speaking. Specifically, in Shahzad et al. [21] a revolutionary lip-reading-based multimodal deepfake detection technique is proposed. The idea is that synthetic lip sequences are frequently out of sync with their audio stream, so one good indicator of whether or not the video information has been altered is the disparity in lip movements. A Multi-modal Multi-scale TRansformer (M2TR) is presented in Wang et al. [22], where the authors utilize a two-stream approach in which the frequency stream uses learnable frequency filters to filter out forgery features in the frequency domain and the RGB stream uses several scales to catch inconsistencies between different regions within an image.

In contrast to previous works, our multimodal approach consists of extracting and processing each video's audio and visual features using two branches. The audio branch is responsible for detecting inconsistencies in the frequency spectrum of the mel-spectrogram input image, while the video branch is responsible for detecting spatiotemporal irregularities on a frame-by-frame basis. Both branches are characterized by a CNN-based model used for feature extraction. The features from each branch are then concatenated and fed into a final shared convolutional block that simultaneously elaborates the information and performs the final multi-label classification.

3. Method

This section describes our proposed *Convolutional Multimodal deepfake detection model (CMDD).* In our approach, we simultaneously consider the information related to audio and video modalities. The proposed architecture, depicted in Figure 1, comprises a different stream for each modality. Each of them extracts the feature vector from the input data using the power offered by the CNNs to investigate the temporal and spatial locality characteristics. Then, the extracted vectors are concatenated, and the resulting tensor is given to another CNN block with a final fully connected (FC) layer for the classification.

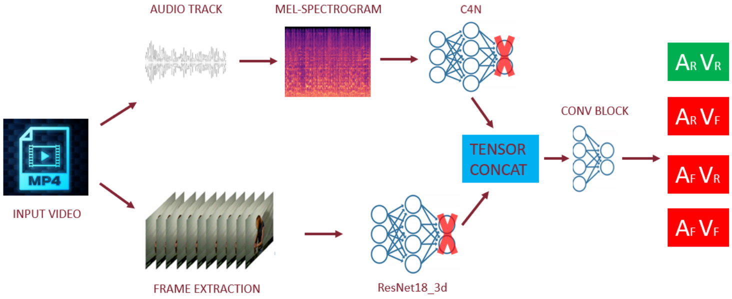

Figure 1: Our proposed Convolutional Multimodal deepfake detection (CMDD) architecture. Label legend: *ArVr* means real audio real video, *ArVf* real audio fake video, *AfVr* fake audio real video, and *AfVf* fake audio and video.

3.1. Preprocessing

Due to its multimodal nature, the proposed model requires a pre-processing phase to detect and extract the necessary information in videos. One network branch is structured to deal with audio information while the other requires visual features. Therefore, the input has to be adapted before being fed into the two audiovisual blocks. Specifically for the auditory part, we extract the audio track from the input video using the FFmpeg[4] media converter and represent it as a waveform digital signal. This signal is converted from the time to the frequency domain using the Fast Fourier Transform (FFT), which is applied to signal segments using shifting windows. The result is a sort of heat map named spectrogram, a visual representation of how the signal's amplitude varies over time at different frequencies. The obtained spectrogram is finally converted from a logarithmic scale to a mel scale using Equation 1, where m indicates the mel scale and f refers to the frequency.

$$m = 2595 * log_{10}(1 + \frac{f}{700}) \tag{1}$$

Regarding the visual part, we divide the input video into frames again using the FFmpeg converter, and then all the obtained frames are stacked into a unique tensor. All mel spectrograms and frames are resized to 224x224 pixels, converted into a tensor, and finally normalized.

3.2. Architecture

Our proposed architecture consists of the audio and the video pipeline. The former uses a *C4N model* consisting of a standard CNN architecture with 4 convolutional blocks. Each block comprises a 2D convolutional layer with a rectified linear unit (ReLU) activation function and a *batchnorm2d* normalization. No adaptive pooling layers or linear classifiers are used. The audio network takes in input tensors containing the information of each mel spectrogram given in an RGBA format. Instead the latter uses a pre-trained ResNet18 [23] adapted for 3D input data to handle sequences of images. Like the audio network, it has no fully connected layers for

[4] https://ffmpeg.org/

classification; therefore the resulting output is a feature vector. The two feature vectors obtained from the audio and video pipelines are then concatenated, and the resulting tensor is finally fed into another CNN block with a 3D convolutional layer, ReLu function, a BatchNorm3D layer, and a last FC layer for the final 4-class classification: label 0 ($ArVr$) for videos classified with both fake audio and fake video, label 1 ($ArVf$) for real audio and fake video, label 2 ($AfVr$) for fake audio and real video, and label 3 ($AfVf$) for both real audio and real video.

4. Experimental settings

This section describes the dataset and implementation details, compares results with state-of-the-art architectures, and highlights differences.

4.1. Dataset

We train and evaluate our architecture using the FakeAVCeleb [24] dataset. Most available datasets are not structured to deal with multimodal deepfake detection. For instance, datasets such as FaceForensics++[25], Deep-FakeTIMIT [26] or KoDF [27] contain manipulated content exclusively on the visual modality. Instead, FakeAVCeleb is created specifically for multimodal audio-visual studies. It is a multimodal, age and gender-balanced dataset with YouTube videos taken from the VoxCeleb2 [28] dataset by selecting videos of 500 celebrities with five racial backgrounds (Caucasian (American), Caucasian (European), African, Asian (East), and Asian (South)). All the speeches are in English, but the different ethnic backgrounds introduce diversity in phonemes and accents partially removing racial bias issues. Each video is a clean recording with only one person's frontal face without occlusion. Additionally, it covers several Deepfake generation techniques, allowing for a better generalization with various detection methods. In particular, the dataset contains 500 real videos and 19500 deepfake ones, with a ratio between fake and real videos equal to 1:39. The provided dataset is extremely unbalanced. Therefore, we remove several deepfakes videos until we reach a perfectly balanced ratio, maintaining the balance in the number of forged videos with the same manipulation technique and number of videos for each class and equality of gender, age, and ethnic backgrounds. The resulting dataset contains videos characterized by different lengths and frame rates. This difference may cause issues during training. As a result, we tested different unimodal models on both the entire video and the first second of the recording. The results were fairly similar, but using the entire video requires significantly more computational power. Thus, we take the decision to consider just the first second of the videos. Therefore, in the pre-processing phase explained in Section 3.1, we extract the first second of audio and the first 25 frames from each video. We set 25 frames because it is the dataset's most common frame rate value, and thus, it is around the number of frames in one second of video.

4.2. Baselines

We develop the CMDD model using the knowledge gained from implementing DeepFakeCVT and DeepMerge baselines.

- **DeepFakeCVT.** The DeepFakeCVT has a similar structure to the CMDD model; the only difference is in the classification block. DeepFakeCVT uses a CCT-3D transformer instead of another convolution block to speed up the computational efficiency of the model.
- **DeepMerge.** The DeepMerge model is instead an ensemble method that directly performs a fusion of the score of each branch of the CMDD model, where the two unimodal networks have the last FC layers for classification.

4.3. Implementation details

We use the proposed baseline architectures to find the best settings for our multimodal architecture. We train our models using AdamW optimizer as an optimization algorithm, the cross-entropy as a loss function, and a learning rate equal to 1e-5. The necessary number of training epochs without introducing overfitting is around 12. We train all models on the Nvidia 1080 Ti with 11GB of memory. Due to the low available computational power, we must set the batch size to 8, specifically, 4 samples show audio information, and the other 4 represent visual features.

4.4. Evaluation metrics

We evaluate our results based on the following metrics.

- **Accuracy.** It represents the ratio of the correct predictions over the whole predicted sample. Calling TP, TN, FP, and FN the *true positive*, *true negative*, *false positive*, and *false negative* examples respectively, we can calculate the accuracy as $\frac{TP+TN}{TP+TN+FP+FN}$.
- **Precision.** It represents the ratio of the correct predictions over the whole correct samples and can be calculated as $\frac{TP}{TP+FP}$.
- **Recall.** It is the ratio of correct predictions for a class to the total number of cases in which it occurs, and is computed as $\frac{TP}{TP+FN}$.
- **F1-score.** It is the Harmonic mean between Precision and Recall and is obtained as $2 * \frac{Precision*Recall}{Precision+Recall}$.
- **Confusion matrix.** It is a matrix that visually represents the distribution of predictions among the classes.

Some metrics are then adapted to deal with more than 2 classes by introducing the concept of macro and micro metrics. Precision-Macro is computed by using the average precision for each predicted class, Recall-Macro is by using the average recall for each actual class, and Precision-Micro and Recall-Micro are by considering all the samples independently from their class. Finally, the F1-Macro score and F1-Micro score are respectively computed using the Macro and Micro metrics of Precision and Recall [29].

5. Experimental Results

In this section, we propose an analysis of the performance of our proposal. Our approach reaches promising results compared to our baselines and the accuracy values of other state-of-the-art

Table 1
Results of the developed multimodal deepfake detection networks. Label legend: **M** stands for **Macro** while **m** for **micro**.

Model	*Accuracy*	*Precision-M*	*Precision-m*	*Recall-M*	*Recall-m*	*F1-M*	*F1-m*
DeepFakeCVT	0.9608	0.9611	0.9608	0.9607	0.9607	0.9606	0.9606
DeepMerge	0.9731	0.9732	0.9731	0.9731	0.9731	0.9730	0.9730
CMDD (proposed)	**0.9895**	**0.9897**	**0.9895**	**0.9895**	**0.9895**	**0.9895**	**0.9895**

(SOTA) methods. Table 1 shows the performance of the proposed CMDD model and the two provided baselines DeepFakeCVT and DeepMerge. It is evident that the CMDD model is the architecture with the best results in all the considered metrics. It achieves a 98.9% accuracy, therefore outperforming the two baselines. Figure 2 shows the confusion matrix related to the CMDD model. It contains just 7 misclassified samples over 668 total examples. Specifically, 3 samples are predicted as FakeVideo-FakeAudio (label 0) instead of FakeVideo-RealAudio (label 1), and the other 3 as RealVideo-FakeAudio (label 2) instead of RealVideo-RealAudio (label 3). Just one recording that belongs to the actual FakeVideo-RealAudio (label 1) class is incorrectly predicted as RealVideo-RealAudio (label 3). Instead, Figure 3 reports the trend of the CMDD's training loss, which rapidly and smoothly decreases till convergence without showing relevant issues.

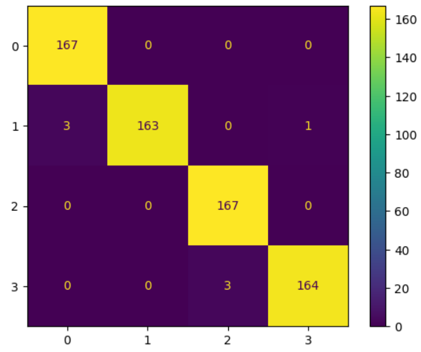

Figure 2: Confusion matrix related to the CMDD model.

The high quality of our results is demonstrated by making comparisons with the SOTA models trained and tested using the FakeAVCeleb dataset reported in Table 2. Our proposed CMDD model takes third place in the SOTA leaderboard. Only DLC [4] and S-Capsule Forensics [2] beat our CMDD and DeepMerge models, while DeepFakeCVT is also beaten by the MIS-AVoiDD [30] model. Differently from our proposed method, the authors of DLC [4] apply a different input processing that consists of extracting the Short Time Fourier Transform signal (STFT) from the input audio track in videos. The STFT gives information about the frequency trend of the extracted signal over time. Instead, we deal with the same frequency information by displaying it in a clever colored image, the mel-spectrogram. Furthermore, we obtain similar results just using the first second of the audio track instead of considering the full audio track. In addition, we represent the frequency information into a single image while [4] considers 4 concatenated

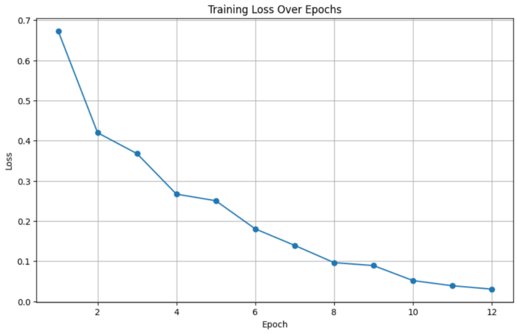

Figure 3: Training loss related to the CMDD model.

Table 2
Comparisons with state-of-the-art models on FakeAVCeleb dataset

Position	*Model*	*Year*	*Accuracy*
1	DLC [4]	2023	0.997
2	S-Capsule Forensics [2]	2023	0.992
3	**CMDD**	**2023**	**0.989**
4	**DeepMerge**	**2023**	**0.973**
5	MIS-AVoiDD [30]	2021	0.962
6	**DeepFakeCVT**	**2023**	**0.960**
7	PVASS-MDD [31]	2023	0.957
8	AVAD [31]	2023	0.942
9	AV-Lip-Sync [21]	2022	0.940
10	AVFakeNet [32]	2023	0.934
11	Multimodaltrace [33]	2023	0.929

images for each second of video, increasing the memory and GPU requirements. In their paper, the authors use a pre-training step to enhance the capabilities of their architecture, while our model is trained from scratch. We also reach comparable results using fewer training epochs, 12 instead of over 100, and batch size, 8 instead of 12. We achieve a quite similar performance despite our limited computing power, thus our model may outperform DLC [4] with equal computational capacity. Instead, S-Capsule Forensics [2] uses mel spectrograms as audio input, referring to the first 4 seconds of the recording instead of the first second as we do. Furthermore, a different visual feature processing is applied. They utilize all the frames of each video, while in our project we set the number of extracted frames to 25 per video. In this way, we have the same number of extracted images, and the network does not consider the length of the recordings as a relevant feature for classification. As well as in [4], we get similar results in accuracy using a lower number of training epochs and batch size. They use 50 epochs and a batch size of 10. Also in the MIS-AVoiDD [30] method, researchers use another type of auditory input. Instead of using the Short Time Fourier Transform (STFT) signal or the Mel spectrogram as we do, they implement an architecture capable of handling the Mel Frequency Cepstrum

Coefficient (MFCC). For the visual part, they use 300 frames per video in the training step and 180 in the testing step. Instead, we obtain higher performances by using 26 frames for each video, specifically 1 image for the audio and 25 for the visual information. Our networks are also shallower than their model, allowing us to quickly extract and use the necessary information for efficient training, without extracting too many features that may introduce some bias.

Further differences in the type of input also appear in PVASS-MDD [31] and in Multimodal-trace [33], where the authors utilize a spectrogram and a Short Time Fourier Transform (STFT) signal respectively. These two papers use unnecessary information in audio and visual features and apply deeper networks than ours. Nevertheless, we beat both models, and even AVFakeNet [32], using less informative content and shallow networks with lower batch size and learning rate values, but the same optimizer and loss function.

In Table 2, all the reported models use the same dataset, but some differences appear in the balancing phase. Some authors correct the unbalanced characteristic of the dataset by adding videos from another small dataset, as done in Shahzad et al. [21], and by using augmentation techniques like horizontal and vertical flipping, blurring, salt and pepper noise, as done in Ilyas et al. [32]. For instance, both preprocessing methods are used by the authors of Raza et al. [33]. In other papers like Yu et al. [4], the authors split the dataset according to the manipulation technique used to forge videos, run the test only on each group separately, and finally make an average of the resulting evaluation metric values. In a few cases, the authors decide to reduce the imbalance of the dataset by removing the deepfakes in excess. This is the case of Shahzad et al. [21] and Hashmi et al. [34]. For this reason, we decided to explore this open branch by applying the same balancing technique, and the resulting ranking of state-of-the-art models is reported in Table 3, where our CMDD model is in the first place, with our two baselines following as the second and third models in the leaderboard.

Table 3
Comparisons with state-of-the-art models on the balanced FakeAVCeleb dataset

Position	*Model*	*Year*	*Accuracy*
1	**CMDD**	**2023**	**0.989**
2	**DeepMerge**	**2023**	**0.973**
3	**DeepFakeCVT**	**2023**	**0.960**
4	AV-Lip-Sync [21]	2022	0.940
5	MFD-Ensemble [34]	2022	0.790

The model closest to the models we proposed in terms of accuracy is the AV-Lip-Sync introduced by Shahzad et al. [21]. In their paper, the authors use a different pipeline for audio information. They use a Wav2Lip tool to generate synthetic lip sequences based on the audio track of each video. Thus, they analyze the two lip sequences, the generated and the real, and detect manipulated videos by looking at the potential mismatch between them. Also, their selected hyper-parameters are different from ours. They use Adam optimizer, a learning rate of 0.0002, and batch size of 32 instead of AdamW, 1e-5 and 8 respectively, as done in our proposed networks. Instead, Hashmi et al. [34] propose an ensemble model that uses different CNN-based networks for the audio and visual feature extraction. Their choice of hyper-parameters is different because they use Adam optimizer, the learning rate of 0.001, cross-entropy loss and a batch size different for each network: 512 for the audio and 64 for the video network. Their results are extremely poor concerning the accuracy values obtained by our models.

6. Conclusions and Future Works

This paper proposes a novel multimodal deepfake detection method called *CMDD*. It allows for the simultaneous detection of forgeries in both auditory and visual channels using two CNN networks for feature extraction and a shared convolutional block for the final multi-class classification. We also implement two different baselines named DeepFakeCVT and DeepMerge to highlight the powerful capabilities of the CMDD model. The CMDD model outperforms both architectures using a shallower network than DeepFakeCVT, which uses a deeper Vision Transformer and a single classification block instead of two as in DeepMerge. CMDD also outperforms several state-of-the-art (SOTA) models, achieving an accuracy score of 98.9%, which places it first in the SOTA ranking when considering methods trained and tested on FakeAVCeleb using our same balancing technique and third when considering all possible versions of the dataset.

In the future, we can enhance our architecture by using more powerful model training and testing tools. For example, increasing batch size, learning rate, or epochs can help us evaluate if our proposed model outperforms the DLC [4] network. Furthermore, the proposed architecture showed a rapid tendency to overfit due to the high number of layers and the dataset type. We plan to expand this study to larger datasets for a more robust evaluation. It may also be interesting to better explore the potential offered by our DeepFakeCVT model by removing the two initial CNNs, and thus reducing the depth of the network by several convolutional steps, but also to test other concatenation and fusion techniques. Future work will, therefore, be oriented towards studying different fusion techniques between the two modalities.

Acknowledgments

This study has been supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and Sapienza University of Rome project “EV2” (003_009_22).

References

[1] S. R. Ahmed, E. Sonuç, M. R. Ahmed, A. D. Duru, Analysis survey on deepfake detection and recognition with convolutional neural networks, in: 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2022, pp. 1–7. doi:10.1109/HORA55278.2022.9799858.

[2] S. Muppalla, S. Jia, S. Lyu, Integrating audio-visual features for multimodal deepfake detection, 2023. arXiv:2310.03827.

[3] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, H.-M. Wang, Avtenet: Audio-visual transformer-based ensemble network exploiting multiple experts for video deepfake detection, 2023. arXiv:2310.13103.

[4] C. Yu, P. Chen, J. Tian, J. Liu, J. Dai, X. Wang, Y. Chai, S. Jia, S. Lyu, J. Han, A unified framework for modality-agnostic deepfakes detection, 2023. arXiv:2307.14491.

[5] A. K. Singh, P. Singh, Detection of ai-synthesized speech using cepstral and bispectral statistics, 2021. arXiv:2009.01934.

[6] T. Liu, D. Yan, R. Wang, N. Yan, G. Chen, Identification of fake stereo audio using svm and cnn, Information 12 (2021) 263. doi:10.3390/info12070263.

[7] T. M. Wani, I. Amerini, Deepfakes audio detection leveraging audio spectrogram and convolutional neural networks, in: G. L. Foresti, A. Fusiello, E. Hancock (Eds.), Image Analysis and Processing – ICIAP 2023, Springer Nature Switzerland, Cham, 2023, pp. 156–167.

[8] R. Wijethunga, D. Matheesha, A. A. Noman, K. D. Silva, M. Tissera, L. Rupasinghe, Deepfake audio detection: A deep learning based solution for group conversations, 2020 2nd International Conference on Advancements in Computing (ICAC) 1 (2020) 192–197. URL: https://api.semanticscholar.org/CorpusID:232071547.

[9] T. Arif, A. Javed, M. Alhameed, F. Jeribi, A. Tahir, Voice spoofing countermeasure for logical access attacks detection, IEEE Access PP (2021) 1–1. doi:10.1109/ACCESS.2021.3133134.

[10] J. Hu, X. Liao, J. Liang, W. Zhou, Z. Qin, Finfer: Frame inference-based deepfake detection for high-visual-quality videos, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 951–959. URL: https://ojs.aaai.org/index.php/AAAI/article/view/19978. doi:10.1609/aaai.v36i1.19978.

[11] Y. Nirkin, L. Wolf, Y. Keller, T. Hassner, Deepfake detection based on the discrepancy between the face and its context, 2020. arXiv:2008.12262.

[12] L. Maiano, L. Papa, K. Vocaj, I. Amerini, Depthfake: a depth-based strategy for detecting deepfake videos, ArXiv abs/2208.11074 (2022). URL: https://api.semanticscholar.org/CorpusID:251741027.

[13] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, L. Verdoliva, Id-reveal: Identity-aware deepfake video detection, 2021. arXiv:2012.02512.

[14] S. A. Khan, H. Dai, Video transformer for deepfake detection with incremental learning, 2021. arXiv:2108.05307.

[15] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: An audio-visual deepfake detection method using affective cues, 2020. arXiv:2003.06711.

[16] A. S, P. Vinod, I. Amerini, V. Menon, A novel deepfake detection framework using audio-video-textual features, 2022. URL: https://doi.org/10.21203/rs.3.rs-2390408/v1. doi:10.21203/rs.3.rs-2390408/v1.

[17] Y. Zhang, X. Li, J. Yuan, Y. Gao, L. Li, A deepfake video detection method based on multi-modal deep learning method, in: 2021 2nd International Conference on Electronics, Communications and Information Technology (CECIT), 2021, pp. 28–33. doi:10.1109/CECIT53797.2021.00014.

[18] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, 2018. arXiv:1711.09577.

[19] K. Chugh, P. Gupta, A. Dhall, R. Subramanian, Not made for each other- audio-visual

dissonance-based deepfake detection and localization, 2021. arXiv:2005.14405.

[20] D. Cozzolino, A. Pianese, M. Nießner, L. Verdoliva, Audio-visual person-of-interest deepfake detection, 2023. arXiv:2204.03083.

[21] S. A. Shahzad, A. Hashmi, S. Khan, Y.-T. Peng, Y. Tsao, 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), 2022, pp. 1885–1892. doi:10.23919/APSIPAASC55919.2022.9980296.

[22] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, S.-N. Lim, Y.-G. Jiang, M2tr: Multi-modal multi-scale transformers for deepfake detection, 2022. arXiv:2104.09770.

[23] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, CoRR abs/1711.11248 (2017). URL: http://arxiv.org/abs/1711.11248. arXiv:1711.11248.

[24] H. Khalid, S. Tariq, M. Kim, S. S. Woo, Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022. arXiv:2108.05080.

[25] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, 2019. arXiv:1901.08971.

[26] P. Korshunov, S. Marcel, Vulnerability assessment and detection of deepfake videos, in: 2019 International Conference on Biometrics (ICB), 2019, pp. 1–6. doi:10.1109/ICB45273.2019.8987375.

[27] P. Kwon, J. You, G. Nam, S. Park, G. Chae, Kodf: A large-scale korean deepfake detection dataset, 2021. arXiv:2103.10094.

[28] J. S. Chung, A. Nagrani, A. Zisserman, Voxceleb2: Deep speaker recognition, in: Interspeech 2018, interspeech 2018, ISCA, 2018, pp. 1086–1090. URL: http://dx.doi.org/10.21437/Interspeech.2018-1929. doi:10.21437/interspeech.2018-1929.

[29] M. Grandini, E. Bagli, G. Visani, Metrics for multi-class classification: an overview, 2020. arXiv:2008.05756.

[30] V. S. Katamneni, A. Rattani, Mis-avoidd: Modality invariant and specific representation for audio-visual deepfake detection, 2023. arXiv:2310.02234.

[31] Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao, A. C. Kot, Pvass-mdd: Predictive visual-audio alignment self-supervision for multimodal deepfake detection, IEEE Transactions on Circuits and Systems for Video Technology (2023) 1–1. doi:10.1109/TCSVT.2023.3309899.

[32] H. Ilyas, A. Javed, K. M. Malik, Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection, Applied Soft Computing 136 (2023) 110124. URL: https://www.sciencedirect.com/science/article/pii/S1568494623001424. doi:https://doi.org/10.1016/j.asoc.2023.110124.

[33] M. Anas Raza, K. Mahmood Malik, Multimodaltrace: Deepfake detection using audiovisual representation learning, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 993–1000. doi:10.1109/CVPRW59228.2023.00106.

[34] A. Hashmi, S. A. Shahzad, W. Ahmad, C. W. Lin, Y. Tsao, H.-M. Wang, Multimodal forgery detection using ensemble learning, in: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2022, pp. 1524–1532. doi:10.23919/APSIPAASC55919.2022.9980255.