Multimodal Forgery Detection Using Ensemble Learning

Ammarah Hashmi*†, Sahibzada Adil Shahzad*‡, Wasim Ahmad*‡, Chia Wen Lin¶, Yu Tsao§, Hsin-Min Wang*
* Social Networks and Human Centered Computing Program, Taiwan International Graduate Program,
Institute of Information Science, Academia Sinica
E-mail: {adilshah275, was_last, whm}@iis.sinica.edu.tw
† Institute of Information Systems and Applications, National Tsing Hua University, Taiwan
E-mail:hashmiammarah0@gmail.com
¶ Department of Electrical Engineering, National Tsing Hua University, Taiwan
E-mail: cwlin@ee.nthu.edu.tw
‡ Department of Computer Science, National Chengchi University, Taipei, Taiwan
§ Research Center for Information Technology Innovation, Academia Sinica, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw

***Abstract*—The recent rapid revolution in Artificial Intelligence (AI) technology has enabled the creation of hyper-realistic deepfakes, and detecting deepfake videos (also known as AI-synthesized videos) has become a critical task. The existing systems generally do not fully consider the unified processing of audio and video data, so there is still room for further improvement. In this paper, we focus on the multimodal forgery detection task and propose a deep forgery detection method based on audiovisual ensemble learning. The proposed method consists of four parts, namely a Video Network, an Audio Network, an Audiovisual Network, and a Voting Module. Given a video, the proposed multimodal and ensemble learning system can identify whether it is fake or real. Experimental results on a recently released multimodal FakeAVCeleb dataset show that the proposed method achieves 89% accuracy, significantly outperforming existing models.**

I. Introduction

The widespread use of smartphones, smart digital devices, and social media has brought a vast amount of online audio, image, and video content. At the same time, recent improvements in machine learning techniques have significantly advanced the capabilities of multimedia (speech, audio, image, and video) processing. Although it has brought convenience to human life, this advancement has also created a new serious problem, which is deepfake. Deepfake refers to digitally altering the media content, e.g., swapping the face of one person with the face of another person in a video clip or altering the person's speech in an audio clip. Several applications of deepfakes have positive contributions to education, art, innovation, movie or film production, artist/actor expression and criminal forensics [1]. However, deepfakes can be potentially harmful if used for malicious purposes, such as political denigration, personal defamation, revenge porn, blackmailing someone or to spreading misinformation [2].

Currently, many AI-based techniques allow users to easily swap faces and expressions, alter lip movements, or morph speech. For example, Face2Face [3], FaceSwap, DeepFakes and Neural Textures [4] are well-known techniques for synthesizing fake multimedia data. Deep generative models, such as Generative Adversarial Networks (GANs) [5] [6] and Variational AutoEncoders (VAEs) [7], have been widely used to generate high-quality forged data. Due to its very real nature, fake multimedia data is increasingly difficult to distinguish from real data. Therefore, deepfakes can easily lead to threatening results. How to effectively solve this problem has become an emerging and important task.

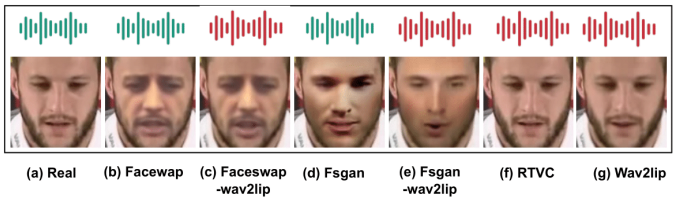

Fig. 1. A real sample and its various manipulated samples from the FakeAVCeleb Dataset. Green audio waveforms represent real audio, while red audio waveforms represent manipulated audio.

Thus far, many machine learning approaches have been proposed for forgery detection [8] [9] [10]. For example, many algorithms are derived to detect forgery in images [8] [9] [4] [11] [12] [13]. For speech, algorithms have been proposed to distinguish between genuine and spoofed audio clips [14] [15]. Similarly, for video, some algorithms toil to detect forged video content [3] [16] [17] [18] [19] [12]. In this study, we focus on the audiovisual video forgery detection task.

Current video forgery detection methods mainly focus on behavior inconsistency, such as eye blinking [20], facial expression [21], head pose [22], lip movement [23], body motion [24], and mismatch between left and right eyes [25]. Obviously, most of these existing methods are unimodal methods using audio modality or visual modality. Moreover, most public datasets focus more on visual manipulations than audio manipulations. We argue that existing techniques may not be

perfect designs for detecting multimodal manipulations.

In this work, we propose a novel multimodal and ensemble learning system for forgery detection. The system consists of four parts, a **Video Network**, an **Audio Network**, an **Audiovisual Network**, and a **Voting Module**. The audio and video networks are unimodal networks that process audio and video data, respectively, to make real/fake predictions. The **Audiovisual Network**, on the other hand, processes multimodal (audio and visual) data for real/fake prediction. Then, the decision-making module combines the predictions of the above three networks to make the final real/fake prediction.

We test the proposed system on the recently released FakeAVCeleb dataset [17]. Fig. 1 shows some samples from the FakeAVCeleb dataset, including a real sample and its various manipulations. Our experimental results show that the proposed system outperforms the baseline [16] reported on the same dataset, with a notable accuracy improvement of 11% (from 78% to 89%). The key contributions of this study are as follows:

- We propose a novel multimodal learning framework that exploits audio and visual information for effective forgery detection.
- We design an ensemble learning framework to combine the predictions from audio-only, video-only, and audiovisual networks.
- To the best of our knowledge, the proposed system achieves state-of-the-art performance on the FakeAVCeleb dataset.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the proposed system. Section IV reports the experimental setup and results. Finally, Section V provides the concluding remarks.

II. Related Work

The word "Deepfake" is a coalesce of "Deep Learning" and "Fake" and refers to the use of deep learning to synthesize fake media data [1]. Deepfakes are the results of the manipulation or fabrication of media data through AI techniques that have proven difficult to distinguish between real and fake. The notation of "deepfakes", also known as AI-generated synthetic media, made a splash in 2017 [1] [26], when forged media content with the swapped faces of celebrities was shared on the Reddit platform by an anonymous user with the account "deepfakes" [5].

Because of the hyper-realistic quality made by the forged media generation techniques [3] [27], manual detection of forged media content becomes a challenging task [28] [29]. Therefore, the detection of fake media content has become an important topic for researchers in the field of machine learning. Many media forensics tools have been developed to identify the authentication of media data, including images, video, text, and audio, to identify fabricated or malicious intent [30] [31] [32]. In this study, we focus our attention on the forgery video (deepfake video) detection.

A. Unimodal Deepfake Detection

Most of the previous forgery video detection approaches are unimodal, such as leveraging facial features [18], employing image/frame-based analysis [18] [33] and exploring statistical inconsistencies and visual artifacts for classification [34] [35]. In [36], forgery video detection is performed based on visual artifacts, including illumination reflections, non-identical eye colors, and missing and incomplete details in eye and tooth regions. Moreover, [11] and [18], respectively, proposed high-level and mesoscopic features, [12] adopted a capsule network, [37] designed a XceptionNet, and [38] derived a two-stream convolutional neural networks (CNNs) for forgery detection. The above forgery video detection works mainly focus on the video part, but audio information also plays an important role. To identify speakers in audio, we often use automatic speaker verification (SV) systems to check the speaker's identity. If the SV result is different from what is claimed, the audio is assumed to be forged. However, it has been reported that current SV systems are easy to deceive by manipulating the audio signal [39] [40]. Several previous works have deeply analyzed this problem and proposed potential solutions [14] [41] [42].

B. Multi-modal Deepfake Detection

Learning using multiple modalities, such as audio, video, and text, refers to multimodal learning. It has been demonstrated that by leveraging information from both audio and visual modalities, better performance can be achieved compared to using single-modality information [43] [44]. For forgery detection, [45] proposed to check the consistency of affective features extracted from the audio and visual information. Similarly, [46] proposed a system that considers the dissimilarity between audio and visual modalities for detecting deepfake videos. Meanwhile, [47] proposed a joint audiovisual model by exploiting the intrinsic synchronization between audio and visual modalities to identify whether a given video is real or not.

C. Deepfake Detection Datasets

Numerous datasets have been developed for video forgery detection. Some well-known examples include FFW [48], UADFV [49], DeepfakeTIMIT dataset [50], FaceForensics++ [51], Celeb-DF [52] , Google DFD [51], DeeperForensics [53], DFDC [54], KoDF [55] and FakeAVCeleb [17]. Except for the DFDC and FakeAVCeleb datasets, all other datasets contain only visual manipulations, which makes them unsuitable for audiovisual forgery detection tasks. The DFDC dataset contains extreme environmental settings, such as low or bright light issues, where sometimes the subject's face is not facing the camera. The FakeAVCeleb dataset is a multimodal, gender and geographically balanced dataset with source videos taken from the VoxCeleb2 dataset [56].

III. Proposed Architecture

Fig. 2 illustrates the overview architecture of our system. We tackle the problem of audiovisual forgery detection with

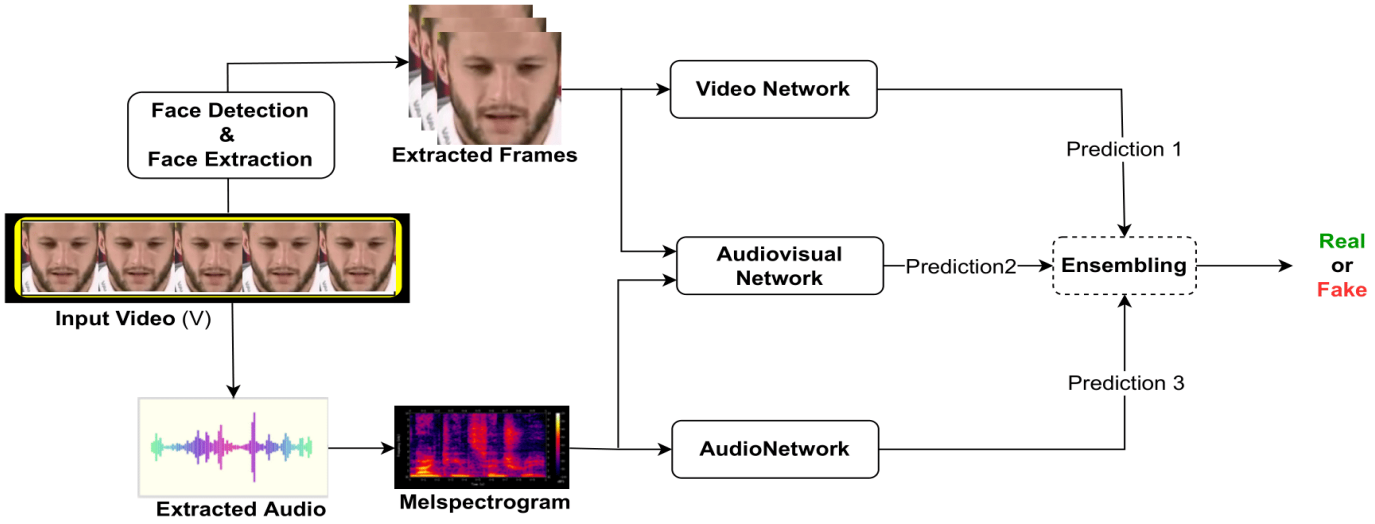

Fig. 2. The proposed multimodal and ensemble learning system for forgery detection.

an ensemble learning technique that consists of a unimodal **Audio Network**, a unimodal **Video Network**, a multimodal **Audiovisual Network**, and a **Voting Module**. Given an input video V, the system aims to predict whether it is real or fake. In this section, we will discuss each part in detail.

A. Audio Network

The **Audio Network** is a unimodal neural network (NN) that takes the Mel-spectral features of the audio extracted from the input video and predicts whether the audio is real or fake. In this study, the **Audio Network** is formed by a simple 2D CNN consisting of four convolutional layers, each followed by a ReLU activation function and batch normalization, as shown in Fig. 3. The final convolutional block is followed by an adaptive pooling layer, which is further followed by a linear classifier that maps 64-dimensional features to 2-dimensional output (real/fake).

B. Video Network

For the **Video Network**, we used the MesoNet [18], a frame-based CNN video classifier, to predict whether a video is real or fake. MesoNet comprises a small number of layers that specifically exploit mesoscopic properties to detect forged video content. It takes as input the frame fragments of a video and determines whether the video is fake or real by hard voting the predictions of all frames.

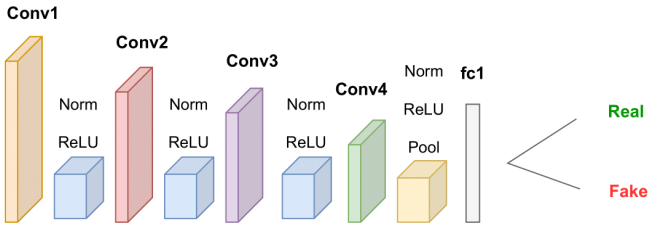

Fig. 3. The **Audio Network**, which is formed by a 2D CNN and designed to predict whether the extracted audio in the video is real or fake.

C. Audiovisual Network

To jointly exploit the audio and visual modalities in a video (V), we employed an audiovisual network in the proposed system. The **Audiovisual Network** is a NN model that combines the audio and video information in a late fusion manner. Fig. 4 shows the architecture of the **Audiovisual Network**. As shown in the figure, the network has two branches, the audio branch and the visual branch. The audio branch is formed by a 2D CNN and takes Mel-spectral features as input, while the visual branch is formed by ResNet3D_18 [57] and takes video frames as input. The outputs of the audio ($\vec{E}_a$) and visual ($\vec{E}_v$) branches are combined via a fusion network to generate the final predictions. The visual branch (ResNet3d_18 [57]) is pretrained on the Kinetics dataset and then fine-tuned on the FakeAVCeleb dataset.

D. Decision-making Module

The ensemble learning technique is a machine learning method that combines predictions from multiple models for

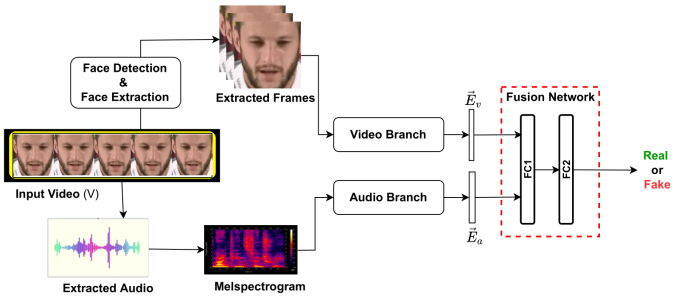

Fig. 4. The **Audiovisual Network**. The audio branch is a 2D CNN, and the visual branch is a ResNet3D_18 network. The outputs of the video branch ($\vec{E}_v$) and audio branch ($\vec{E}_a$) are combined in a late fusion manner.

better results [58]. As shown in Fig. 2, the decision-making module combines the predictions from **Audio Network**, **Video Network**, and **Audiovisual Network** to get the final prediction. There are multiple ways to make the final decision. Among them, hard voting is an intuitive way based on a majority winning solution; on the other hand, soft voting considers the average of multiple predictions. In the proposed system, we simply employ hard voting to generate the final predictions.

IV. Experiments and Results

We present the experimental setup, training data preparation, training hyperparameters, and results in this section.

A. Experimental Setup

1) Dataset: As mentioned earlier, we evaluated the proposed system using the FakeAVCeleb dataset [17]. We selected this dataset for several reasons. First, the FakeAVCeleb dataset contains manipulations in both the audio and video modalities. Second, the dataset is gendered balanced and racially/geographically unbiased. The dataset is divided in four different categories i.e. FakeVideo-FakeAudio (FVFA), RealVideo-FakeAudio (RVFA), FakeVideo-RealAudio (FVRA) and RealVideo-RealAudio (RVRA). Moreover, the latest deepfake generation and synthetic voice generation techniques have been used to generate this dataset, including faceswap, faceswap-wav2lip, fsgan, fsgan-wav2lip, real-time-voice-cloning (RTVC), and wav2lip.

2) Data Analysis and Preprocessing: For better performance, we preprocessed the FakeAVCeleb dataset. For video data, although the videos in the FakeAVCeleb dataset have been face-centered, we used a CNN based face extractor (a.k.a. dlib [59]) to extract the face regions and ignore the rest (e.g. shoulders or background). The extracted video frames are then used as input to the **Video Network** and the video branch of the **Audiovisual Network**. Instead of the entire video sequence, we used a stack of video frames (25 frames) as the video input. On the other hand, we extracted the audio of each video at a sampling rate of 16kHz and stored it in a WAV format. We then extracted the Mel-spectral features as the input of the **Audio Network** and the audio branch of the **Audiovisual Network**. We used the FakeAVCeleb dataset to evaluate several detection systems, including unimodal (audio-alone and video-alone), multimodal (audiovisual) and the proposed ensemble systems. For a more comprehensive evaluation, we constructed eight test sets. Testset-I is the major test set that contains the same number of manipulated videos from each category i.e. RVFA, FVFA, and FVRA in the Fake class. Testset-II contains the same number of video samples from each manipulation technique i.e. faceswap, fsgan, faceswap-wav2lip, fsgan-wav2lip, RTVC, and wav2lip in the Fake class. The remaining test sets are each based on individual manipulation techniques, each containing the same number of video samples from a specific manipulation technique. To avoid bias, each test set contains 70 real videos and 70 fake videos, which are not present in the training set.

TABLE I
Training Sets of Different Classifiers Composed from the FakeAVCeleb Dataset

Classifier	Class	Category	Samples	Catg-Samples	Training-Samples
Audio Network	Fake	RVFA	430	9,841	18,669
		FVFA	9,411		
	Real	FVRA	8,398	8,828	
		RVRA	430		
Video Network	Fake	FVFA	9,411	17,809	35,618
		FVRA	8,398		
	Real	RVFA	430	17,809	
		RVRA	430 + 16,949		
Audiovisual Network	Fake	FVFA	9,411	18,239	36,411
		FVRA	8,398		
		RVFA	430		
	Real	RVRA	430 + 17,742	18,172	

Following the preprocessing steps for all the test sets, we obtained 3,500 video frames; 1,750 of them are real and 1,750 are fake. Meanwhile, we obtained a total of 140 Mel-spectral features, 70 for each class (fake and real).

3) Evaluation Metrics: For performance evaluation, we use precision, recall, f1-score, and accuracy as evaluation metrics,

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

where TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively.

4) Training Hyperparameters: We trained all three classifiers using the Adam optimizer with a learning rate of 0.001 and a cross-entropy loss. The batch sizes for training the audio network, video network, and audiovisual network are 512, 64, and 6, respectively.

B. Training Sets for Different Models

We used the FakeAVCeleb dataset for training; however, the training samples for three networks, namely the unimodal Audio classifier, the unimodal Video classifier, and the multimodal Audiovisual classifier, are different. The FakeAVCeleb dataset has four main categories of videos, namely FVFA (FakeVideo-FakeAudio), RVFA (RealVideo-FakeAudio), FVRA (FakeVideo-RealAudio) and RVRA (RealVideo-RealAudio). For binary classification, we need to compose training sets for unimodal and multimodal classifiers according to their respective requirements. The training sets for these classifiers are shown in Table I.

1) Audio Network: As shown in Table I, we trained the unimodal Audio classifier on a total of 18,669 audio samples, 9,841 from the fake class and 8,828 from the real class. The real class includes videos of the FVRA and RVRA categories because the audio modality is not manipulated in these two categories, while the fake class includes videos of the RVFA and FVFA categories because the audio modality is manipulated in these two categories.

2) Video Network: For the Video classifier, we considered the videos of the FVFA and FVRA categories as the fake class because the video modality in both categories is fake, and the videos of the RVRA and RVFA categories as the real class due to their real video modality in both categories. As shown in Table I, there are only 860 training samples for the real class (430 video samples in RVRA and RVFA, respectively), but a total of 17,809 training samples for the fake class (9,411 video samples in FVFA and 8,398 video samples in FVRA) in the FakeAVCeleb dataset. The number of training samples for the two classes is extremely unbalanced. To overcome this issue, we augmented the training samples of the real class with 16,949 video samples (shown as red in I) from the VoxCeleb1 dataset [60]. They are all real videos belonging to the RVRA category. Finally, the number of training samples for each class is the same. The unimodal Video classifier has trained on a total of 35,618 video samples, with 17,809 training video samples for each class (real and fake).

3) Audiovisual Network: In multimodal fake video detection, a video is considered fake if any modality of the video is fake. Therefore, a video is considered real only if its audio and video modalities are both real. The training samples for the fake class include the video clips in FVFA, FVRA, and RVFA, while the training samples for the real class only include the video clips in RVRA. Again, the number of training samples for the two classes is extremely unbalanced. As shown in Table I (in red), we augmented the training samples for the real class with 17,742 videos from the VoxCeleb1 dataset [60]. We trained two multimodal networks, namely the Audiovisual network and the Freeze Audiovisual network, on a total of 36,411 video samples, with 18,239 video samples for the fake class and 18,172 video samples for the real Class. The Freeze Audiovisual network is a variant of the Audiovisual network with some parameters fixed during training.

C. Results

1) Audio-only Detection Results: The audio-only detection results evaluated on eight different test sets consisting of various manipulation techniques are reported in Table II. From the table, we note that the audio-only model (**Audio Network**) performs well on the fsgan-wav2lip, RTVC and faceswap-wav2lip test sets with accuracies of 0.98, 0.97 and 0.96, respectively. The high accuracy on these test sets is because their Fake class consists of video clips with manipulated audio. For the Testset-I, wav2lip and Testset-II test sets, the Fake class consists of video clips with or without audio manipulation; thus **Audio Network** yields moderate performance with accuracies of 0.81, 0.79 and 0.78. **Audio Network** performs poorly on the faceswap and fsgan test sets, because the Fake class consists of video clips with only video manipulation. Therefore, all test video clips should be judged as Real by Audio Network, and the model did get about 50% accuracy. The results of this experiment show that the audio-only network can effectively identify whether audio is manipulated, but can not detect visual manipulation.

TABLE II
AUDIO-ONLY DETECTION RESULTS

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset- I	Real	0.73	0.97	0.83	0.81
	Fake	0.96	0.64	0.77	
Testset- II	Real	0.71	0.94	0.81	0.78
	Fake	0.91	0.61	0.74	
faceswap	Real	0.50	0.96	0.65	0.49
	Fake	0.40	0.03	0.05	
faceswap-wav2lip	Real	0.97	0.96	0.96	0.96
	Fake	0.96	0.97	0.96	
fsgan	Real	0.50	0.96	0.65	0.49
	Fake	0.40	0.03	0.05	
fsgan-wav2lip	Real	0.99	0.97	0.98	0.98
	Fake	0.97	0.99	0.98	
RTVC	Real	0.99	0.96	0.97	0.97
	Fake	0.96	0.99	0.97	
wav2lip	Real	0.71	0.97	0.82	0.79
	Fake	0.95	0.60	0.74	

TABLE III
VIDEO-ONLY DETECTION RESULTS

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset- I	Real	0.76	0.98	0.86	0.83
	Fake	0.97	0.66	0.79	
Testset- II	Real	0.82	0.98	0.90	0.88
	Fake	0.98	0.78	0.87	
faceswap	Real	0.84	0.99	0.91	0.90
	Fake	0.99	0.81	0.89	
faceswap-wav2lip	Real	0.98	0.99	0.99	0.98
	Fake	0.99	0.98	0.98	
fsgan	Real	0.95	0.99	0.97	0.97
	Fake	0.99	0.95	0.97	
fsgan-wav2lip	Real	0.99	0.99	0.99	0.99
	Fake	0.99	0.99	0.99	
RTVC	Real	0.50	0.99	0.66	0.50
	Fake	0.43	0.01	0.01	
wav2lip	Real	0.95	0.99	0.97	0.96
	Fake	0.99	0.93	0.95	

2) Video-only Detection Results: Table III shows the results of the video-only network (Meso4) evaluated on the eight test sets. We note a similar trend to the results of the audio-only network: **Video Network** performs well on the fsgan-wav2lip, faceswap-wav2lip, fsgan and wav2lip test sets, since their Fake class consists of video clips with visual manipulation. However, relatively poor performance is obtained on the faceswap, Testset-II and Test-I test sets since the Fake class includes video clips with or without visual manipulation. The model has the worst performance on RTVC with 50% accuracy, because the Fake class only contains video clips with audio manipulation, while the visual parts are all real. The results of this experiment show that the video-only network can effectively identify whether the video is manipulated, but can not detect audio manipulation.

3) Audiovisual Detection Results: Table IV shows the results of **Audiovisual Network** evaluated on the eight test sets. For the faceswap-wav2lip, fsgan-wav2lip and RTVC test set,

TABLE IV
AUDIOVISUAL DETECTION RESULTS

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset- I	Real	0.83	0.96	0.89	0.88
	Fake	0.95	0.80	0.87	
Testset- II	Real	0.71	0.96	0.82	0.79
	Fake	0.93	0.61	0.74	
faceswap	Real	0.49	0.96	0.65	0.49
	Fake	0.25	0.01	0.03	
faceswap-wav2lip	Real	1.00	0.96	0.98	0.98
	Fake	0.96	1.00	0.98	
fsgan	Real	0.50	0.96	0.65	0.49
	Fake	0.40	0.03	0.05	
fsgan-wav2lip	Real	1.00	0.96	0.98	0.98
	Fake	0.96	1.00	0.98	
RTVC	Real	0.99	0.96	0.97	0.97
	Fake	0.96	0.99	0.97	
wav2lip	Real	0.82	0.96	0.88	0.87
	Fake	0.95	0.79	0.86	

TABLE V
RESULTS OF THE FREEZE AUDIOVISUAL MODEL

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset- I	Real	0.81	0.84	0.83	0.82
	Fake	0.84	0.80	0.82	
Testset- II	Real	0.75	0.84	0.79	0.78
	Fake	0.82	0.71	0.76	
faceswap	Real	0.54	0.84	0.66	0.57
	Fake	0.65	0.29	0.40	
faceswap-wav2lip	Real	0.66	0.84	0.74	0.90
	Fake	0.78	0.56	0.65	
fsgan	Real	0.66	0.84	0.74	0.70
	Fake	0.78	0.56	0.65	
fsgan-wav2lip	Real	0.98	0.84	0.91	0.91
	Fake	0.86	0.99	0.92	
RTVC	Real	0.81	0.84	0.83	0.82
	Fake	0.84	0.80	0.82	
wav2lip	Real	0.83	0.84	0.84	0.84
	Fake	0.84	0.83	0.83	

the Fake class contains video clips with both audio and visual manipulations. Therefore, Audiovisual Network performs best on these three test sets. Similarly, the performance on Testset-I, wav2lip and Testset-II is reasonable, with accuracies of 0.88, 0.87 and 0.79, respectively. However, the Audiovisual classifier performs the worst on faceswap and fsgan with an accuracy of 0.49. Since the Fake class of both test sets contains video clips with only video manipulation, we believe that the poor performance is due to the dominance of the audio modality in the final prediction.

Additionally, we trained another Audiovisual network. We froze the feature extractors of the audio and visual networks and fine-tuned only the classifier part. Table V shows the results of the Freeze Audiovisual classifier evaluated on the eight test sets. Comparing Tables IV and V, we can see that the accuracy on faceswap and fsgan increased from 0.49 to 0.57 and 0.70, respectively, but the accuracy on the other test sets dropped quite a bit. Since the Freeze Audiovisual classifier has a more balanced performance than the Audiovisual classifier, we believe that the Freeze Audiovisual classifier should be used for ensemble learning.

4) Ensemble Learning Detection Results: The results of the Ensemble classifier evaluated on the eight test sets are shown in Table VI. Comparing Table VI with Tables II- V, it is clear that the Ensemble classifier generally performs better than the Audio-only network, Video-only network, Audiovisual network and Freeze Audiovisual network. Among all the classifiers, the Ensemble classifier performs best on 4 of the 8 test sets, including Testset-I, Testset-II, faceswap-wav2lip,

TABLE VI
RESULTS OF THE ENSEMBLE AUDIOVISUAL MODEL

Test-Set type	Class	Precision	Recall	F1-Score	Accuracy
Testset- I	Real	0.83	0.99	0.90	0.89
	Fake	0.98	0.80	0.88	
Testset- II	Real	0.82	0.99	0.90	0.89
	Fake	0.98	0.79	0.87	
faceswap	Real	0.64	0.99	0.78	0.72
	Fake	0.97	0.46	0.62	
faceswap-wav2lip	Real	0.99	0.99	0.99	0.99
	Fake	0.99	0.99	0.99	
fsgan	Real	0.81	0.99	0.89	0.88
	Fake	0.98	0.77	0.86	
fsgan-wav2lip	Real	1.00	0.99	0.99	0.99
	Fake	0.99	1.00	0.99	
RTVC	Real	0.82	0.99	0.90	0.89
	Fake	0.98	0.79	0.87	
wav2lip	Real	0.88	0.99	0.93	0.93
	Fake	0.98	0.87	0.92	

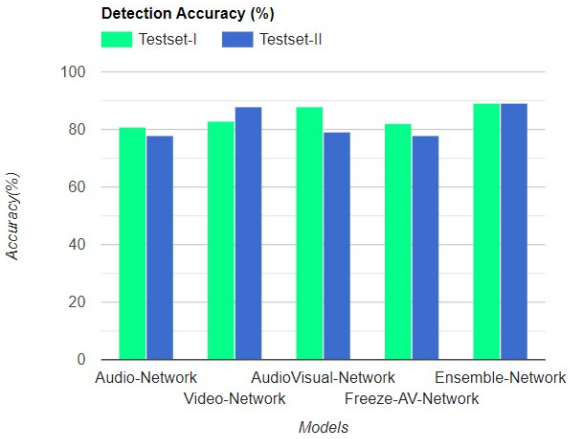

Fig. 5. Comparison of the results of our models.

and fsgan-wav2lip. Obtaining the best performance on Testset-I and Testset-II is an important indicator for selecting the best method, since these two test sets cover fake videos of all categories (RVFA, FVRA, and FVFA) through various manipulation techniques. In addition, the Ensemble classifier provides more balanced results than the Freeze Audiovisual network, improving the accuracy to 0.72 and 0.88 on faceswap and fsgan, respectively.

5) Comparison of our Models: Among the 8 test sets, Testset-I is a main test set consisting of the same number of video samples of all four video categories in the FakeAVCeleb dataset, namely RVRA, RVFA, FVRA, and FVFA, while Testset-II is another main test set consisting of the same number of video samples from each manipulation technique, including faceswap, faceswap-wav2lip, fsgan, fsgan-wav2lip, RTVC and wav2lip. Fig. 5 shows the accuracy of five different deepfake detection models evaluated on Testset-I and Testset-II. As shown in the experiments above, **Audio-Network** can accurately detect fake audio, but it is only effective for detecting audio manipulation in video. Similarly, **Video-Network** can accurately detect fake video, but it is only effective for detecting visual manipulation in video. Since the audio or visual parts in an input video may be altered, only **AudioVisual-Network**, **Freeze-AV-Network** and **Ensemble-Network** that can detect audio and visual manipulations in the video can meet the needs of practical applications. While **AudioVisual-Network** has higher accuracy than **Freeze-AV-Network**, **Freeze-AV-Network** is more stable in detecting various types of manipulations in video. Among all the five models, **Ensemble-Network** performs the best on Testset-I and Testset-II with an accuracy of 0.89.

6) Comparison of our Ensemble Model and other Models: Finally, we compare our ensemble model with a variety of existing unimodal, multimodal, and ensemble methods, as shown in Fig. 6. All models were evaluated on Testset-I. The detailed results are shown in Table VII. The Unimodal VGG16 model trained on the visual modality outperformed all other video unimodal methods in [16], but only achieved 0.81 accuracy. Xception achieved an accuracy of 0.76 by leveraging

TABLE VII
RESULTS OF OUR ENSEMBLE MODEL AND SEVERAL EXISTING UNIMODAL, ENSEMBLE AND MULTIMODAL METHODS. THE DFD IN THE FIRST COLUMN REFERS TO THE DEEPFAKE DETECTION METHOD. "V", "A" AND "AV" STAND FOR VISUAL, AUDIO AND AUDIOVISUAL MODALITIES, RESPECTIVELY.

DFD Method	Model	Modality	Class	Precision	Recall	F1-score	Accuracy
Unimodal [16]	VGG16	V	Real	0.6935	0.8966	0.7821	0.8103
			Fake	0.8724	0.7750	0.8208	
Unimodal [16]	Xception	A	Real	0.8750	0.6087	0.7179	0.7626
			Fake	0.7033	0.9143	0.7950	
Unimodal [23]	LipForensics	V	Real	0.70	0.91	0.80	0.76
			Fake	0.88	0.61	0.72	
Ensemble (Soft-Voting) [16]	VGG16	AV	Real	0.6935	0.8966	0.7821	0.7804
			Fake	0.8948	0.6894	0.7788	
Ensemble (Hard-Voting) [16]	VGG16	AV	Real	0.6935	0.8966	0.7821	0.7804
			Fake	0.8948	0.6894	0.7788	
Multimodal-1 [16]	Multimodal-1	AV	Real	0.000	0.000	0.000	0.5000
			Fake	0.496	1.000	0.663	
Multimodal-2 [16]	Multimodal-2	AV	Real	0.710	0.587	0.643	0.674
			Fake	0.648	0.760	0.700	
Multimodal-3 [16]	CDCN	AV	Real	0.500	0.068	0.120	0.515
			Fake	0.500	0.940	0.651	
Multimodal-4 [45]	Not-made-for-each-other	AV	Real	0.62	0.99	0.76	0.69
			Fake	0.94	0.40	0.57	
Multimodal (ours)	**Ensemble**	**AV**	**Real**	**0.83**	**0.99**	**0.90**	**0.89**
			Fake	**0.98**	**0.80**	**0.88**	

only the audio modality. The LipForensics model used high-level features in the form of lip movements and achieved an accuracy of 0.76 due to unimodality (i.e., visual lip movements). In [16], the authors reported results from various multimodal and ensemble models, including Ensemble (soft- and hard-voting), Multimodal-1, Multimodal-2 and Multimodal-3. Despite exploiting both modalities, their multimodal and ensemble models yield poor results. We trained the audiovisual dissonance-based model in [45] on the FakeAVCeleb dataset. The model (Multimodal-4) achieves 0.69 accuracy. It is clear from Table VII that our ensemble model outperforms all existing unimodal, multimodal and ensemble models with an accuracy of 0.89. This is a new state-of-the-art performance on the FakeAVCeleb dataset.

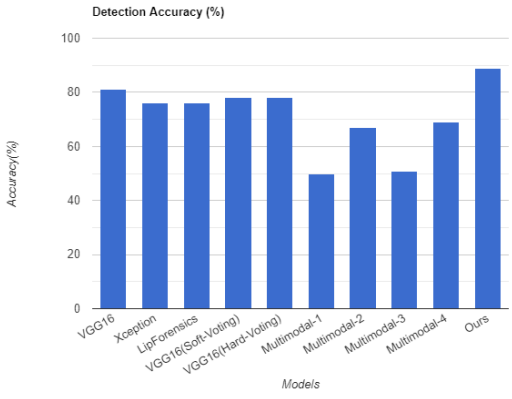

Fig. 6. Comparison of our ensemble model and several existing unimodal, ensemble and multimodal models.

V. CONCLUSION

In this paper, we propose a novel ensemble learning technique for audiovisual deepfake detection. It aims to leverage audio and visual manipulations in video clips for better forgery detection. Furthermore, the proposed method uses decision fusion for higher prediction performance. Our experiments show that the proposed method achieve state-of-the-art results on the FakeAVCeleb dataset, which is a recently released multi-modal manipulation dataset. We believe that our work is a useful step towards effective audiovisual deepfake detection. In the future, we intend to use Transformer with other ensemble models for effective forgery detection.

ACKNOWLEDGMENT

This research work was supported by the NSTC-Taiwan Grant 111-2221-E-001-002.

REFERENCES

[1] M. Westerlund, The emergence of deepfake technology: A review, Technology Innovation Management Review 9 (11) (2019). 1, 2
[2] C. Vaccari, A. Chadwick, Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news, Social Media+ Society 6 (1) (2020) 2056305120903408. 1
[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2387–2395. 1, 2
[4] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: Image synthesis using neural textures, ACM Transactions on Graphics (TOG) 38 (4) (2019) 1–12. 1
[5] J. Kietzmann, L. W. Lee, I. P. McCarthy, T. C. Kietzmann, Deepfakes: Trick or treat?, Business Horizons 63 (2) (2020) 135–146. 1, 2
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014). 1
[7] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013). 1
[8] H. Farid, Image forgery detection, IEEE Signal processing magazine 26 (2) (2009) 16–25. 1
[9] C.-M. Pun, X.-C. Yuan, X.-L. Bi, Image forgery detection using adaptive oversegmentation and feature point matching, ieee transactions on information forensics and security 10 (8) (2015) 1705–1716. 1
[10] J. A. Redi, W. Taktak, J.-L. Dugelay, Digital image forensics: a booklet for beginners, Multimedia Tools and Applications 51 (1) (2011) 133–162. 1

[11] B. Bayar, M. C. Stamm, A deep learning approach to universal image manipulation detection using a new convolutional layer, in: Proceedings of the 4th ACM workshop on information hiding and multimedia security, 2016, pp. 5–10. 1, 2

[12] H. H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2307–2311. 1, 2

[13] C.-C. Hsu, Y.-X. Zhuang, C.-Y. Lee, Deep fake image detection based on pairwise learning, Applied Sciences 10 (1) (2020) 370. 1

[14] A. Tomilov, A. Svishchev, M. Volkova, A. Chirkovskiy, A. Kondratev, G. Lavrentyeva, Stc antispoofing systems for the asvspoof2021 challenge, in: Proc. ASVspoof 2021 Workshop, 2021, pp. 61–67. 1, 2

[15] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, E. Khoury, Generalization of audio deepfake detection., in: Odyssey, 2020, pp. 132–137. 1

[16] H. Khalid, M. Kim, S. Tariq, S. 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, 2021, pp. 7–15. 1, 2, 6, 7

[17] H. Khalid, S. Tariq, M. Kim, S. S. Woo, Fakeavceleb: a novel audio-video multimodal deepfake dataset, arXiv preprint arXiv:2108.05080 (2021). 1, 2, 4

[18] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE international workshop on information forensics and security (WIFS), IEEE, 2018, pp. 1–7. 1, 2, 3

[19] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, arXiv preprint arXiv:1811.00656 (2018). 1

[20] T. Jung, S. Kim, K. Kim, Deepvision: Deepfakes detection using human eye blinking pattern, IEEE Access 8 (2020) 83144–83154. 1

[21] G. Mazaheri, A. K. Roy-Chowdhury, Detection and localization of facial expression manipulations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1035–1045. 1

[22] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8261–8265. 1

[23] A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, Lips don't lie: A generalisable and robust approach to face forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5039–5049. 1, 7

[24] R. Yasrab, W. Jiang, A. Riaz, Fighting deepfakes using body language analysis, Forecasting 3 (2) (2021) 303–321. 1

[25] T. Jung, S. Kim, K. Kim, Deepvision: Deepfakes detection using human eye blinking pattern, IEEE Access 8 (2020) 83144–83154. 1

[26] S. Lyu, Deepfake detection: Current challenges and next steps, in: 2020 IEEE international conference on multimedia & expo workshops (ICMEW), IEEE, 2020, pp. 1–6. 2

[27] I. Korshunova, W. Shi, J. Dambre, L. Theis, Fast face-swap using convolutional neural networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3677–3685. 2

[28] B. Chesney, D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, Calif. L. Rev. 107 (2019) 1753. 2

[29] N. C. Köbis, B. Doležalová, I. Soraperra, Fooled twice: People cannot detect deepfakes but think they can, Iscience 24 (11) (2021) 103364. 2

[30] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, Protecting world leaders against deep fakes., in: CVPR workshops, Vol. 1, 2019. 2

[31] K. Bhagtani, A. K. S. Yadav, E. R. Bartusiak, Z. Xiang, R. Shao, S. Baireddy, E. J. Delp, An overview of recent work in media forensics: Methods and threats, arXiv preprint arXiv:2204.12067 (2022). 2

[32] L. Verdoliva, Media forensics and deepfakes: an overview, IEEE Journal of Selected Topics in Signal Processing 14 (5) (2020) 910–932. 2

[33] A. Ajoy, C. U. Mahindrakar, D. Gowrish, A. Vinay, Deepfake detection using a frame based approach involving cnn, in: 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, 2021, pp. 1329–1333. 2

[34] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE, 2019, pp. 83–92. 2

[35] S. Agarwal, H. Farid, O. Fried, M. Agrawala, Detecting deep-fake videos from phoneme-viseme mismatches, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 660–661. 2

[36] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), IEEE, 2019, pp. 83–92. 2

[37] 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. 2

[38] P. Zhou, X. Han, V. I. Morariu, L. S. Davis, Two-stream neural networks for tampered face detection, in: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), IEEE, 2017, pp. 1831–1839. 2

[39] R. K. Das, T. Kinnunen, W.-C. Huang, Z. Ling, J. Yamagishi, Y. Zhao, X. Tian, T. Toda, Predictions of subjective ratings and spoofing assessments of voice conversion challenge 2020 submissions, arXiv preprint arXiv:2009.03554 (2020). 2

[40] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., Tacotron: Towards end-to-end speech synthesis, arXiv preprint arXiv:1703.10135 (2017). 2

[41] X. Wang, J. Yamagishi, A comparative study on recent neural spoofing countermeasures for synthetic speech detection, arXiv preprint arXiv:2103.11326 (2021). 2

[42] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, et al., Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection, arXiv preprint arXiv:2109.00537 (2021). 2

[43] S. Agarwal, H. Farid, O. Fried, M. Agrawala, Detecting deep-fake videos from phoneme-viseme mismatches, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 660–661. 2

[44] P. Korshunov, S. Marcel, Speaker inconsistency detection in tampered video, in: 2018 26th European signal processing conference (EUSIPCO), IEEE, 2018, pp. 2375–2379. 2

[45] K. Chugh, P. Gupta, A. Dhall, 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, 2020, pp. 439–447. 2, 7

[46] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, D. Manocha, Emotions don't lie: An audio-visual deepfake detection method using affective cues, in: Proceedings of the 28th ACM international conference on multimedia, 2020, pp. 2823–2832. 2

[47] Y. Zhou, S.-N. Lim, Joint audio-visual deepfake detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 14800–14809. 2

[48] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, C. Busch, Fake face detection methods: Can they be generalized?, in: 2018 international conference of the biometrics special interest group (BIOSIG), IEEE, 2018, pp. 1–6. 2

[49] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8261–8265. 2

[50] P. Korshunov, S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection, arXiv preprint arXiv:1812.08685 (2018). 2

[51] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11. 2

[52] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3207–3216. 2

[53] L. Jiang, R. Li, W. Wu, C. Qian, C. C. Loy, Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2889–2898. 2

[54] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C. C. Ferrer, The deepfake detection challenge (dfdc) dataset, arXiv preprint arXiv:2006.07397 (2020). 2

[55] P. Kwon, J. You, G. Nam, S. Park, G. Chae, Kodf: A large-scale korean deepfake detection dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10744–10753. 2

[56] J. S. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep Speaker Recognition, in: Proc. Interspeech 2018, 2018, pp. 1086–1090. doi:10.21437/Interspeech.2018-1929. 2
[57] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6450–6459. 3
[58] D. Opitz, R. Maclin, Popular ensemble methods: An empirical study, Journal of artificial intelligence research 11 (1999) 169–198. 4
[59] D. E. King, Dlib-ml: A machine learning toolkit, The Journal of Machine Learning Research 10 (2009) 1755–1758. 4
[60] A. Nagrani, J. S. Chung, A. Zisserman, Voxceleb: a large-scale speaker identification dataset, in: INTERSPEECH, 2017. 5