A Novel Smart Deepfake Video Detection System

Marwa Elpeltagy[1], Aya Ismail[2]*, Mervat S. Zaki[3], Kamal Eldahshan[4]
Systems and Computers Department, Al-Azhar University, Egypt[1]
Mathematics Department, Tanta University, Egypt[2]
Mathematics Department, Al-Azhar University (Girls Branch), Egypt[3]
Mathematics Department, Al-Azhar University, Egypt[4]

***Abstract*—Rapid advancements in deep learning-based technologies have developed several synthetic video and audio generation methods producing incredibly hyper-realistic deepfakes. These deepfakes can be employed to impersonate the identity of a source person in videos by swapping the source's face with the target one. Deepfakes can also be used to clone the voice of a target person utilizing audio samples. Such deepfakes may pose a threat to societies if they are utilized maliciously. Consequently, distinguishing either one or both deepfake visual video frames and cloned voices from genuine ones has become an urgent issue. This work presents a novel smart deepfake video detection system. The video frames and audio are extracted from given videos. Two feature extraction methods are proposed, one for each modality; visual video frames, and audio. The first method is an upgraded XceptionNet model, which is utilized for extracting spatial features from video frames. It produces feature representation for visual video frames. The second one is a modified InceptionResNetV2 model based on the Constant-Q Transform (CQT) method. It is employed to extract deep time-frequency features from the audio modality. It produces feature representation for the audio. The corresponding extracted features of both modalities are fused at a mid-layer to produce a bimodal information-based feature representation for the whole video. These three representation levels are independently fed into the Gated Recurrent Unit (GRU) based attention mechanism helping to learn and extract deep and important temporal information per level. Then, the system checks whether the forgery is only applied to video frames, audio, or both, and produces the final decision about video authenticity. The newly suggested method has been evaluated on the FakeAVCeleb multimodal videos dataset. The experimental results analysis assures the superiority of the new method over the current–state-of-the-art methods.**

Keywords*—Deepfake; deepfake detection; bimodal; XceptionNet; InceptionResNetV2; constant-Q transform; CQT; Gated Recurrent Unit; GRU; video authenticity; deep learning; multimodal

I. Introduction

The rapid development of artificial intelligence techniques: autoencoders, Generative Adversarial Networks (GANs), and variational autoencoders facilitated the generation of hyper-realistic fake videos, images, and audio. A deepfake indicates a synthetic image or video AI-generated by swapping an individual's face with another. Applications such as ZAO [1], and DeepFaceLab [2] enable individuals to rapidly generate forged images and videos easily. Recently, a human's voice can be cloned using advanced AI techniques. AI-based audio manipulation is a category of deepfake that clones a human's voice and shows that human saying things that he never said. Overdub, iSpeech, and VoiceApp are instances of voice cloning open-access platforms that can generate synthesized deepfake sounds that nearly resemble the target human's speech [3]. The work of [4] is an example of these manipulation methods, which involves the creation of highly realistic deepfake videos with a precise lip-sync using a group of AI technologies; FaceSwap, FaceSwap GAN, DeepFaceLab, SV2TTS [5], and Wav2Lip [6].

The majority of deepfake videos are created by cloning sounds, synchronizing lips, and frame-by-frame synthesizing faces. Nevertheless, they lack natural emotions, pauses, and breathing behaviour. Additionally, they suffer from discontinuity and faces' flickering among frames. Deepfake can be misused to impersonate individuals, configure an opinion towards a public figure, and spread falsified news. Therefore, a deepfake detection method is needed to cope with the progress in the deepfake generation process and to distinguish the fakes in video frames, audio, and the whole video including video frames and audio.

This paper introduces a smart deepfake detection method that captures the manipulation in a video (multimodal by nature) on three levels; video frames, audio, and the whole video. It distinguishes whether a given video is a deepfake or not. Two proposed feature extraction methods are employed to extract features from video frames and audio modalities. The first method applied to the visual video frames modality is the XceptionNet with some newly introduced modifications. The Xception network achieved effective results in distinguishing the manipulated videos [7, 8]. The suggested modifications to the Xception network produce useful spatial information of the video frames and improve the deepfake detection method performance. The second method applied to the audio modality is a modified InceptionResNetV2 model based on the CQT method to produce deep time-frequency information of the audio segments and improve the detection method performance. The CQT is a time-frequency analysis method that produces higher time resolution at high-frequency areas and higher frequency resolution at low-frequency areas [9]. Its efficiency has been proven in music signal processing tasks [10], speaker verification systems [11], acoustic scenes and events detection and classification [12], anti-spoofing [13], synthetic speech detection [14, 15], and speech emotion recognition [9]. The corresponding features extracted from the two modalities are fused at a mid-layer to create a bimodal information-based feature representation for the whole video. Finally, the GRU-based attention mechanism is applied to these three levels of representation independently. This assists to learn instructive temporal information for each level and

*Corresponding Author.

detect deepfake videos. The GRU performs well in tasks of sequence learning and overcomes the gradient vanishing and explosion problems of the standard recurrent neural network [16]. The proficiency of the attention mechanism has been proven in several areas including machine translations, image captioning, question answering, speech recognition [17], and event detection [18]. A comparative study with recent state-of-the-art deepfake detection methods is conducted in terms of accuracy, Area Under Receiver Operating Characteristic (AUROC) curve metric, precision, recall, F1-score, sensitivity, and specificity.

The rest of this work is organized as follows: Section II presents the literature review for deepfake video detection methods. Section III presents the newly proposed method for deepfake video detection. Section IV is dedicated to the experimental results and analysis. The conclusion and future work are presented in Section V.

II. Literature Review

The progress of AI-based video and voice generation methods raised the ease of creating natural and highly realistic deepfakes that can never be distinguished. Since deepfakes violate security and pose a real threat to society, several researchers have directed their interest to create methods for detecting deepfakes. However, they concentrate on detecting the deepfakes either in video frames or audio modality.

Some of the existing deepfake visual video detection methods spot the manipulation by targeting specific spatial and temporal artifacts that are generated during the fake creation process. Some other detection methods are data-driven that do not target any specific artifacts and distinguish the manipulation by classification [3]. The deepfake visual video detection methods can be categorized into Convolution Neural Network (CNN)-based methods [19, 20, 21, 22], methods that are based on CNN with a temporal network [23, 24, 25, 26, 27], handcrafted feature-based methods [28], and handcrafted feature-based methods with deep networks [29, 30]. This is illustrated in Fig. 1.

The work of [19] detected the deepfakes by exploiting artifacts left by the generation methods when warping the target image to be consistent with the source video. It used four pre-trained CNN models for detecting fake contents; ResNet101, VGG16, ResNet50, and ResNet152. Since deepfake videos suffer from inconsistency among the inter-frames, Hu et al. [20] introduced two branches that are based on CNNs to capture those local and global inconsistencies and then detect deepfakes. Rana and Sung [21] proposed a deep ensemble learning method for detecting deepfake videos. Their method depended on combining several deep base-learners and then training a CNN on these learners to build an ameliorated classifier. In [22], a fine-tuned InceptionResNetV2 model followed by the XGBoost model was employed to capture discrepancies in the spatial domain of fake videos and then individuate deepfakes. The FakeApp creates forged videos that had intra-frame and temporal inconsistencies between frames. Such inconsistencies were detected using InceptionV3 CNN and long short-term memory (LSTM) models [23]. As AI-generated fake videos lack normal eye blinking, Li et al. [24] introduced the VGG16-LSTM to capture the temporal regularities in the eye blinking process and then distinguish the deepfakes. Most deepfake videos are created frame-by-frame where each forged face is created independently. This causes incoherence in the temporal domain of the face region; discontinuity and flickering. As a result, Zheng et al. [25] introduced a fully temporal convolution network that aimed to learn the temporal discrepancies while removing spatial ones. Then, a temporal transformer encoder followed by a multi-layer perceptron was employed to learn the long-range inconsistencies along the time dimension, and then distinguish the deepfakes. In [26], a 2D CNN-based Spatio-temporal learning model was introduced to learn and capture spatial and temporal inconsistencies of forged videos. This temporal inconsistency was captured from both vertical and horizontal directions over adjacent frames and helped in detecting the fakes. The work of [27] introduced a fine-tuned EfficientNet-b5 model followed by the bidirectional LSTM model and densely connected layer. It aimed to discover the Spatio-temporal inconsistencies in deepfake videos and then distinguish the authenticity of videos. Deepfakes were created by joining the generated face into the source image. This produced errors in facial landmark locations that were detected by estimating the 3D head poses for real and deepfake videos. Then, the estimated difference of head poses was fed into the Support Vector Machine (SVM) for deepfake detection [28]. Khalil et al. [29] proposed a model that employed the local binary patterns descriptor to analyze the texture of real and fake videos. Additionally, a CNN-based enhanced high-resolution network was used to automatically capture informative multi-resolution representations of these videos. Then, the output of both was fed into the capsule network to individuate deepfakes. Ismail et al. [30] introduced a hybrid method in which two feature extraction methods were employed to learn and extract enrich spatial features from the detected face frames of video. These methods were a CNN that was based on the Histogram of Oriented Gradient (HOG) method and the improved XceptionNet. Their outputs were merged to be fed into GRUs sequence to extract the spatiotemporal features and detect the fake videos.

The deepfake audio detection methods can be categorized into handcrafted feature-based methods [31, 32], methods that are based on low-level features with CNN [14, 33, 34, 35], methods that rely on using low-level features with CNN and temporal network [37, 38], and end-to-end deep networks-based methods [39]. This is presented in Fig. 2.

The work of [31] extracted several low-level-features; Constant-Q Cepstral Coefficients (CQCC), Cepstrum, Mel-Frequency Cepstrum Coefficients (MFCC), inverted MFCC, Linear Predictive Cepstral Coefficients (LPCC), and LPCC-residual features. These features were utilized along with the Gaussian Mixture Model (GMM) to detect the forged audio.

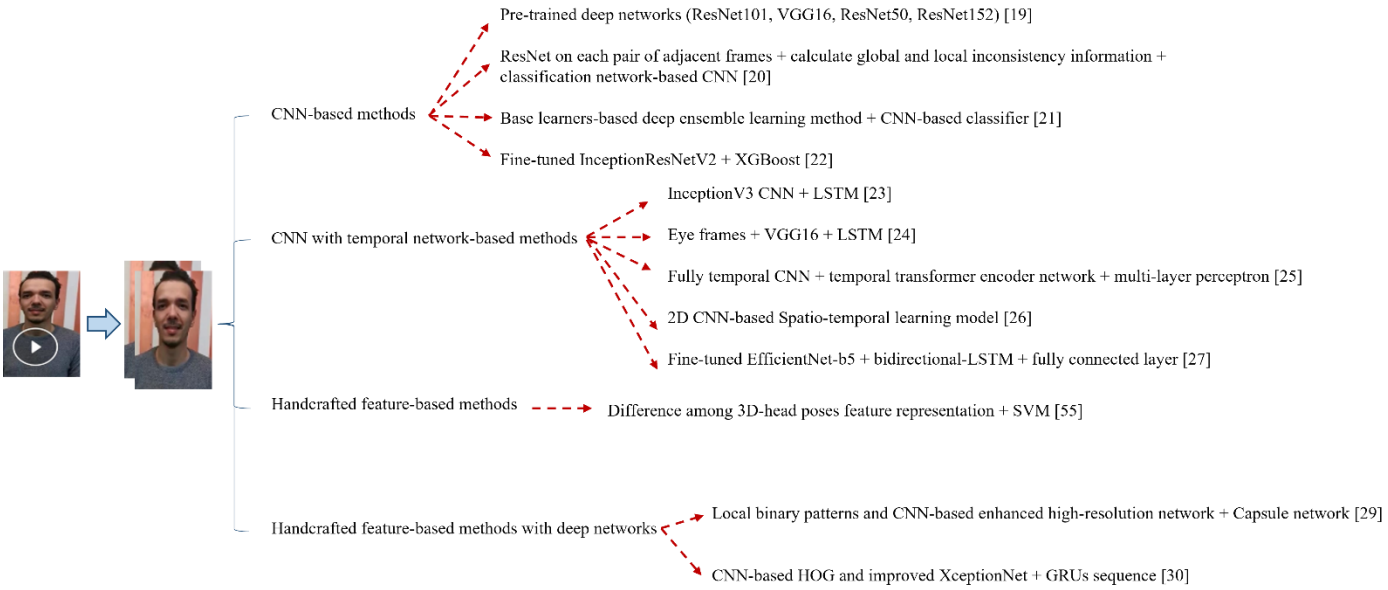


Fig. 1. The deepfake visual video detection methods categorization.

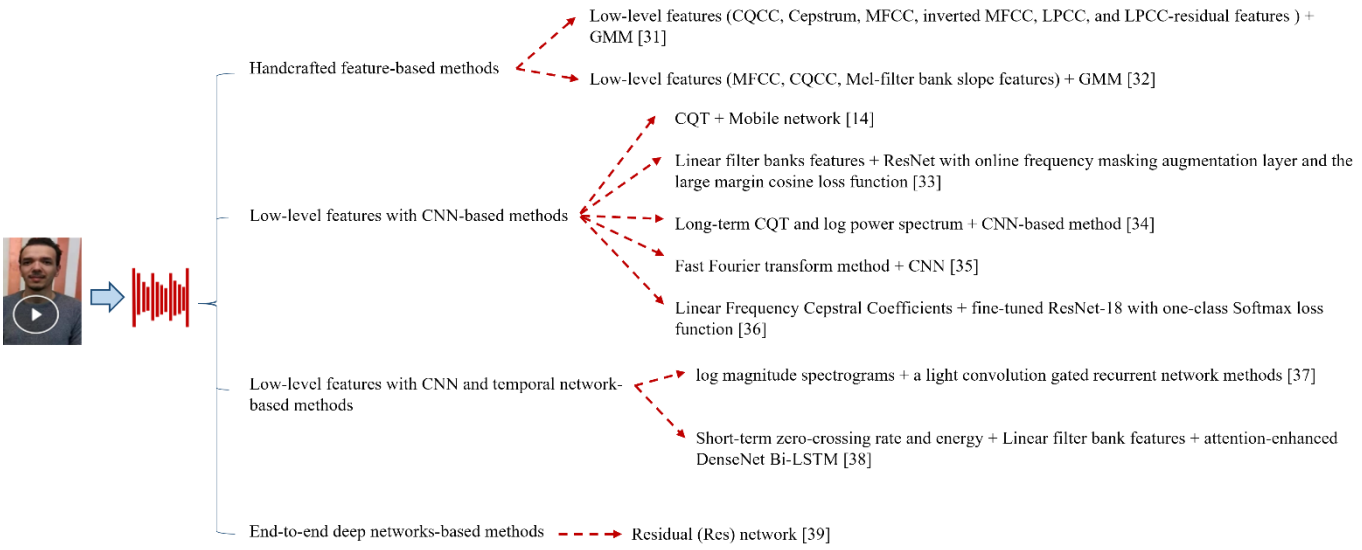


Fig. 2. The deepfake audio detection methods categorization.

In [32], MFCC, CQCC, and Mel-filter bank slope features were employed to train the GMM to capture the vocal tract information and then distinguish the fake audio. Reimao [14] employed the CQT method to convert the audio signals into visual audio representations. These produced representations were fed into the Mobile network model to detect the synthetic speech. The linear filter banks' low-level features were extracted from audios. Then, these features were fed into the ResNet model to produce deep feature representations and detect audio manipulation. In addition, the online frequency masking augmentation layer and the large margin cosine loss function were employed during training the Residual network to learn more robust key feature embeddings [33]. WU et al. [34] employed the long-term CQT and log power spectrum to extract audios representation. This representation was used as an input to the feature genuinization method. This method learned a transformer with a CNN which was based on genuine speech characteristics. It aimed to maximize the difference between the distribution of genuine and synthetic speeches. Then, the transformed features were utilized with a light CNN model to detect the synthetic speech. In [35], the audio signals were converted into spectrogram images using the Fast Fourier transform method. These images were fed as an input into a CNN to validate audio signal authenticity. The work of [36] utilized Linear Frequency Cepstral Coefficients to convert raw audios into feature vector representations. Then, these representations were fed to a fine-tuned ResNet-18. In addition, a one-class Softmax loss function was proposed to

learn an embedding feature space in which the genuine speech had a compact boundary while the fake data was isolated from the genuine one by a certain margin. In [37], the log magnitude spectrograms were extracted from audio files. Then, a light convolution gated recurrent network was employed on these spectrograms to produce deep features and discriminate the real speech from the spoofed one. The work of [38] employed the short-term zero-crossing rate and energy to select the silent segments from each speech signal. Then, the linear filter bank features were extracted and fed into an attention-enhanced DenseNet Bi-LSTM model to identify audio manipulations. In [39], an end-to-end model which is based on the Residual network was proposed to extract deep features of audio data and then detect the synthetic speech.

Some researchers introduced approaches based on learning from different modalities to detect deepfakes. These approaches, which are often known as deepfake multimodal-video detection methods, can be categorized into CNN-based methods [40, 41, 42, 43], and methods that are based on using CNN and temporal network [44, 45]. This is depicted in Fig. 3.

The work of [40] exploited the perceived emotion cues from speech and face modalities to detect any manipulation in a video. It employed the OpenFace-V1 technique to extract the facial features and the PyAudioAnalysis library to extract the MFCC speech features. Then, the Siamese network-based architecture and the triplet loss were utilized to model the similarity between both modalities within a video and distinguish the fake content. Since any modification of visual video frames or audio modality within a video lead to a loss of lip synchronization, and abnormal lip and facial movements, a multimodal video deepfake detection method was introduced [41]. This method was based on computing the dissimilarity score between visual video and auditory segments. The 3D-Residual network-based architecture was used for extracting visual video features from face segments, and the raw audio segments were converted into MFCC features and then fed into CNN. The contrastive loss was estimated over audio and visual video features for each segment, which forced the real representation of both modalities to be closer than the manipulated one. Additionally, the cross-entropy loss was applied on every single modality to confirm that each one independently learns informative features. The work of [42] presented a multimodal video deepfake detection method based on discovering the defects in manipulated mouth areas via employing genuine audio as a reference. The audios were aligned and clipped into partitions based on phonemes, and Mel-scale spectrograms were extracted and used as audio features. The mouth frames were extracted from videos based on facial landmarks using the dlib python library. Then, each mouth frame with a particular phoneme interval was matched to a fixed-length audio partition to produce auditory-visual video pairs. After that a CNN architecture was trained on these pairs to capture the synchronization degree between lip movements and speech by measuring the similarity score of auditory-visual video pairs. Zhou and Lim [43] employed the asynchrony property between fake visual video, especially mouth movements, and speech to detect any modification within a video. The Multi-Task CNN (MTCNN) was utilized for detecting the face from video frames and the Residual (2+1)D-18 network was applied for extracting visual feature representations of these frames. For audio, a simple 1D convolution network was utilized for extracting 1D waveform signal feature representation. In addition, a sync-stream was built by applying central connections to visual video and audio network feature representations between low-level features; spatial and temporal information, to higher-level semantic representations. At each layer, the representations of visual video and auditory modalities were fused with the current layer of sync-stream. The output of this was utilized as an input to the fusion at the next layer. This helped in modelling the synchronization patterns of both modalities and distinguishing the deepfakes. Based on the observation that machines cannot recreate human emotions naturally in manipulated videos, Gino [44] introduced a deepfake detection method depending on exploiting emotion features from visual video and audio modalities. The low-level descriptors (LLDs) were extracted from raw audio segments using the OpenSmile toolkit and passed to the LSTM architecture to extract emotional features of speech. In addition, the face frames were detected from videos using the BlazeFace tool and then passed into 3D-CNN architecture to extract visual emotional features. After that, two approaches were followed in the final deepfake detection phase. In the first one, the visual and auditory emotional features were combined horizontally. Then, these features were fed either into the LSTM network or into Lazypredict models. In the second, the average between the prediction scores returned by training the LSTM and Lazypredict models on the visual video and auditory modalities separately was computed. The work of [45] detected the fake content in videos by extracting visual video and auditory emotional features and passing them to a deep network. The OpenFace-V2 toolkit was employed for extracting 31 visual features related to the intensity of facial muscle actions, eye gaze, and head position. The python_speech_features library was used to extract 13 MFCC features and their respective derivatives; delta MFCC, and delta-delta MFCC, as audio features. The visual video and auditory features were normalized and concatenated to be passed into CNN blocks that were followed by two Bi-LSTM networks and dense layers for deepfake detection.

A few deepfake detection methods are concerned with multimodal videos. However, they do not consider whether a video is manipulated only on the visual video frames level, audio level, or bimodal level which combines visual frames and audio. Consequently, this paper introduces a novel smart deepfake video detection system that can check whether the manipulation is just applied to video frames, audio, or both. It then produces the final decision for detecting the deepfakes on these three levels: visual frames, audio, and the whole video.

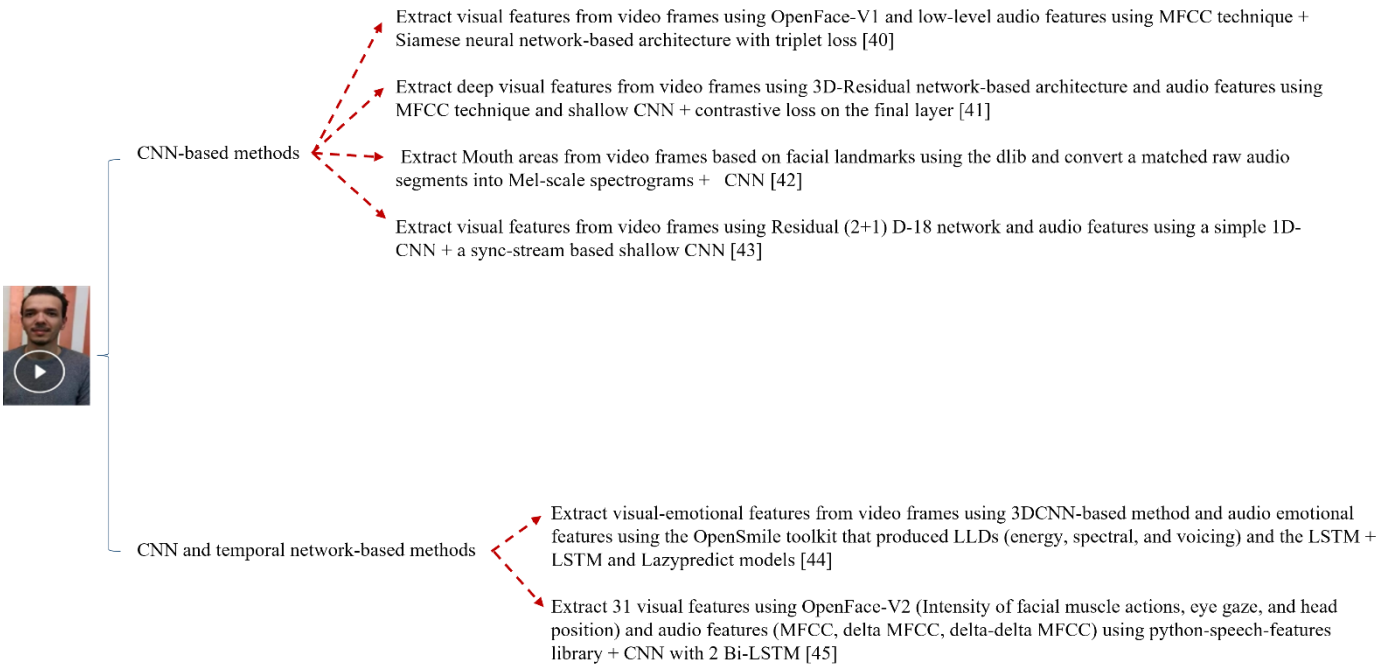


Fig. 3. The deepfake multimodal-video detection methods categorization.

III. Proposed Method

The suggested deepfake video detection method consists of three base stages: pre-processing, feature extraction using unimodal and bimodal information, and classification. These stages are shown in Fig. 4 and each one will be described hereafter.

A. Pre-processing

The visual video and audio modalities are pre-processed individually. The face frames are extracted from videos and saved separately. They are rescaled to the size 224 × 224, and their pixel values are normalized into [-1,1]. These pre-processed face frames will become an input to the next stage for learning and extracting deep visual video features. The raw audio files are extracted from videos and stored separately in a wave format. Then, the audio files are segmented. The CQT method is applied to every audio segment to produce a time-frequency representation of these segments. The CQT method is used for transforming audio signals from the time domain to the time-frequency domain. In CQT, frequency bins are geometrically spaced and ratios between centre frequencies and bandwidths, which are called Q-factors, of all bins are equal [47, 12]. The CQT of a discrete audio signal x(n) in the time domain is computed by the following formula [10, 46, 47, 48]:

$$X(m,n) = \sum_{k=n-\lfloor N_m/2 \rfloor}^{n+\lfloor N_m/2 \rfloor} x(k)\, a_m^*\left(k - n + \frac{N_m}{2}\right) \quad (1)$$

where $m = 1, 2, \dots, M$ represents the m^{th} index of frequency bin, $\lfloor . \rfloor$ denotes the floor function, and x(k) represents the k^{th} sample of a speech time-domain frame. The symbol $N_m = \frac{f_{sr}}{f_m} Q \in R$ indicates window lengths, f_{sr} represents the sampling rate frequency, and $f_m = f_1 2^{\frac{m-1}{b}}$ indicates the centre frequency of the m^{th} bin. The symbol f_1 denotes the centre frequency of the lowest bin, b refers to the bins number per octave and practically it determines a trade-off between time and frequency resolution. The factor $Q = \frac{f_m}{f_{m+1} - f_m} = \frac{1}{2^{\frac{1}{b}} - 1}$ produces a constant frequency to resolution ratio for each bin. The term $a_m^*(n)$ represents the complex conjugate of the complex-valued time-frequency atoms $a_m(n)$ which is defined as follows:

$$a_m(n) = \frac{1}{C}\, w\left(\frac{n}{N_m}\right) e^{i\left(2\pi n \frac{f_m}{f_{sr}} + \Phi_m\right)} \quad (2)$$

where $w(t)$ denotes a window function; Hann (Wang et al. 2019) [49], which is sampled at points specified by $\frac{n}{N_m}$. It is zero when t does not belong to [0,1]. The $C = \sum_{l=-\lfloor N_m/2 \rfloor}^{\lfloor N_m/2 \rfloor} w\left(\frac{l + N_m/2}{N_m}\right)$ represents a scaling factor, and Φ_m represents a phase offset.

The CQT computations are implemented using the librosa python library. The audio files are resampled to 22,050 Hz. A frequency bins number of 84 with 12 bins per octave, a hop length of 128 samples, and a minimum frequency value of approximately 65 Hz are used during the CQT calculations. In addition, the Hann window function is applied. The output of the CQT is then transformed into a log scale; decibels, to cope with the wide range of sound intensity. This produces a decibel-scaled spectrogram that has the shape T×84 per audio segment where T relies on the audio file duration. The duration of audio files adopted here is three seconds and accordingly, T is equal to 65. The spectrograms are normalized into the range [-1,1] and then reshaped to (65,84,1) as three-channel images. They will become an input into the next stage to learn and extract deep auditory features.

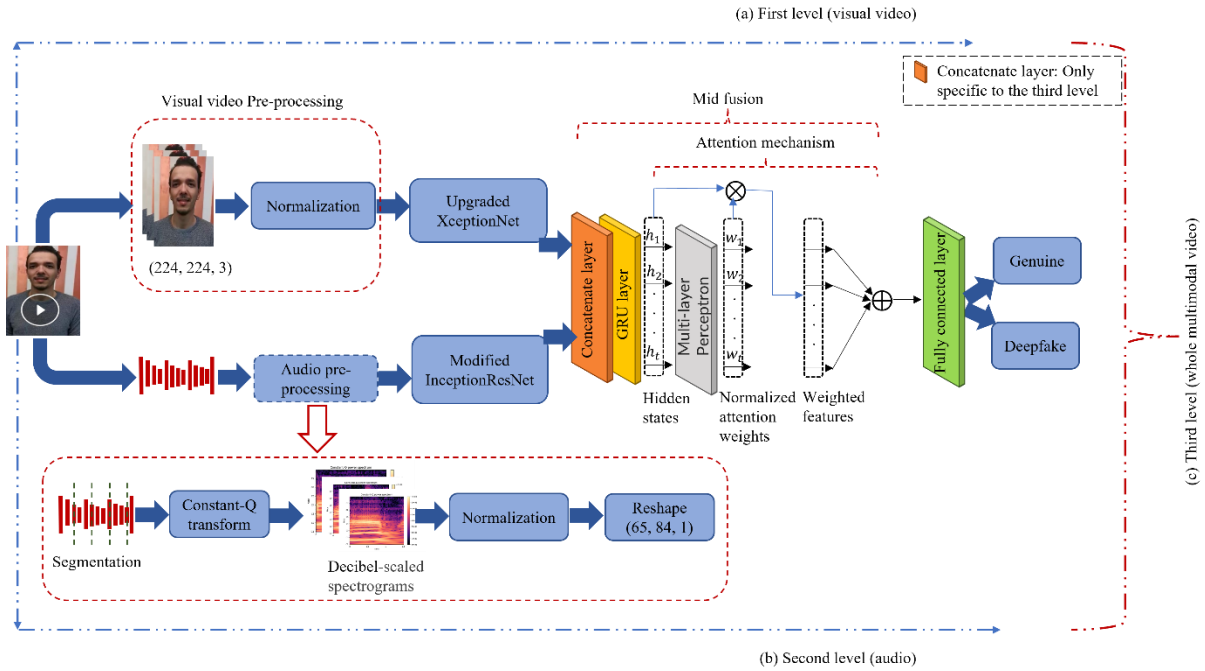


Fig. 4. The proposed smart deepfake video detection system architecture.

B. *Feature Extraction using Unimodal and Bimodal Information*

In this stage, the problem of deepfake detection is handled based on proposing two feature extraction methods for visual video and audio modalities. An upgraded XceptionNet is suggested to extract instructive deep spatial features from pre-processed face frames of videos. It outputs a visual feature representation of the unimodality; video frames. A modified InceptionResNetV2 is suggested to apply on the CQT spectrograms representing audio files to extract deep time-frequency features from audios. It produces a feature representation of the unimodality; audio. Then, the corresponding extracted feature representations from these modalities are first fused. This outputs a feature vector representation of the whole video using bimodal information. After that, these various resultant representations are independently passed into the GRU-based attention mechanism. This helps to learn the significant temporal information from the sequential feature representation per video on three levels: visual video frames, audio, or the whole video. Finally, a fully connected layer is applied to produce the final prediction about video authenticity. These components are explained in detail in the following subsections.

1) Visual video frames features: The processed face frames of videos with the shape (h×w×3) are received as an input to the proposed upgraded Xception network where h=224, w=224, and 3 denote the height, width, and RGB channels per frame. The Xception original architecture consists of 36 convolutional layers divided into 14 modules. All modules have shortcut residual connections around them except for the first and last ones. The Xception comprises depth-wise separable convolution layers, which reduce the cost of convolution operation dramatically [50, 51]. The proposed upgraded Xception network architecture is depicted in Fig. 5. The original XceptionNet is upgraded by first injecting seven layers before the last rectified linear unit (ReLU) activation layer of the last module. These seven layers are convolution with 1536 filters, batch normalization, ReLU activation, convolution with 1024 filters, batch normalization, ReLU activation, convolution with 1024 filters, and batch normalization. The convolution layers produce more informative and exclusive feature maps that help to differentiate between real and fake visual videos. The batch normalization layers, which standardize the input, have the effect of drastically speeding up the training and improving the model's performance by providing a modest regularization. The ReLU activation layers, which give a value of zero for all negative input feature values, add a nonlinear property to the model allowing it to understand and learn complex structures in data. Then, the dropout layer that randomly drops out units with a rate of 0.2 is injected between the last ReLU activation and the global average pooling layers to prevent overfitting and boost the model's generalization. After that two layers are injected after the global average pooling layer; the fully connected layer with 1024 units and ReLU activation function, and the dropout layer with a rate of 0.5. After applying the upgraded XceptionNet to the face frames of videos, the output becomes a vector representation of 1024 features per frame. The suggested modifications to the Xception network attempt to generate an instructive spatial hierarchical representation of frames. This helps to improve the deepfake detection method performance in real-world circumstances; number equations consecutively.

2) Audio features: The CQT spectrograms of audio files with the shape (65, 84, 1) per segment are received as an input

to the proposed modified InceptionResNetV2. The InceptionResNetV2 original architecture is built by joining the inception blocks and the skip connections. Each InceptionResNet block contains convolutions of different-sized filters that are combined by skip connections. These skip connections prevent the degradation problem that occurred via deep structures and reduce the time of training [52].

The proposed modified InceptionResNetV2 architecture is depicted in Fig. 6. The original InceptionResNetV2 is modified first by decreasing the repeating times' number of Inception ResNet blocks; A, B, and C, from 5, 10, and 5 to 4, 7, and 3, respectively. Then, some layers are injected after the last InceptionResNet block C and before the global average pooling layer. These layers are convolution with 512 filters on a kernel size of 1×1, batch normalization, ReLU activation, a couple of convolutions with 1024 filters on a kernel size of 1×1 where each one is followed by batch normalization and ReLU activation, and a dropout with a rate of 0.2. After that a fully connected layer with 1024 units and ReLU activation function is injected between global average pooling and dropout layers. In addition, filter units, kernel size, and stride for some layers are altered as shown in Fig. 6. After applying the modified InceptionResNetV2 to audio files segments, the output becomes a vector representation of 1024 auditory features per segment. The proposed modifications to the InceptionResNetV2 aim to generate an informative deep time-frequency representation of audio segments. This aids to enhance the performance of the proposed deepfake detection method.

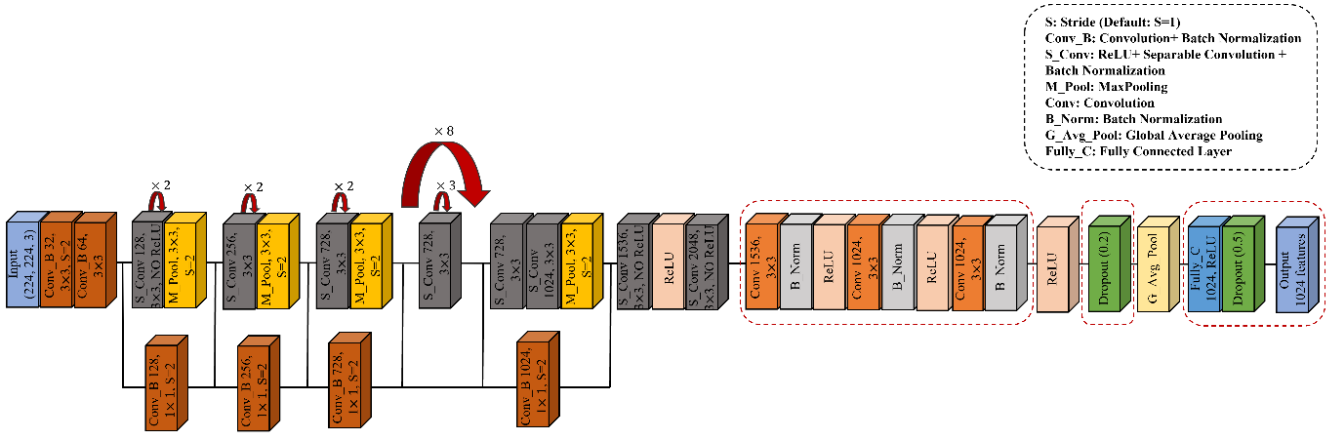


Fig. 5. The proposed upgraded xception network architecture.

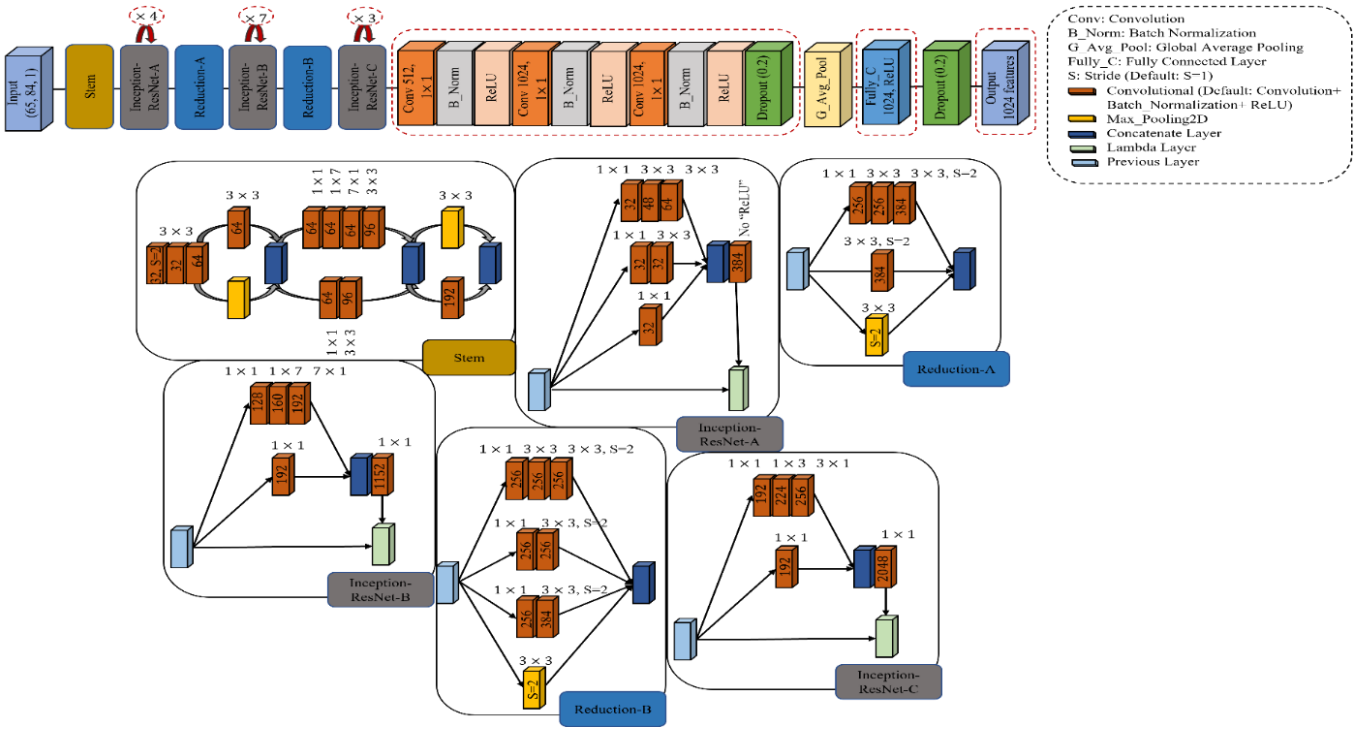


Fig. 6. The proposed modified inceptionResNetV2 architecture.

3) Bimodal information-based video features: The deep extracted features from visual video frames and audio modalities using the above-mentioned unimodality-based feature extraction methods are mid-fused at a concatenate layer. This produces a feature vector representation for the whole video, which is based on bimodal information.

4) Temporal information extraction-based attention mechanism: Most deepfake videos are generated based on synthesizing faces frame-by-frame, cloning voices, and synchronizing lips. They suffer from flickering and discontinuity of the face frames and lack of normal emotions, breathing, pauses, and the pace at which the target subject speaks among audio segments. As a result, the GRU-based attention mechanism is applied to the three levels of the extracted features independently; visual video frames, audio, and the whole video. This aims to capture the instructive temporal information that helps to differentiate real videos from fake ones.

The GRU architecture is composed of two gates; update (upd) and reset (res), that modulate the information flow from the previous time step to the current step. At each time step t, the update gate decides the amount of previous information that should be retained, and the reset gate determines the amount of information that needs to be forgotten [53]. The GRU hidden state h at the time t is defined by the following formulae [54]:

$$\mathrm{upd}_t = S\left(W_{\mathrm{upd}} x_t + U_{\mathrm{upd}} h_{t-1}\right) \quad (3)$$

$$\mathrm{res}_t = S(W_{\mathrm{res}} x_t + U_{\mathrm{res}} h_{t-1}) \quad (4)$$

$$\acute{h}_t = \tanh(W_h x_t + \mathrm{res}_t \circ U_h h_{t-1}) \quad (5)$$

$$h_t = (1 - \mathrm{upd}_t) \circ \acute{h}_t + \mathrm{upd}_t \circ h_{t-1} \quad (6)$$

where x refers to the input, and W and U represent the weight matrices. The symbol $S(.)$ represents the sigmoid function, $\tanh(.)$ represents the Hyperbolic Tangent, $\circ$ denotes the Hadamard product, and $\acute{h}_t$ denotes the candidate hidden state. As can be seen in Fig. 4, a single GRU is applied to the above-mentioned feature representations on the three levels. It produced a matrix of hidden state vectors at each time step t, which represents the learned temporal information per visual video, audio, or the whole video. The hidden state vector is defined as follows:

$$H = [h_1, h_2, \ldots, h_t] \quad (7)$$

The attention mechanism uses the weights to concentrate on the important features from the input sequence H. It is defined by the following equations [17, 55]:

$$u_t = \tanh(W h_t + b) \quad (8)$$

$$\mathrm{alpha}_t = \mathrm{softmax}(u_t) \quad (9)$$

$$c_t = \mathrm{alpha}_t h_t \quad (10)$$

$$v = \sum_t c_t \quad (11)$$

where u_t is a result of feeding a hidden vector h_t into a single-layer Multi-Layer Perceptron (MLP) with the tanh activation function. W represents the weight matrix, and b refers to the bias term. The symbol alpha_t represents the normalized attention weights that are produced by applying the softmax layer to u_t. v is a video representation that is formed by summing hidden vectors h_t weighted by attention weights alpha_t.

C. Classification

After the instructive temporal features are produced from the GRU-based attention mechanism, a fully connected layer is used as an output layer with two classes. Softmax function is used to decide deepfake videos from real ones. The Softmax formula is defined as follows:

$$\mathrm{Softmax}\,(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}} \quad (12)$$

where $y_i$denotes the values resulting from the output layer neurons.

D. Dataset

The proposed method has been evaluated on the FakeAVCeleb multimodal videos dataset. This dataset consisted of 490 celebrity genuine videos that were selected from the VoxCeleb2 dataset based on various ethnic groups, gender, and age. Its genuine videos are face-centered and cropped. The fake videos of the FakeAVCeleb dataset were generated using DeepFaceLab, Faceswap, and FSGAN, while fake audios were generated using a real-time voice cloning tool (SV2TTS). Additionally, the Wav2Lip was applied to the deepfake videos to re-enact these videos based on the cloned audios. Thus, the FakeAVCeleb dataset had more realistic deepfakes. The FakeAVCeleb was divided into four groups; genuine visual videos with genuine audios, genuine visual videos with deepfake audios, deepfake visual videos with genuine audios, and deepfake visual videos with deepfake audios [4].

To evaluate the proposed method, 1215 genuine and deepfake videos of the FakeAVCeleb dataset are employed. These videos are divided into three subsets: training, validation, and testing.

IV. Experimental Results and Analysis

The proposed deepfake video detection method is evaluated by the FakeAVCeleb dataset. Its performance is assessed using the following evaluation metrics [56]:

$$\mathrm{precision} = \frac{\mathrm{True\_Positives}}{\mathrm{True\_Positives} + \mathrm{False\_Positives}} \quad (13)$$

$$\mathrm{sensitivity} = \mathrm{recall} = \frac{\mathrm{True\_Positives}}{\mathrm{True\_Positives} + \mathrm{False\_Negatives}} \quad (14)$$

$$F_1 - \mathrm{score} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \quad (15)$$

$$\mathrm{accuracy} = \frac{\mathrm{True\_Positives} + \mathrm{True\_Negatives}}{\mathrm{True\_Positives} + \mathrm{True\_Negatives} + \mathrm{False\_Negatives} + \mathrm{False\_Positives}} \quad (16)$$

$$\mathrm{specificity} = \frac{\mathrm{True\_Negatives}}{\mathrm{True\_Negatives} + \mathrm{False\_Positives}} \quad (17)$$

$$\mathrm{AUROC} = \int_0^1 \mathrm{sensitivity}\left((1 - \mathrm{Specificity})^{-1}(x)\right) dx = p(x_2 > x_1) \quad (18)$$

where True_Positives denotes deepfake samples' number that is correctly predicted. The False_Positives represents genuine samples' number that is incorrectly predicted. The False_Negatives denotes deepfake samples' number that is incorrectly predicted. The True_Negatives refers to genuine samples' number that is correctly predicted. The symbol x_2 represents the predicted deepfake samples and x_1 denotes the predicted genuine samples. The higher the AUROC curve metric, the better the fake video detection method's performance at individuating the deepfake videos from the genuine ones.

The following three experiments are applied to the FakeAVCeleb dataset:

Experiment 1: This experiment represents applying the proposed method to the FakeAVCeleb videos dataset for two levels; visual video frames and audio. The visual video frames and audio modalities are trained end-to-end separately. Thus, a single GRU-based attention mechanism with 1024 units is independently applied to the visual video features that are extracted using the proposed upgraded XceptionNet and the audio features that are extracted using the proposed CQT based modified IncetionResNetV2. This learns the instructive temporal features for each unimodality. The visual video modality is trained for 32 epochs using the stochastic gradient descent (SGD) optimizer [57] with a learning rate of $2e^{-3}$ which is decayed by $4e^{-10}$, and a momentum of 0.9. The audio modality is trained for 27 epochs using the adaptive moment (Adam) optimizer [58] with a learning rate of e^{-3}. The batch size is 32. Then, the predictions are produced per modality. The performance of visual video and audio on the FakeAVCeleb dataset is shown in Table I and Table II, respectively. The proposed upgraded XceptionNet with a single GRU-based attention mechanism for the visual video modality has achieved 98.51% accuracy and 98.45% AUROC outperforming recent state-of-the-art methods by a large margin. Additionally, the proposed CQT based modified InceptionResNetV2 with a single GRU-based attention mechanism for the audio modality has achieved 97.52% accuracy and 97.62% AUROC outstanding other state-of-the-art methods by a large margin.

Experiment 2: In this experiment, the prediction results of visual video frames and audio modalities from experiment 1 are employed to produce the prediction for the whole video. Thus, the whole multimodal-video prediction is decided to be genuine if both modalities are predicted as genuine, otherwise, it's deepfake. Experiment 2 performance for multimodal video deepfake detection is recorded in Table III. It has yielded 96.04% accuracy and 95.49% AUROC.

Experiment 3: This experiment represents applying the proposed method to the FakeAVCeleb videos dataset for the third level; whole multimodal video. As the FakeAVCeleb dataset is distributed into four groups: genuine visual videos and audios, genuine visual videos with fake audios, fake visual videos with genuine audios, fake visual videos and audios, the whole video label (y_i) is considered genuine if the label of both visual video (y_{iv}) and audio (y_{ia}) modalities are genuine, otherwise, it's fake. This can be defined as follows:

$$y_i = \begin{cases} 0\,, \text{if } y_{iv} = y_{ia} = 0 \\ 1, \quad \text{otherwise} \end{cases} \tag{19}$$

The single GRU-based attention mechanism with 3572 units is applied to the bimodal information-based video features. This helps to learn the instructive temporal features for the whole multimodal video. The details of GRU-based attention mechanism layers that are applied on top of bimodal information-based video features are described in Table IV. The proposed method is trained for 24 epochs using the SGD optimizer with a learning rate of $2e^{-3}$ and a decay factor of $4e^{-10}$, and a momentum of 0.9. This is employed to update the weight parameters and is aimed to minimize the difference between actual and predicted labels. The batch size is set to 64.

TABLE I. THE PERFORMANCE OF THE UPGRADED XCEPTIONNET METHOD WITH SINGLE GRU-BASED ATTENTION MECHANISM FOR DETECTING THE DEEPFAKE VISUAL VIDEO UNIMODALITY COMPARED TO RECENT STATE-OF-THE-ART METHODS ON THE FAKEAVCELEB DATASET

Model	Unimodality	
	Visual video	
	Accuracy	AUCROC
Experiment 1 (The proposed upgraded XceptionNet with GRU-based attention mechanism for the first level)	**98.51%**	**98.45%**
VGG16 [60]	81.03%	81.04%
Xception [7]	73.06%	73.07%

TABLE II. THE PERFORMANCE OF THE MODIFIED INCEPTIONRESNETV2 METHOD WITH SINGLE GRU-BASED ATTENTION MECHANISM FOR DETECTING THE DEEPFAKE AUDIO UNIMODALITY COMPARED TO RECENT STATE-OF-THE-ART METHODS ON THE FAKEAVCELEB DATASET

Model	Unimodality	
	Audio	
	Accuracy	AUCROC
Experiment 1 (The proposed CQT based modified InceptionResNetV2 with GRU-based attention mechanism for the second level)	**97.52%**	**97.62%**
Mel-frequency cepstrum (MFC)+ VGG16 [60]	67.14%	67.13%
MFC+ Xception [60]	76.26%	76.25%

(CQT [61] + MobileNet) [14]	82.67%	82.38%

TABLE III. THE PERFORMANCE OF THE PROPOSED METHOD FOR DETECTING WHOLE MULTIMODAL VIDEO DEEPFAKES COMPARED TO RECENT STATE-OF-THE-ART METHODS ON THE FAKEAVCELEB DATASET

Model	Bimodal	
	Visual video and audio	
	Accuracy	AUCROC
Experiment 2	**96.04%**	**95.49%**
Experiment 3 (The proposed method for the third level: whole multimodal video)	97.52%	97.21%
Ensemble Soft/ hard voting based VGG16 [60]	78.04%	78.05%
Two CNN blocks (one per modality) [60]	67.4%	67.2%
Xception [7]	43.94%	43.73%

TABLE IV. THE GRU-BASED ATTENTION MECHANISM LAYERS DETAILS

Layer (type)	Output shape	Parameters number
main_input (Input Layer)	[(None, 8, 4096)]	0
gru (GRU)	(None, 8, 3572)	82191720
attention (attention)	(None, 3572)	3580
Total parameters: 82,195,300 Trainable parameters: 82,202,446 Non-trainable parameters: 0		

The cross-entropy loss (l) function is utilized to measure the efficiency of the suggested deepfake video detection method on three levels: video frames, audio, and the whole video. Its formula [59] is defined as follows:

$$l = -\frac{1}{M}\sum_{k=1}^{M}(y_k \log(p_k) + (1 - y_k)\log(1 - p_k)) \quad (20)$$

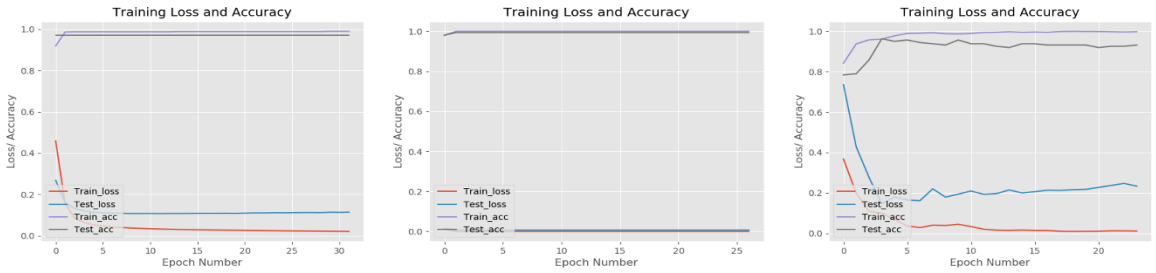
where M refers to the number of visual videos, audios, or whole videos. The y_k and p_k denote the actual label and predicted probability corresponding to the k^{th} video. It can be seen in Table III that the proposed method, which represents experiment 3, for whole multimodal video deepfake detection has achieved 97.52% accuracy and 97.21% AUROC. Its performance exceeds that of experiment 2 because experiment 2 is unable to learn intercorrelations between different modalities. Additionally, it outperforms recent state-of-the-art methods by an average growth of 34.4% accuracy and 34.2% AUROC as can be seen in Table III.

The experiments are carried out using an OMEN HP laptop with a 16-gigabyte Intel (R) Core (TM) i7-9750H CPU, a 6-gigabyte RTX 2060 GPU, and Windows 11. The proposed method is implemented using the Python programming language. Python libraries such as Keras, OpenCV, Random, Tensorflow, Numpy, OS, and Librosa are used during the implementation.

The accuracy and loss curves of the proposed method on the training and validation subsets of the FakeAVCeleb dataset for the three levels; visual video frames, audio, and whole multimodal videos, are shown in Fig. 7. Additionally, the proposed method confusion matrix for deepfake video detection on the three levels is depicted in Fig. 8. Furthermore, Fig. 9 shows the receiver operating characteristic (ROC) curve and the AUROC curve of the proposed method performance. As shown in Fig. 9, the ROC curve is extremely close to the top left ensuring the high performance of the proposed method.

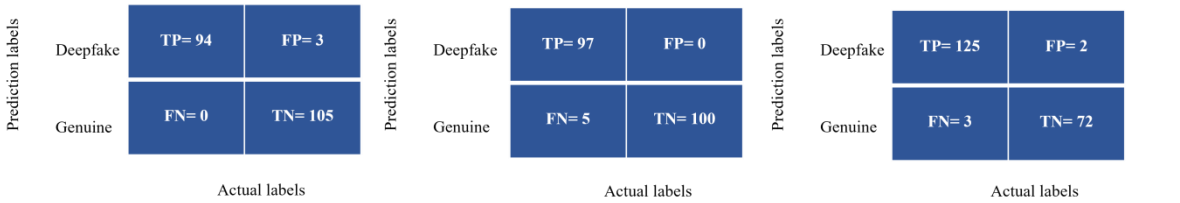
Fig. 10 provides a comparison of the proposed method with contemporary state-of-the-art methods using evaluation metrics. As shown in Fig. 10, the proposed method has yielded better performance in comparison to the other methods on the three levels. It has a precision of 96.91%, recall of 100%, F1-score of 98.43%, and specificity of 97.22% for detecting visual videos. Additionally, it has a precision of 100%, recall of 95.10%, F1-score of 97.49%, and specificity of 100% for detecting audios. Further, it has a precision of 98.43%, recall of 97.66%, F1-score of 98.04%, and specificity of 97.30% for detecting whole multimodal videos.

It can be concluded that the proposed upgraded XceptionNet generated a useful spatial hierarchical representation of faces, which contributed to distinguishing between genuine and fake videos. As well, the proposed CQT-based modified InceptionResNetV2 produced a valuable deep time-frequency representation of audio. This assisted to reveal deepfake videos and improved the detection method's effectiveness. Moreover, a concatenate layer that is applied to the features extracted from visual video and audio modalities produced an informative bimodal representation of videos. In addition, the GRU-based attention mechanism, which is applied to the visual video, audio, and bimodal features, assisted in capturing the most important temporal information of videos. This in turn helped to detect the deepfakes. Furthermore, it can be inferred that correlating features from different modalities can improve the chances of achieving accurate deepfake video detection.



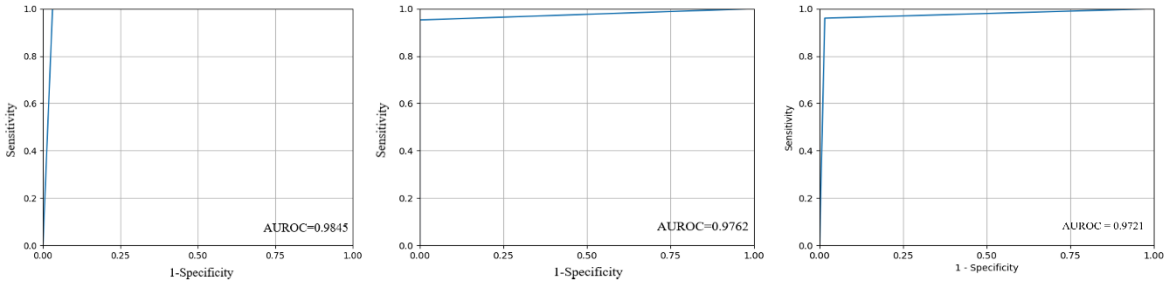
(a) First level (visual video) (b) Second level (audio) (c) Third level (whole multimodal video)

Fig. 7. The accuracy and loss curves of the proposed deepfake video detection method on training and validation sets.



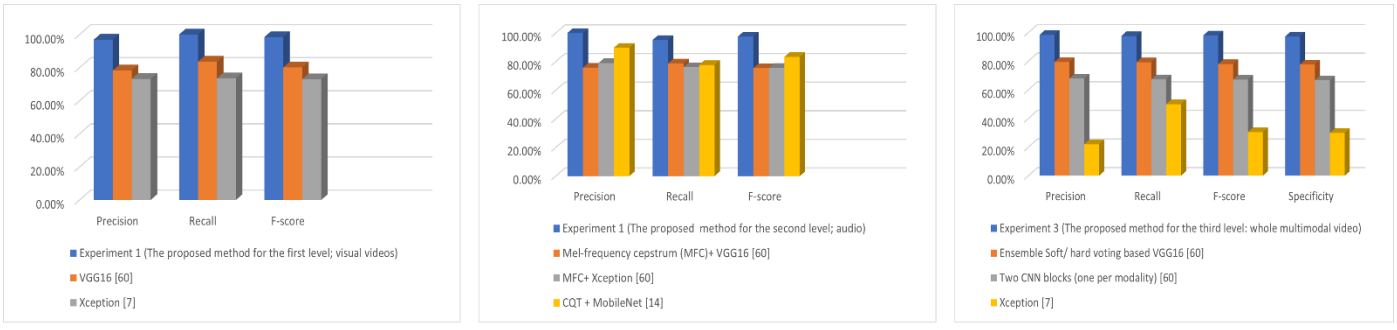
(a) First level (visual video) (b) Second level (audio) (c) Third level (whole multimodal video)

Fig. 8. The confusion matrix visualization of the proposed deepfake video detection method.



(a) First level (visual video) (b) Second level (audio) (c) Third level (whole multimodal video)

Fig. 9. The ROC curve and the AUROC curve of the proposed deepfake video detection method performance.



(a) First level (visual video) (b) Second level (audio) (c) Third level (whole multimodal video)

Fig. 10. The evaluation metrics of the proposed deepfake video detection method compared to recent state-of-the-art methods on the FakeAVCeleb dataset.

V. Conclusion and Future Work

A newly smart system for detecting video deepfakes has been presented. Two methods were proposed to extract features from visual video frames and audio modalities, respectively. These methods produced useful spatial information for visual video and valuable time-frequency information for audio, which improved the performance of the deepfake detection method. In addition, the feature representations of both modalities were passed into a mid-layer to produce an informative bimodal representation per video. It proved that using bimodal information boosts learning during training compared to the method that ignores intercorrelation between modalities. The GRU-based attention mechanism was then applied to the different feature representations to extract the most significant temporal information and detect the

deepfakes. The proposed method has been evaluated on the FakeAVCeleb multimodal videos dataset. It achieved 98.51% accuracy, 98.45% AUROC, 96.91% precision, 100% recall, 98.43% F1-score, and 100% sensitivity on the first level; visual videos. Additionally, it yielded 97.52% accuracy, 97.62% AUROC, 100% precision, 95.10% recall, 97.49% F1-score, and 95.10% sensitivity on the second level; audios. Moreover, it attained 97.52% accuracy, 97.21% AUROC, 98.43% precision, 97.66% recall, 98.04% F1-score, 97.66% sensitivity, and 97.30% specificity on the third level; whole multimodal videos. Consequently, the proposed method outperformed the current state-of-the-art methods by a large margin.

In the future, several optimization algorithms can be employed to enhance the performance of the proposed deepfake video detection method. Furthermore, a huge multimodal video dataset may be utilized to improve the detection method's performance.

REFERENCES

[1] Antoniou A (2019) Zao's deepfake face-swapping app shows uploading your photos is riskier than ever. The Conversation.

[2] Perov I, Gao D, Chervoniy N, Liu K, Marangonda S, Umé C, Dpfks M, Facenheim SC, RP L, Jiang J, Zhang S, Wu P, Zhou B, Zhang W (2020) DeepFaceLab: a simple, flexible and extensible face swapping framework.

[3] Masood M, Nawaz M, Malik KM, Javed A, Irtaza A (2021) Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. arXiv preprint arXiv:2103.00484

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

[5] Jia Y, Zhang Y, Weiss RJ, Wang Q, Shen J, Ren F, Chen Z, Nguyen P, Pang R, Moreno IL, Wu Y (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. arXiv preprint arXiv:1806.04558.

[6] Prajwal KR, Mukhopadhyay R, Namboodiri VP, Jawahar CV (2020, October) A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the 28th ACM International Conference on Multimedia, pp 484-492.

[7] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1-11.

[8] Kumar A, Bhavsar A, Verma R (2020) Detecting deepfakes with metric learning. In 2020 8th international workshop on biometrics and forensics (IWBF), IEEE, pp 1-6.

[9] Singh P, Saha G, Sahidullah M (2021, January) Non-linear frequency warping using constant-Q transformation for speech emotion recognition. In 2021 International Conference on Computer Communication and Informatics (ICCCI), IEEE, pp 1-6.

[10] Schörkhuber C, Klapuri A (2010, July) Constant-Q transform toolbox for music processing. In 7th sound and music computing conference, Barcelona, Spain, pp 3-64.

[11] Delgado H, Todisco M, Sahidullah M, Sarkar AK, Evans N, Kinnunen T, Tan ZH (2016, December) Further optimisations of constant Q cepstral processing for integrated utterance and text-dependent speaker verification. In 2016 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp 179-185.

[12] Lidy T, Schindler A (2016, September) CQT-based convolutional neural networks for audio scene classification. In Proceedings of the detection and classification of acoustic scenes and events 2016 workshop (DCASE2016), IEEE Budapest, Vol 90, pp 1032-1048.

[13] Halpern BM, Kelly F, van Son R, Alexander A (2020). Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure.

[14] Reimao RAM (2019) Synthetic speech detection using deep neural networks.

[15] Li X, Li N, Weng C, Liu X, Su D, Yu D, Meng H (2021, June) Replay and synthetic speech detection with res2net architecture. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp 6354-6358.

[16] Shen G, Tan Q, Zhang H, Zeng P, Xu J (2018) Deep learning with gated recurrent unit networks for financial sequence predictions. Procedia computer science 131: 895-903.

[17] Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016, August) Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th annual meeting of the association for computational linguistics, Vol 2, pp 207-212.

[18] Cheng Q, Fu Y, Huang J et al (2022) Event detection based on the label attention mechanism. Int. J. Mach. Learn. & Cyber. https://doi.org/10.1007/s13042-022-01655-y.

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

[20] Hu Z, Xie H, Wang Y, Li J, Wang Z, Zhang Y (2021, January) Dynamic Inconsistency-aware DeepFake Video Detection. In IJCAI.

[21] Rana MS, Sung AH (2020, August) Deepfakestack: a deep ensemble-based learning technique for deepfake detection. In 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), IEEE, pp 70-75.

[22] Ismail A, Elpeltagy M, S Zaki M, Eldahshan K (2021) A New deep learning-based methodology for video deepfake detection using XGBoost. Sensors 21(16): 5413.

[23] Güera D, Delp EJ (2018, November) Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal-based surveillance (AVSS), IEEE, pp 1-6.

[24] Li Y, Chang MC, Lyu S (2018, December) In ictu oculi: exposing ai created fake videos by detecting eye blinking. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, pp 1-7.

[25] Zheng Y, Bao J, Chen D, Zeng M, Wen F (2021) Exploring temporal coherence for more general video face forgery detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 15044-15054.

[26] Gu Z, Chen Y, Yao T, Ding S, Li J, Huang F, Ma L (2021, October) Spatiotemporal inconsistency learning for deepFake video detection. In Proceedings of the 29th ACM International Conference on Multimedia, pp 3473-3481.

[27] Ismail A, Elpeltagy M, Zaki M, ElDahshan KA (2021) Deepfake video detection: YOLO-Face convolution recurrent approach. PeerJ Computer Science 7: e730.

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

[29] Khalil SS, Youssef SM, Saleh SN (2021) iCaps-Dfake: an integrated capsule-based model for deepfake image and video detection. Future Internet 13(4): 93.

[30] Ismail A, Elpeltagy M, Zaki MS, Eldahshan K (2022) An integrated spatiotemporal-based methodology for deepfake detection. Neural Comput & Applic. https://doi.org/10.1007/s00521-022-07633-3.

[31] Witkowski M, Kacprzak S, Zelasko P, Kowalczyk K, Galka J (2017, August) Audio replay attack detection using high-frequency features. In Interspeech, pp 27-31.

[32] Saranya MS, Padmanabhan R, Murthy HA (2018, July) Replay attack detection in speaker verification using non-voiced segments and decision level feature switching. In 2018 International Conference on Signal Processing and Communications (SPCOM), IEEE, pp 332-336.

[33] Chen T, Kumar A, Nagarsheth P, Sivaraman G, Khoury E (2020, November) Generalization of audio deepfake detection. In Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, pp 132-137.

[34] Wu Z, Das RK, Yang J, Li H (2020) Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. arXiv preprint arXiv:2009.09637.

[35] Bartusiak ER, Delp EJ (2021) Frequency domain-based detection of generated audio. Electronic Imaging 2021(4): 273-1-273-7.

[36] Zhang Y, Jiang F, Duan Z (2021) One-class learning towards synthetic voice spoofing detection. IEEE Signal Processing Letters 28: 937-941.

[37] Gomez-Alanis A, Peinado AM, Gonzalez JA, Gomez AM (2019, September) A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In Proc. Interspeech, Vol 2019, pp 1068-1072.

[38] Huang L, Pun CM (2020) Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM network. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28: 1813-1825.

[39] Hua G, Bengjinteoh A, Zhang H (2021) Towards end-to-end synthetic speech detection. IEEE Signal Processing Letters.

[40] Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020, October) Emotions don't lie: an audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM international conference on multimedia, pp 2823-2832.

[41] Chugh K, Gupta P, Dhall A, Subramanian R (2020, October) Not made for each other-audio-visual dissonance-based deepfake detection and localization. In Proceedings of the 28th ACM International Conference on Multimedia, pp 439-447.

[42] Gu Y, Zhao X, Gong C, Yi X (2020, November) Deepfake video detection using audio-visual consistency. In International Workshop on Digital Watermarking, Springer, Cham, pp 168-180.

[43] Zhou Y, Lim SN (2021) Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 14800-14809.

[44] Gino J (2021) Audio-video deepfake detection through emotion recognition.

[45] Mozley K (2021) Don't believe everything you hear: combining audio and visual cues for deepfake detection.

[46] Blankertz B (2001) The constant Q transform. URL http://doc. ml. tu-berlin. de/bbci/material/publications/Bla_constQ. Pdf.

[47] Schörkhuber C, Klapuri A, Sontacchi A (2012, September) Pitch shifting of audio signals using the constant-q transform. In Proc. of the DAFx Conference.

[48] Todisco M, Delgado H, Evans NW (2016, June) A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients. In Odyssey, Vol 2016, pp 283-290.

[49] Wang M, Wang R, Zhang XL, Rahardja S (2019, November) Hybrid constant-Q transform based CNN ensemble for acoustic scene classification. In 2019 Asia-pacific signal and information processing association annual summit and conference (APSIPA ASC), pp 1511-1516.

[50] Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251-1258.

[51] Ghosh S, Pal A, Jaiswal S, Santosh KC, Das N, Nasipuri M (2019) SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving. International Journal of Machine Learning and Cybernetics 10(11): 3145-3154.

[52] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017, February) Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

[53] Arfianti UI, Novitasari DCR, Widodo N, Hafiyusholeh M, Utami WD (2021) Sunspot number prediction using gated recurrent unit (GRU) algorithm. IJCCS (Indonesian Journal of Computing and Cybernetics Systems) 15(2): 141-152.

[54] Le XH, Ho HV, Lee G (2019, September) Application of gated recurrent unit (GRU) network for forecasting river water levels affected by tides. In International Conference on Asian and Pacific Coasts, Springer, Singapore, pp 673-680.

[55] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016, June) Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1480-1489.

[56] Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process 5(2): 1.

[57] Achlioptas P (2019) Stochastic gradient descent in theory and practice.

[58] Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[59] Ho Y, Wookey S (2019) The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. IEEE Access 8: 4806-4813.

[60] Khalid H, Kim M, Tariq S, Woo SS (2021b, October) 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, pp 7-15.

[61] Aly M, Alotaibi NS (2022) A New Model to Detect COVID-19 Coughing and Breathing Sound Symptoms Classification from CQT and Mel Spectrogram Image Representation using Deep Learning. International Journal of Advanced Computer Science and Applications, pp 601-611.