

(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.