

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

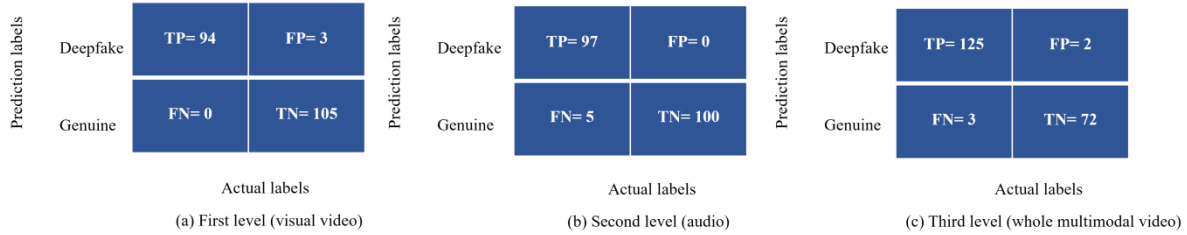


Fig. 8. The confusion matrix visualization of the proposed deepfake video detection method.

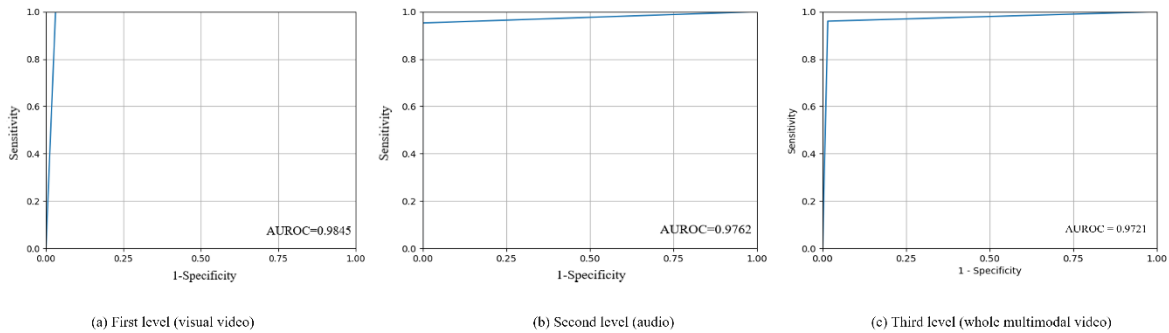


Fig. 9. The ROC curve and the AUROC curve of the proposed deepfake video detection method performance.

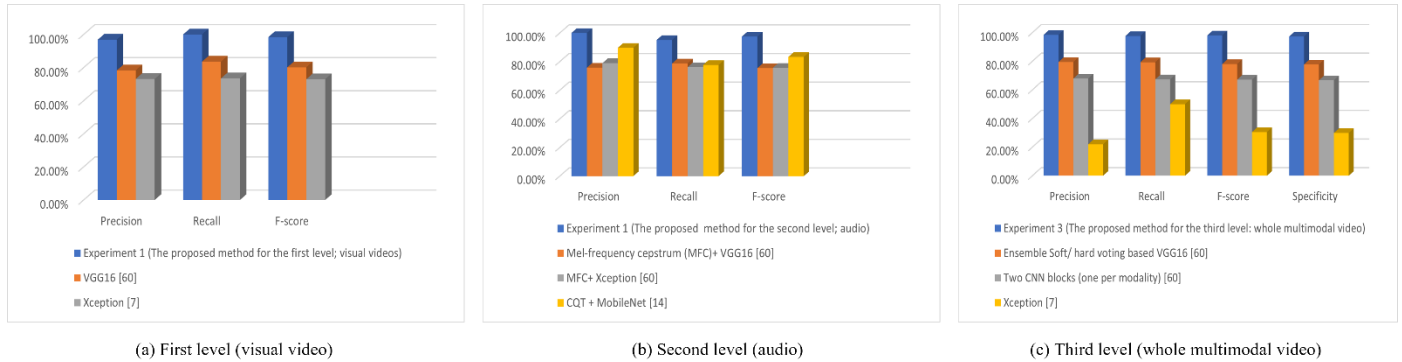


Fig. 10. The evaluation metrics of the proposed deepfake video detection method compared to recent state-of-the-art methods on the FakeAVCeleb dataset.

## 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