Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition

Alaa Alsaeedi[1], Amal AlMansour[2], Amani Jamal[3]
Computer Science Department, King Abdulaziz University, Jeddah, Saudi Arabia[1, 2, 3]
Computer Science Department, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia[1]

***Abstract*—Recently, there has been a significant reliance on the Internet. This creates a fertile environment for various risks, including fraud, privacy violations, and theft. The most common and dangerous risks at present are known as deepfakes. The development of deepfake technologies relies on advancements in artificial intelligence. Deepfake content can greatly affect privacy and security, posing a significant risk to many fields. Therefore, recent research has focused on mechanisms to detect real content from fake content. These mechanisms are classified into two main types: single-modal and multimodal detection. It is worth noting that the widespread deepfake technology has recently become more complex. This may hinder traditional single-mode detection methods in detecting video clips. In this study, we designed an effective multimodal fusion mechanism that integrates pre-trained audio, visual, and textual features. Our framework is based on three considerations: audio features, visual features, and emotion recognition. Emotion recognition focuses on three considerations: audio emotion, facial emotion, and sentiment of speech. We take advantage of the sentiment of speech to ensure there is consistency between audio and visual emotion with the meaning of words. As we achieved, the sentiment of speech makes our model more accurate and robust than when we used the audio-visual emotion inconsistency measures only. In our experiment, we used the FakeAVCeleb dataset, and we achieved 95.24% accuracy, proving our assumption of the impact of the sentiment of speech, the emotion of audio tone, and facial expressions to detect deepfakes.**

***Keywords*—Machine learning; deepfake; multimodal; sentiment of speech; emotion recognition**

I. Introduction

Recently, artificial intelligence technologies (AI) have witnessed significant development. This development has substantial uses, as AI technologies have become part of most areas of our daily lives [1]. Like any major advancement, AI development has both positive and negative aspects, leading to its use being categorized into these two sides [1]. Among the positive uses of AI, some have been used to enhance innovation and progress in various fields such as medicine, education, and industry. For example, AI has contributed to the development of medical technologies that help diagnose diseases more accurately, improve education by providing smart educational methods, and increase production efficiency in manufacturers [1]. On the contrary, some have exploited AI unethically to achieve harmful goals such as espionage or violating privacy, spreading misleading information, or manipulating public opinions by creating fake content, which is called deepfake [2], [1]. Deepfakes are one of the most concerning uses of these advancements. That involves digital manipulation of media such as text, audio, and visuals [2]. According to Spector [3], it is estimated that a substantial portion, around 50% of the billions of audio files, images, and videos uploaded daily on social and professional platforms, is manipulated. The term deepfake combines "deep learning" and "fake", indicating manipulation based on AI [2]. AI-based manipulations such as Variational Auto-Encoders (VAEs), Generative Adversarial Networks (GANs), and Diffusion Models (DMs) aim to reach unprecedented levels of realism [4]. Notably, deepfakes are utilized across various fields, including hyper-realistic computer-generated imagery (CGI), virtual reality (VR), augmented reality (AR), education, animation, art, and cinema. However, their deceptive nature also makes them susceptible to malicious use [5]. The major theft incident that occurred in 2021 sheds light on the severe financial risks associated with deepfake technology. Thieves employed AI-based voice cloning to mimic a bank official. That enabled them to fraudulently obtain $35 million from a UAE bank [6]. Facial expressions and speech are essential to human interaction, and it is crucial in biometric-based identity recognition. That means altered faces and voices present significant challenges to the integrity of online information and security systems [2].

Deepfakes can manifest in various media, all aiming to alter current content for the benefit of the manipulator. A deep fake can occur on a single modal such as text, audio, or visual, as shown in Fig. 1 and 2.

Otherwise, multimodal, typically shown with videos, such as audio-visual or text-visual manipulation, which is a more complicated deepfake technology, which is shown in Fig. 3 and 4, respectively [4].

Text-based deepfake means manipulating unprecedented volumes of online misinformation, such as fake news and rumors circulating across the internet, influencing public perceptions of significant social events [11]. Recent advances in neural generative models, like GPT-2, have exacerbated this issue, as these models can generate highly fluent and coherent text, potentially allowing adversaries to produce convincing fake news [11]. Audio-based deepfakes are artificially generated or altered audio recordings that convincingly mimic real speech [12]. Lastly, a visual-based deepfake is the replacement of a targeted individual's face in a video with that of another person by splicing a synthesized facial region into the original footage [5].

Fig. 1. Video deepfake from Celeb-DF dataset [7].

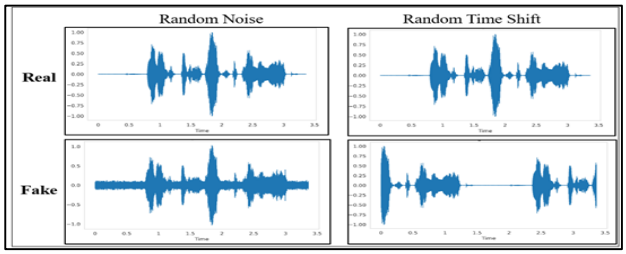

Fig. 2. Audio deepfake from ASVspoof 2019 TFRecord dataset[8]. Random noise means the background noise, and random time shift means changing the audio order.

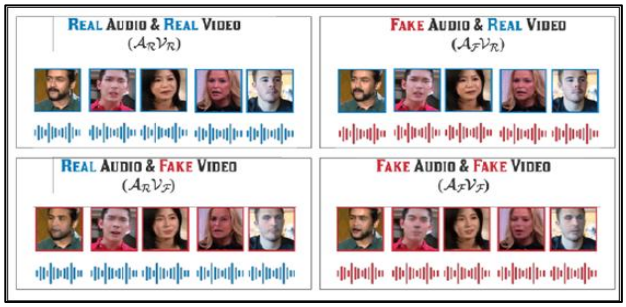

Fig. 3. Audio and visual multimodal manipulation [9].

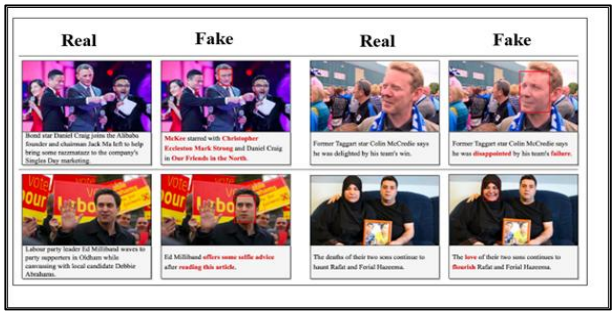

Fig. 4. Text and visual multimodal manipulation[10] .

Deepfake detection has become substantial in the digital age for various reasons. First, it helps stop misinformation, which can convince public opinion and have significant social and political effects [13]. Second, the integrity of news, social media, and legal evidence to support digital media authenticity [14]. Finally, since deepfakes can impersonate people for harmful purposes, it protects against financial fraud and identity theft [15]. To deepfake detection, researchers are developing and evaluating algorithms that can effectively identify manipulated media [4]. These detection models are categorized into two main groups based on the manipulation type [4]. Single-modal detection models focus on analyzing one specific type of data, such as text, image, or audio. Multimodal detection models integrate the analysis of many types of manipulations, combining images, audio, and text. Multi-modal better handles real-world deepfakes [4].

As noted above, multimodal detection is superior to single-modal detection due to the employment of many levels of detection. Thus, studies indicate that integrating emotion recognition, which means increasing detection levels, into deepfake detection provides a promising enhancement in the fight against advanced AI-generated media [16]. Emotion recognition is an aspect of artificial intelligence designed to interpret human emotions through nonverbal analysis [17]. Emotional recognition can be accomplished through several forms, including text, visual, and audio forms [17]. Text emotion recognition systems assess written text for emotions [18]. Classifying emotions like happiness, rage, and sadness includes processing textual signals, including word choice, syntax, and language patterns [18]. Visual emotion recognition employs computer vision to examine facial features in images and videos [17]. In audio emotion identification, AI algorithms use tone, pitch, and rhythm to infer emotions [17]. Conventional detection techniques frequently depend on perceptual discrepancies or technical defects. As the quality of deepfakes enhances, these indicators become increasingly difficult to discern. Consequently, examining emotional authenticity, including the intensity and genuineness of expressions, will enhance the distinction between authentic and altered information [16], [19], and [20].

Recent studies have shown that multimodal detection strategies, either visual-audio or text-audio, can give the deep learning model complementary information. However, the current strategies are still struggling with detecting fake content, especially with the datasets that use complex manipulation tools. The related works that used an emotional indicator achieved better accuracy than those without emotion detection. Nevertheless, there is no investigation about the concatenation of emotions with the sentiment of speech, "what I say and how I say it". Thus, we will get better information than the traditional emotion detection when we integrate the sentiment of speech—what I say—with audio emotion and facial emotion—how I say it.

Motivated by this gap, we introduce our novel framework, titled "Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition". A framework for developing an efficient multimodal deepfake detection system. Our method detects deepfakes based on emotional mismatches in three considerations.

The main contributions are as follows:

- Compute the mismatch between the audio tone and the facial emotion.
- Compute the mismatch between the audio tone and the sentiment of speech.
- Compute the mismatch between the facial emotion and the sentiment of speech.

- Perform an ablation study to analyze the impact of each field in our model with five experiments (audiovisual emotion model, sentiment of speech only model, audio emotion with sentiment of speech model, facial emotion with sentiment of speech model, and our approach that includes all of above model).

The structure of this study is as follows: Section II presents a review of the literature. Section III describes the proposed framework and methodology. Section IV details the experiments and reports the results. Section V provides a discussion of the findings. Finally, Section VI concludes the study and outlines future works.

II. Literature Review

As we mentioned earlier, deepfake techniques have recently become more accurate and sophisticated due to advances in artificial intelligence. Therefore, it is important to keep up with this development with effective detection tools. We categorize the literature reviews according to the approaches to detecting deepfake content. We break it into three subsections: single-modal, multimodal, and emotional deepfake detection approaches. Then we will present the existing datasets. This section will conclude with research gaps and our contribution.

A. Single-Modal Detection Approach

As we know, in single-modal detection, there are three types of methods for detecting deepfakes: textual, audio, or visual detection. In this subsection, we will present the literature reviews that utilize a single type of deepfake detection. Divided into sub-subsections are text-based, audio-based, and visual-based. This subsection concludes with an analytical table of the literature mentioned there.

1) Text-based approach. The following studies aim to propose a deepfake detection model based on textual content. Saravani et al. [21] examined a technique for identifying social media bots through the analysis of tweets. Utilizing the deepfake dataset comprising 25,572 tweets from bots mimicking human accounts, the model processed data and employed a classification architecture that included two fully connected layers. The authors attained a two per cent enhancement in accuracy over prior models by incorporating bidirectional encoder representations from transformers (BERT) for text representation that used natural language processing (NLP), a bidirectional long-short-term memory (BiLSTM) layer is a type of neural network used to preserve word order, and the neural network expand vector of locally aggregated descriptors (NeXtVLAD) layer for effective information summarization. Li et al. [22] used large language models (LLMs) that closely resemble human writing to study the challenges of distinguishing human-authored from machine-generated writing. The authors built their identification system using human-authored texts from Reddit comments, news headlines, and academic works. To ensure diversity, they used 27 LLMs and three prompt types to generate machine-generated writings. The detection system provided a probability score for classifications and used the Longformer neural network, which outperformed related models. Jensen-Shannon distance revealed that text variety inhibited the detection of linguistic trends. The detection method had an 86.54% recall rate on GPT- 4 texts. Zhong et al. [11] examined the growing problem of online deception, focusing on GPT-2-generated fake news. The researchers developed FAST, a graph-based method for analyzing literary compositions' factual architecture to distinguish human-generated from machine-generated material. They examined the news-style (GROVER) and web text-style (GPT-2) datasets for binary classification. We found that human-authored texts cited the same elements in their sentences, but machine-generated texts incorporated unnecessary sections, making differentiation easier.

2) Audio-based approach. The following research efforts aim to develop a deepfake detection model based on voice analysis. A study by Mcuba et al. [23] used deep learning techniques, especially convolutional neural networks (CNNs), to make it easier to spot deepfake sounds. They used the Baidu Silicon Valley AI Lab's VCTK and LibriSpeech speaker datasets for research. During preprocessing, they converted audio recordings into images and extracted features like mel-spectrograms and MFCCs for analysis. Hamza et al. [12] studied deepfake audio identification. The authors used advanced machine learning techniques, specifically MFCCs for audio evaluation, to improve detection approaches. To ensure high-quality preprocessing for model training, the authors used the Fake-or-Real (FoR) dataset with approximately 195,000 samples of real and synthetic speech. This study focused on feature extraction and classification models, with the Support Vector Machine (SVM) achieving the greatest accuracy of 98.83%. Shorter audio segments increased classification performance, with SVM outperforming in clean and loud circumstances. Pianese et al. [24] focused on speaker biometrics and trained the algorithm on real data to ensure real-world performance. Centroid-based (CB) and maximum similarity (MS) were the study's key approaches. The study tested advanced models such as RawNet2- antispoofing on ASVSpoof2019, FakeAVCelebV2, and In-The-Wild Audio Deepfake. Centroid-based (CB) testing compares test audio to the real audio average. Second, maximum-similarity testing (MS) compares test audio to real audio with the highest similarity score. The MS method performed well across datasets, especially in complicated, real-world circumstances. Muller et al. [25] examined text-to-speech synthesis advances and their effects on audio deepfakes, which may impersonate human voices. The authors used ASVspoof 2019 and IWA. They created 37.9 hours of audio samples from real and fictitious prominent people to test the models. ADAM optimizes cross-entropy loss-trained deep learning models.

TABLE I. ANALYTICAL REVIEW OF SINGLE-MODAL APPROACH

Study	Approach	Model	Datasets	Performance Metrics	Evaluation
Saravani et al. [21]	Text- based	BERT, BiLSTM, and NeXtVLAD	Deepfake	Accuracy	92%
Li et al. [22]	Text- based	Longformer	Customized (Texts produced by GPT-4)	Accuracy	86.54%
Zhong et al. [11]	Text- based	Graph Convolutional Network (GCN), LSTM, and Next Sentence Prediction (NSP)	GROVER and GPT-2	Accuracy	GROVER =87.97% GPT-2= 93.1%
Mcuba et al [23]	Audio-based	CNN with Adam, SDG, and Adadelta optimizers	VCTK with the LibriSpeech speakers	Accuracy	Adadelta =85.9% SDG = 83.6% Adam = 72.2%
Hamza et al. [12]	Audio-based	SVM	FoR	Accuracy	98.83%
Pianese et al. [24]	Audio-based	Supervised method, CB, and MS	ASVSpoof20 19, FakeAVCele bV2, and IWA	Calculate the average area under the curve (AUC) for all datasets	Supervised method = 79.6% CB = 88.4% MS= 91.3%
Muller et al. [25]	Audio-based	Multiple neural network	ASVspoof20 19 and IWA dataset	Equal Error Rate (EER)	ASVspoof2019 =9.85% IWA= 60.10%
Wodajo and Atnafu [5]	Visual-based	CViT (CNN and Vision Transformer (ViT))	DFDC	Accuracy	91.5%
Cozzolino et al. [10]	Visual-based	ID-Reveal which uses temporal ID network	VoxCeleb2 dataset	Accuracy	80.4%
Li et al. [26]	Visual-based	DSP-FWA by using CNNs	Own dataset (Celeb-DF) and others as UADFV and DFDC	AUC	Celeb-DF=64.6% UADFV=97.7% DFDC=75.5%

Comparing neural network topologies requires feature extraction, according to the authors. They found that raw audio features outperformed processed features and that longer audio segments reduced model error rates from 19.89% to 9.85%.

3) Visual-based approach. The studies below focus on developing a deepfake detection model based on image analysis. Wodajo and Atnafu [5] developed CViT to detect false content. The study required three steps for preparation: we recognized faces from video frames, scaled them to 224 x 224 pixels in RGB format for consistent input, and optimized content to reduce background interference. Rotation, flipping, and color changes boosted dataset variety and model generalization. The CViT used convolutional layers and transformer processes to record spatial hierarchy in images and apply self-attention to detect essential parts. The CViT had 91.5% accuracy on the enlarged dataset, although dataset heterogeneity and visual artifacts affected it. Cozzolino et al. [10] used the Temporal ID Network and 3D Morphable Model (3DMM) Generative Network to create the ID-Reveal method for detecting deepfake movies. The temporal ID network developed an embedded vector that identifies an individual using convolutional layers and similarity score computations, while the 3DMM generative network created authentic 3D characteristics. Test videos were compared to pristine reference videos for authenticity. The ADAM optimizer gave ID-Reveal 80.4% accuracy and 0.91 AUC for high-quality videos. Li et al. [26] proposed detection methods and the Celeb-DF dataset of high-quality deepfake videos of celebrities. They examined nine detection methods, including CNNs and head movement assessment. Deep learning-based spatio-temporal features for video forgery detection with attention (DSP-FWA) performed best. DSP-FWA looks at videos' spatial and temporal data using an attention method to bring out important parts and make detection more accurate.. Table I shows that deepfake detection techniques are either in text, audio, or visual domains. With each approach demonstrating unique strengths. Text-based models, Saravani et al. [21], used BERT, BiLSTM, and the NeXtVLADGCN, showed good accuracy in detecting manipulated texts, equal to 92%. While audio-based detection, Hamza et al. [12], used SVM, which outperformed other methods with an accuracy of 98.83%. On the other side, visual-based models, Wodajo and Atnafu [5] used the CViT and achieved good accuracy around 91%. Overall, selecting the right detection method depends on the type of deepfake being analyzed.

B. Multimodal Detection Approach

In multimodal detection, the detection is achieved by analyzing more than one type of manipulation. As we mentioned, there are two main combination types of methods for multimodal detecting deepfake: textual-visual or audio-visual detection. In this subsection, we will present the literature reviews that utilize multimodal deepfake detection. We divided this into subsections: text-visual based models and audio-visual-based models. This subsection concludes with an analytical table of the literature mentioned there.

1) Text-visual-based approach. Text-visual multimodal detection provides a framework for deepfake detection that integrates textual and visual analysis. Shao et al.[27] studied system localized modifications in coupled photos and text, with a specific focus on news stories. They have created the multimodal media manipulation dataset called Detecting and Grounding MultiModal Media Manipulation (DGM4). This dataset of 230,000 samples, including 152,574 modified pairs, examines the impact of simultaneous picture and text alterations. The detection model hierarchical multimodal manipulation reasoning transformer (HAMMER) and its enhanced version, HAMMER++, have proposed approaches to align and analyze image-text interactions for more accurate

modification detection. Liu et al.[4] developed the Unified Frequency-Assisted Trans Former Framework (UFAFormer) to discover and evaluate media modifications in images and text. It includes picture and text encoders, a frequency encoder that leverages DWT, and a Forgery-Aware Mutual Module (FAMM) to promote feature integration. UFAFormer improved accuracy on DGM4. Change recognition and classification improved as the authors devised components that incorporate multiple sources of information. Wang et al. [28] suggested a framework that used unimodal feature extractors to extract unique photo and text properties. The model used dual-branch cross-attention (DCA). It improved manipulation identification by analyzing both modalities simultaneously. The study focused on manipulation grounding, which involved identifying altered visual parts and text phrases. The data preparation included standardizing text to fifty tokens and scaling pictures to 256x256. Their method reached 95.11%.

2) Audio-visual-based approach. The following efforts provide a framework for deepfake detection that integrates audio and visual analysis. Khalid et al. [29] introduced the evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. They used the FakeAVCeleb dataset, which had 600 clips starring various celebrities. The authors employed three evaluative methodologies: unimodal, ensemble, and multimodal. Unimodal techniques examined audio or video separately, employing baseline models such as VGG16 and Xception; nevertheless, they encountered difficulties when one modality was authentic and the other was fabricated. Ensemble approaches amalgamate predictions from distinct classifiers for audio and video, enhancing accuracy. Nonetheless, the multimodal strategy, which concurrently assessed both data kinds, demonstrated diminished efficacy, indicating that existing techniques were insufficient for intricate deepfake identification. Lewis et al. [30] used multimodal deep learning to detect false videos utilizing spatial, spectral, and temporal anomalies. NOLANet uses multiple detection methods, including audio and visual attributes to classify content as real or false to identify deepfakes. They used Facebook's Deepfake Detection Challenge dataset (FDDC). Multi-stage detection used LSTM networks for categorization. Preprocessing the films involved extracting frames at 30 frames per second while maintaining the audio. They used BlazeFace to trim and recognize faces at 128x128 pixels. Face landmarks were extracted using FANet. Multiple LSTM networks processed integrated visual and auditory data after DCT analysis. Aligning aural and visual data helped discover deepfake discrepancies. Raza and Malik [31] proposed a multimodal representation learning multimodal trace for deepfake detection. The framework classified audio and video inputs as authentic or fake using a structured architecture. Essential components include audio and video feature extractors, mixers to combine them, and a classification layer. Data is prepared for analysis after normalization and feature extraction by Fast Fourier transform and ResNet. IntrAmodality Mixer Layers (IAML) evaluated audio and video inputs separately to discover patterns. Multi-label categorization using IEML unifies audio and video input modalities. Muppalla et al. [32] presented an integrated audio-visual feature for multimodal deepfake detection. That aimed to improve the identification of multimodal deepfakes by amalgamating audio and visual data via a distinct classification methodology. This entailed compressing video frames to a standardized dimension. Which is about 300x300 pixels and transforming audio into Mel spectrograms. The approach applied deep learning networks for feature extraction and implemented a multi-task learning technique. The assessment utilized two principal datasets. The FakeAVCeleb and the TMC datasets. They tested the framework using two models, the capsule network and the Swin transformer; the capsule network has the best accuracy. Cheng et al. [7] introduced a voice-face matching detection (VFD) system by looking at how well the voices and faces in deepfake videos match each other. They developed a method that first trains the model on regular data. Then fine-tune it for deepfake detection. The authors trained their model with the Voxceleb2 dataset, which includes over 17 million audio utterances from more than 6,000 celebrities. In the testing, they used many datasets, such as DFDC and FakeAVCeleb. Feng et al. [33] proposed self-supervised video forensics by audio-visual anomaly detection. The model was trained on real, unlabeled speech videos from the Lip-Reading Sentences 2 (LRS2) and Lip-Reading Sentences 3 (LRS3) datasets. By learning the normal patterns of audio-visual synchronization.

This led to the classification of anomalies by flagging videos that deviated from the established distribution of features representing the relationship between audio and visual elements. A key component of the method involved calculating synchronization scores for audio-video pairs using a specialized network. The model maximized synchronization scores for real pairs and identified fake ones by analyzing their probability distributions. The evaluation of the model was conducted using two datasets. The FakeAVCeleb and the large-scale Korean-language deepfake detection (KoDF) datasets. Oorloff et al. [34] proposed an audio-visual feature fusion (AVFF) framework for detecting video deepfakes. By employing a self-supervised learning strategy to capture the relationship between audio and visual elements. A trained classifier differentiated between real and fake videos by examining characteristics from both audio and visual elements. The training procedure used the LRS3 dataset for feature extraction. The FakeAVCeleb and KoDF datasets are used to evaluate the model. Table II shows multimodal deepfake detection methods which shows promising results by combining different inputs like text, audio, or visuals. Wang et al. [28] achieved the highest result for text-visual models at 91.42%. While Muppalla et al. [32] reported an impressive 99.20% multimodal result for audiovisual models. The ensemble based on the voting boosts detection robustness in Khalid et al. [29]. Overall, integrating multiple modalities enhances performance. However, effectiveness depends on the datasets, algorithms used, and the integration method used to merge two single models.

TABLE II. ANALYTICAL REVIEW OF MULTIMODAL APPROACH

Study	Approach	Model	Datasets	Performance metrics	Evaluation	
Shao et al. [27]	Text- visual based	Image detection by BBox Detector	DGM4	Intersection over Union mean (IoUmean)	Image = 76.51%	
		Text detection by Token Detector		Recall	Text= 72.14%	
		Multimodal by Multi-Label Classification		Mean Average Precision (mAP)	Multi-label =86.29%	
Liu et al. [4]	Text- visual based	Image encoder by vision transformer (ViT-B/16)	DGM4	Intersection over Union mean (IoUmean)	Image = 78.33%	
		Text encoder by BETR base model		Recall	Text= 70.61%	
		Multimodal by unified decoder to manipulate detection and grounding across image, text, and image-text pairs		Mean Average Precision (mAP)	Multi-label =87.85%	
Wang et al. [28]	Text- visual based	Image encoder by vision transformer (ViT-B/16)	DGM4	Intersection over Union mean (IoUmean)	Image = 80.83%	
		Text encoder by RoBERTa base model		Recall	Text= 70.73%	
		Binary classifier to determine whether manipulation is present. Fine-grained classifier to identify specific types of manipulation.		Mean Average Precision (mAP)	Multi-label =91.42%	
Khalid et al. [29]	Audio-visual based	Audio using many deep learning methods	FakeAV Celeb	Accuracy	Audio = 76%	
		Image using many methods such as VGG16			Image = 81%	
		Ensemble approach based on voting method. Multimodal using CNN			Ensemble = 78% Multimodal= 67.3%	
Lewis et al. [30]	Audio-visual based	Audio using LipSpeech sub-network	FDDC	Accuracy	Audio = 59.21%	
		Image using VSNet sub- network			Image = 61.59%	
		Multimodal using LSTM network detector			Multimodal= 65.18%	
Raza and Malik [31]	Audio-visual based	Audio IAML sub- network	FakeAVCeleb	Accuracy	Audio = NA	
		Image IAML sub- network			Image = NA	
		Multimodal using mixer layer with IEML			Multimodal= 92.9%	
Muppall a et al. [32]	Audio-visual based	Audio details not found	FakeAV Celeb and TMC	Accuracy of Capsule network	FakeAVCeleb	TMC
		Image details not found			Audio =99.80%	Audio =99.68%
		Multimodal using capsule network and the swin transformer			Image =61.59%	Image = 96.43%
					Multimodal= 65.18%	Multimodal= 99.20%
Cheng et al.[7]	Audio-visual based	VFD uses a dual-stream network (audio and visual)	DFDC and FakeAVCeleb	Accuracy	DFDC	FakeAV Celeb
		Multimodal using a matching function			Audio = NA	Audio = NA
					Image = NA	Image = NA
					Multimodal= 80.96%	Multimodal= 81.52%

Feng et al. [33]	Audio-visual based	Audio using VGG-M for encoding Multimodal using a matching function. Image using ResNet-18 for encoding Multimodal using autoregressive method	FakeAVCeleb and KoDF	AUC	FakeAVCeleb Audio = NA Image = NA Multimodal= 95.8%	KoDF Audio = NA Image = NA Multimodal= 86.9%
Oorloff et al. [34]	Audio-visual based	Unimodal encoders (audio, visual) Multimodal use: Cross-modal fusion: (audio-to-visual) and (visual-to-audio) Complementary masking while one modality is masked, the other is visible.	FakeAVCeleb and KoDF	AUC	FakeAVCeleb Audio = NA Image = NA Multimodal= 94%	KoDF Audio = NA Image = NA Multimodal= 95.5%

C. Emotion-Based Detection Approach

This subsection will focus on the related deepfake detection models that utilize emotional recognition features to detect fake content. Mittal et al. [19] proposed affective cue-based audio-visual deepfake detection. They found inconsistencies in actual and fake videos of the same person using a Siamese neural network and embedding vectors. Deepfakes are detected by evaluating emotional variations between speech and facial expressions. Their model was evaluated on DF-TIMIT and DFDC, two well-known deepfake detection datasets. Hosler et al. [20] suggested that using emotional inconsistencies to detect deepfakes is meaningful. Two dimensions were utilized to model emotion: valence (positive/ negative effect) and arousal. Real and fake videos have different emotional patterns, according to the authors. Deepfake technology struggled to recreate real emotions. First, video emotional features were extracted for detection. To express the subject's emotions across time, use valence and arousal. Finalizing the video's deepfake detection utilizing these emotional signals. The tests used two datasets. An emotion prediction model is trained on SEMAINE, and deepfake detection on DFDC. An LSTM model predicted emotional states and captured video emotional dynamics using extracted features. Conti et al. [35] presented a semantic approach for deepfake speech detection through emotion recognition (SER). The model identified emotional inconsistencies in synthetic audio by extracting emotional features and feeding them into a classifier. The system identified four primary types of emotions in speech, which are angry, happy, sad, and neutral. The system effectively distinguished reality from synthetic speech.

TABLE III. ANALYTICAL REVIEW OF EMOTIONAL APPROACH

Study	Approach	Model	Datasets	Performance Metrics	Evaluation	
Mittal et al. [19]	Audio-visual based	Emotions extraction Memory fusion network (MFN) to extract emotions from audio and visual modalities Detection by Siamese network to detect inconsistencies.	DF-TIMIT and DFDC	AUC	Without emotion indicator DFTIMIT= 94.8% DFDC=8 2.8%	With emotion indicator DFTIMIT= 96.3% DFDC= 84.4%
Hosler et al. [20]	Audio-visual based	Emotions extraction by Extracting low-level features for face and voice. Detection by Uses LSTM models to predict continuous emotional states (valence and arousal) over time.	DFDC	Accuracy	Without emotion indicator = 87.5%	With emotion indicator = 99.5%
Conti et al. [35]	Audio based	Emotions extraction by Speech Emotion Recognition (SER) component that extracts emotional features from speech using a 3D- Convolutional Recurrent Detection by Neural Network (CRNN) Synthetic Speech Detector (SSD) classifies the input as real or deepfake using a Random Forest classifier.	ASVspoof 2019	AUC	Without emotion indicator = NA	With emotion indicator = 98%

Testing showed high accuracy, especially in clean audio, and demonstrated better performance than other models. The core idea was that synthetic speech often failed to replicate authentic emotional patterns, making SER-based features valuable for detection. Table III shows that incorporating emotional recognition significantly enhances the performance of deepfake detection models. Mittal et al. [19] and Hosler et al. [20] proposed detection models using audio-visual deepfake detection. Demonstrated that including emotional features improved the performance. Hosler et al. [20] achieved impressive growth from 87% to around 99%. Conti et al. [35] also showed high performance with emotion extraction by speech emotion recognition. These findings suggest that emotion extraction can be a crucial factor in improving the performance of deepfake detection methods. However, there

are some considerations to enhance the performance, including selecting datasets, algorithms used, and merge methods.

The literature analysis presented above shows that several studies have been conducted to detect deepfake content with different types of content. As we mentioned, as deepfake technology advances, methods for detecting these intricate alterations must also improve. As we observed in Table II, multimodal detection in most cases has improved the performance of deepfake detection, while the model in Muppalla et al. [30] achieved an accuracy of around 99%, the poor detection accuracy of the visual-based model affected the overall multimodal detection accuracy when they used the FakeAVCeleb dataset. They reached an accuracy of around 99% on audio and 61% on image, leading to the overall detection of the multimodal dropping to around 66%. This situation necessitates further investigation into the methods of merging the individual models. Furthermore, notice in Table III that when emotion analysis is added to the model, it enhances the performance. Looking at Mittal et al. [19] and Hosler et al. [20] on how the emotion-based analysis increased performance results, however, there are limited studies that utilize this feature. In this study, we will aim to design audio-visual multimodal deepfake detection leveraging emotional recognition. The emotion extraction will be in three points: audio, image, and sentiment of speech. And the inconsistency between them. Considering the side effects of the multimodal dropped performance. That may occur due to the ineffective selection of the algorithm of audio detection and visual detection, as well as the integration methods.

III. Framework and Methodology

This section explores the architecture of our proposed framework, as shown in Fig. 5, which is designed to detect deepfake content by leveraging emotion recognition in multimodality.

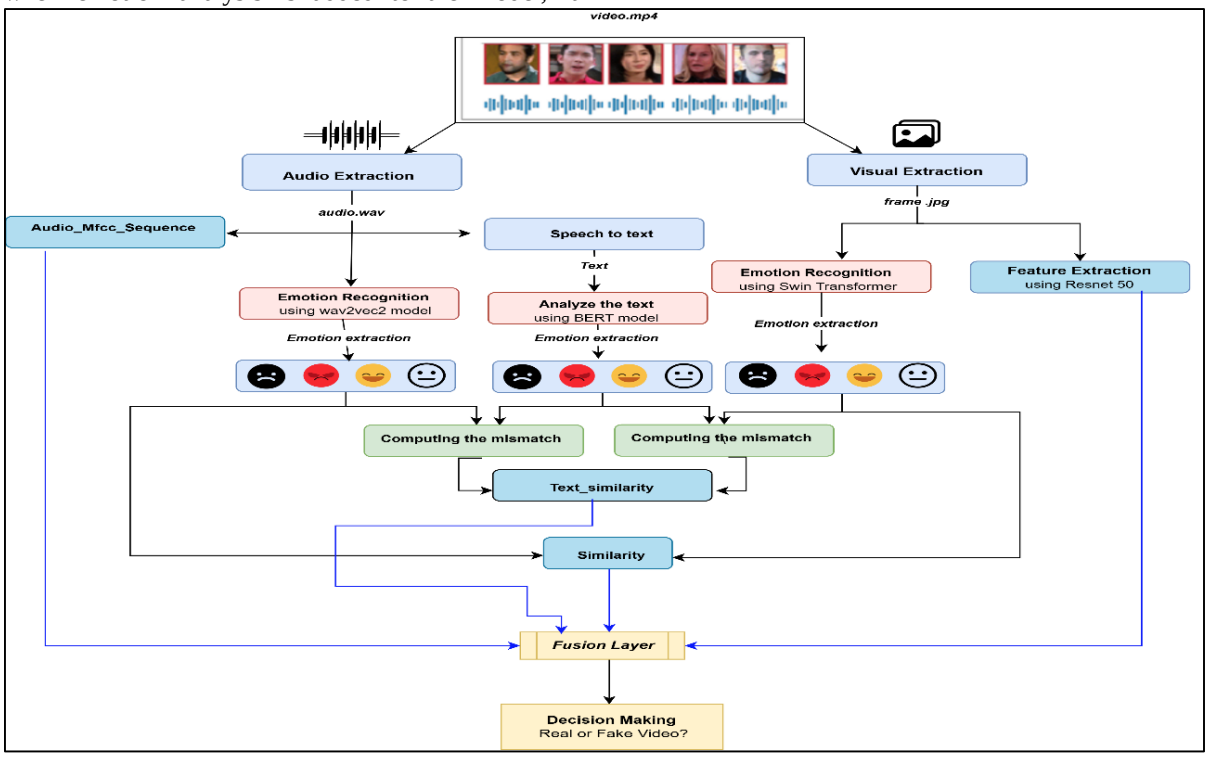

Fig. 5. The model framework.

A. Audio Features Extraction

The first component of our framework is the audio extraction and analysis, containing two main models. The first model aims to extract the main features of (.wav) audio, which includes thirteen Mel-Frequency Cepstral Coefficients (MFCC). The second model seeks to convert the speech to text, aka speech recognition, by using the pre-trained model called wav2vec2 [refer Appendix A] [36]. After that, the audio section involves the emotion extraction stage of each audio and text, aka the sentiment of speech. For the emotion recognition of the audio tones, we used the wav2vec2 pretrained model [37]. For sentiment analysis, we used a pretrained model using the bidirectional encoder representations from transformer (BERT) [38].

B. Visual Features Extraction

The second component is a visual analysis part that extracts the frames (.jpg) with image size 224*244 pixels and then extracts the visual features by a pretrained model using Residual Network - 50 layers (ResNet50) model to extract the noise, sharpness, and edge density. After that, the visual emotion extraction using the pretrained model used the shifted window transformer (Swin transformer) [39].

C. Cross-Model Emotional Consistency Stage

The consistency computation will be between the three modalities: text, audio, and visual. As follows, between the sentiment of speech and the emotion of the audio tone. Between the sentiment of speech and the emotion of the visual

part. Last, compute the mismatch or consistency between the emotion of the audio tone and the emotion of the visual part.

D. The Fusion Layer

Our framework's fusion technique is based on concatenates modality-specific characteristics from the text, audio, and visual streams before passing them through a classification. In particular, a Long Short-Term Memory (LSTM) network with a hidden size of 256 is used to process the temporal data from the video frames in order to identify dynamic patterns over time. After that, this output is coupled with emotion embeddings from each modality, visual quality metrics (noise, edge density, sharpness), and high-level features taken from the audio signal (MFCC and spectral contrast). The final classification output (real or false) is obtained by passing the resultant feature vector through two fully connected layers using ReLU activations.

This straightforward yet effective fusion method enables the model to integrate emotional signals and quality-related features from multiple modalities into a unified decision space. Although more sophisticated approaches—such as attention-based mechanisms—could further enhance performance, we intentionally opted for a lightweight architecture. Our goal was to ensure reproducibility and maintain interpretability, providing a clear and stable baseline for multimodal deepfake detection.

We implemented a modular pipeline to extract and align multimodal features from the FakeAVCeleb dataset. For each video, seven evenly spaced frames were extracted, alongside full audio and transcribed text. Visual features were obtained using a pretrained ResNet50 model, and temporal dependencies were captured using an LSTM module. Additional visual quality metrics (sharpness, edge density, and noise) were computed per frame. Audio features included MFCCs and spectral contrast, while emotional vectors for audio, text, and visual modalities were parsed from pre-extracted sentiment scores. Cosine similarity was used to measure the alignment between emotion vectors. A full architecture diagram is shown in Fig. 5.

This fusion strategy is theoretically grounded in the concept of cross-modal emotional coherence, which hypothesizes that authentic content exhibits natural alignment between facial expressions, vocal tone, and verbal sentiment. By explicitly modeling emotional alignment across modalities—rather than fusing only raw or semantic features—the framework captures inconsistencies that are typical in manipulated content. When compared to partially fused techniques, the ablation studies empirically show that integrating emotional consistency results in significant performance benefits. Our method employs emotional incompatibility as deception cues, providing a new layer of semantic analysis in fake content detection, in contrast to classic multimodal fusion, which only considers spatial or temporal correlations.

IV. Experiments and Results

In this section, we will explore our experiments and results and how we adopt a structured training approach that ensures stability and generalizability to train our deepfake detection model successfully.

A. Experiments

In this subsection, we will present the experiment setup, the dataset that has been used and the ablation study that was conducted in our experiments

1) Experimental setup. The model was trained for 10 epochs using the Adam optimizer with a learning rate of 1e-4. We applied cross-entropy loss and set the batch size to 4, with a fixed randomization seed of 101 in all experiments. The training dataset consisted of 80% of the total data, while the remaining 20% was used 10% for validation and 10% for testing.

2) Dataset. The dataset that has been used in our study is called the FakeAVCeleb dataset [9]. That contains 20,500 video samples, which are divided into the following categories: 500 real videos from the VoxCeleb2 dataset, featuring real audio. 500 videos that combine synthetic (fake) audio with actual footage. 9,000 videos with real audio and fake video. 10,000 videos containing fake audio and video. The average length of each video is roughly 7.8 seconds. The dataset offers a variety of actual and altered audiovisual material combinations to facilitate multimodal deepfake detection research. Fig. 6 below shows the categories of the FakeAVCeleb dataset. The original FakeAVCeleb dataset contains 20,500 samples. For this study, we selected a representative and balanced subset consisting of 998 video samples, equally divided into 499 real and 499 fake videos. The samples were randomly selected from the full dataset while ensuring class balance. To streamline the training and evaluation process, we manually created a CSV file that includes the file paths and corresponding labels for each video sample. This setup allows for efficient data loading and consistent processing across experiments.

Fig. 6. The FakeAVCeleb dataset [9].

3) Ablation study. To ensure the productivity of our contribution, we conducted the ablation study for each emotion feature in our multimodal deepfake detection model as follows:

a) Audio-visual emotion only (No sentiment): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise,

sharpness, and edge density. Furthermore, the emotion consistency computation between audio and visual modalities. Without any sentiment information. Table IV shows the detailed results of the experiment.

TABLE IV. AUDIO-VISUAL WITH EMOTION RECOGNITION RESULT

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with emotion recognition	76.19%	72.15%	73.3%

b) Audio emotion and text sentiment (No visual Emotion): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise, sharpness, and edge density. Furthermore, the emotion consistency computation is between audio and text modalities without any computation between audio and visual emotion. Table V shows the detailed results of the experiment.

TABLE V. AUDIO-VISUAL WITH AUDIO EMOTION AND TEXT SENTIMENT

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with audio emotion and text sentiment	76.19%	69.21%	68.33%

c) Visual emotion and text sentiment (No audio emotion): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise, sharpness, and edge density. Furthermore, the emotion consistency computation is between visual and text modalities without any computation between audio and visual emotion. Table VI shows the detailed results of the experiment.

TABLE VI. AUDIO-VISUAL WITH VISUAL EMOTION AND TEXT SENTIMENT

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with visual emotion and text sentiment	76.19%	69.21%	68.33%

d) Sentiment fusion only (No cross-audio-visual emotion): In this experiment, we focused on the features extracted from audio, which are thirteen MFCC, and from visual, which are noise, sharpness, and edge density. Furthermore, the emotion consistency computation between visual and text modalities and between audio and text modalities without any computation between audio and visual emotion. Table VII shows the detailed result of the experiment.

e) Full multimodal emotion and sentiment fusion (Our approach): In this experiment, it was conducted on the complete framework features, which are the features extracted from audio, which are thirteen MFCC, and from visual which are noise, sharpness, and edge density, the emotion consistency computation between visual and text modalities, and between audio and text modalities. Furthermore, the computation between audio and visual emotion. Table VIII shows the detailed results of the experiment.

TABLE VII. AUDIO- VISUAL WITH SENTIMENT ANALYSIS

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Audio-visual with sentiment analysis	80.95%	78.57%	81.67%

TABLE VIII. OUR APPROACH

Model	Performance		
	Accuracy	*F1-Score*	*Recall*
Our Approach	95.24 %	95.24%	95.45%

These results empirically validate the contribution of sentiment indicators in multimodal fusion, showing that their absence significantly degrades performance. This supports our hypothesis that emotional coherence between modalities is a strong signal for deepfake detection. These ablation results are further discussed and interpreted in Section V, where we reflect on their implications and underlying patterns.

B. Results

In this subsection, we present the performance evaluation of our proposed model through three comparative perspectives. First, we compare our results with existing studies that utilized the same dataset (FakeAVCeleb) to provide a fair benchmark under similar data conditions. Second, we examine studies that incorporated emotion-related features, particularly those focusing on emotional or affective cues from audio or visual streams, to highlight the novelty of integrating the sentiment of speech into the detection process. Finally, we compare our model against state-of-the-art baseline methods, including both visual-only and multimodal approaches such as FaceForensics++[40] and LipForensics[41]. These comparisons aim to position our work within the broader deepfake detection landscape and demonstrate the effectiveness of our multimodal sentiment-aware framework.

1) Comparative results with the related used same dataset. In this part, we will explore the related works that use the FakeAVCeleb dataset to detect the deepfake content. Table IX shows the comparison details.

2) Comparative results with emotion-based approach. In this part, we will present the related works that focused on emotion recognition. Table X below shows the details of the comparison.

Although the model results were not the highest compared to previous studies, it opened a new area of research in how the sentiment of speech can affect the detection of deepfakes if considered along with the emotions of images and audio. Thus, when all the experimental strategies were fixed and the ablation study was conducted, it was found that the sentiment of speech indicator increased the productivity of the model in detecting deepfakes. The model accuracy without sentiment of speech reached 76%, however, when adding the sentiments, the accuracy increased to 95.24%.

3) Comparative results with state-of-the-art baseline models. We also compare our model with well-known baseline methods like FaceForensics++[40] and LipForensics [41].

Including both visual-only and multimodal approaches. This helps to place our work in context and shows how our use of audio, video, and sentiment features adds value to deepfake detection. Table XI shows the details of the comparison.

As shown in Table XI, FaceForensics++ [40] is a visual-only approach that uses XceptionNet and achieved 95.7% AUC on the FF++ dataset. LipForensics [41], on the other hand, detect a deepfake based on inconsistencies in lip movements and performs even better, reaching 97.1% AUC. While both are strong and widely used methods, they focus only on the visual aspect of the content and don't take audio or emotional signals into account. This is where our approach brings something new, by combining audio, video, and sentiment features to better capture manipulations that might be subtle or spread across different modalities. On the other hand, our approach looks at the problem a bit differently. Instead of relying on just the visual side, we combine both audio and video inputs — and go a step further by including the sentiment behind the speech. This gives the model a better chance at catching those subtle manipulations that might slip past visual-only methods. When we tested it on a balanced sample from the FakeAVCeleb dataset, the model reached 95.24% accuracy, which shows how adding emotional cues from speech can really make a difference in spotting deepfakes, especially when the fakeness is hard to notice.

TABLE IX. COMPARING WITH RELATED WORKS ON THE SAME DATASET

Study	Deepfake detection	Emotion recognition	Sentiment of speech	Result
Pianese et al. [24]	Audio-based	No	No	91%
Khalid et al. [29]	Audio-visual based	No	No	81%
Raza and Malik [31]	Audio-visual based	No	No	92.9%
Muppalla et al. [32]	Audio-visual based	No	No	65.18%
Cheng et al. [7]	Audio-visual based	No	No	81.52%
Feng et al. [33]	Audio-visual based	No	No	95.8%
Oorloff et al. [34]	Audio-visual based	No	No	94%
Our study	**Audio-visual based**	**YES**	**YES**	**95.24%**

TABLE X. COMPARING WITH RELATED USED EMOTION RECOGNITION

Study	Deepfake detection	Dataset	Emotions based	Sentiment of speech	Result
Mittal et al. [19]	Audio-visual	DF-TIMIT	Yes	NO	94.8
Hosler et al. [20]	Audio-visual	DFDC	Yes	NO	99.5%
Conti et al. [35]	Audio	ASVspoof2019	Yes	NO	98%
Our study	**Audio-visual**	**FakeAVCeleb**	**Yes**	**Yes**	**95.24%**

TABLE XI. COMPARING WITH BASELINE MODELS

Study	Deepfake detection	Dataset	Result
FaceForensics [40]	Visual	FaceForensics++	AUC = 95.7%
LipForensics[41]	Visual (Lip-based)	FaceForensics++/ DFDC	AUC= 97.1%
Our study	**Audio, Visual, Sentiment**	**FakeAVCeleb**	**Accuracy= 95.24%**

V. DISCUSSION

As mentioned before, the proposed framework depends on the extraction and analysis of emotions from visual, textual, and audio modalities. BERT is used for speech sentiment, Swin Transformer is used for visual emotion, and wav2vec2 is used for audio tones. An LSTM-based fusion layer is then utilized to fuse the consistency calculated by comparing these emotional cues across modalities. To ensure the accuracy of the results and emotional studies separately, we conducted the ablation study. Table XII shows the results from the ablation study.

TABLE XII. RESULTS FROM THE ABLATION STUDY

Experiment	Accuracy	F1-score	Recall
Audio-visual with emotion recognition	76.19%	72.15%	73.3%
Audio- visual with audio emotion and text sentiment	76.19%	69.21%	68.33%
Audio- visual with visual emotion and text sentiment	76.19%	69.21%	68.33%
Audio- visual with sentiment analysis	80.95%	78.57%	81.67%
Our Approach	95.24 %	95.24%	95.45%

As we noted above, we conducted a study that focused on emotions in five experiments. The first experiment contained models of audio, image features, and emotional mismatch between audio and image, which is what came in previous studies. It's got modest results in terms of accuracy, F1-score, and recall. Moreover, in the second and third experiments, we observe that both F1-score and recall decreased compared to the first experiment. This indicates that partial modality fusion is insufficient to enhance the distinction between real and fake content. This justifies how the fourth experiment improved performance significantly, as we removed the partial modality fusion. Now the model measures the mismatch from three aspects: first, between the text and the audio, let's call it A. Second, between the text and the image, let's call it B, and third, between A and B. In the last experiment, which is our approach, we achieved the highest results, with a consistency across all metrics, which are accuracy, F1-score, and recall. Fig. 7 below shows the consistency across all metrics between all experiments. The final experiment demonstrates a strong balance across all evaluation metrics, with accuracy (95.24%), F1-score (95.24%), and recall (95.45%) closely aligned. This consistency indicates that the model maintains a reliable tradeoff between accuracy and recall. In contrast, earlier experiments show a notable gap between accuracy and other metrics. That suggests that partial fusion strategies (emotion only, sentiment-audio only, or sentiment-visual only) may lead to biased predictions or limited generalization. The balanced results in our approach highlight their robustness and reliability in detecting deepfakes across various modalities.

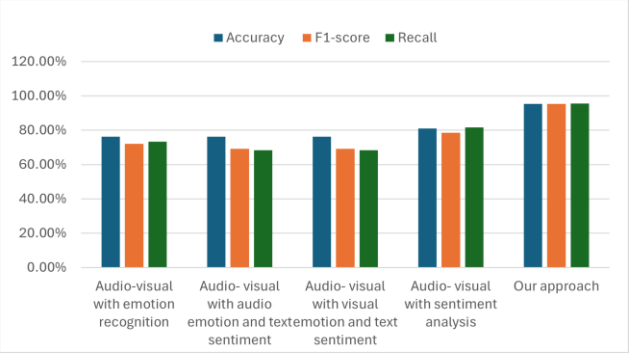

Fig. 7. Evaluation metrics of experiments.

We also explore additional enhancements which is not adding significant improvements. Table XIII shows the detailed results.

TABLE XIII. ENHANCEMENTS

Experiment	Accuracy	F1-score	Recall
The approach with gating mechanism	90.48%	90.45%	90.45%
The approach with the FGSM	95.24 %	95.24%	95.45%
The approach without enhancement	95.24 %	95.24%	95.45%

Finally, using the FakeAVCeleb dataset, this study examines how speech sentiment and emotion recognition affect detection accuracy. When compared to previous studies utilizing the same dataset, our model performed effectively, achieving an accuracy of 95.24%. Opening a new viewpoint of the effects of sentiment of speech on deepfake detection productively. To the best of our knowledge, no prior research using this dataset has examined this emotional aspect. Furthermore, in the domain of deepfake detection, no previous study has considered both speech sentiment and emotion recognition simultaneously. Our research presents this combination on the FakeAVCeleb dataset that provides a novel and useful viewpoint for developing multimodal deepfake analysis with the sentiment of speech.

VI. CONCLUSION

People's reliance on the Internet and social media has increased significantly in this era. With the development of artificial intelligence technologies, data manipulation has become accessible to everyone, and there are even free manipulation tools. In this study, we explored the framework for detecting deepfakes in video by using emotion recognition in three modalities: audio, images, and the sentiment of transcribed speech.

Our research contributed to this combination of emotions, achieved a high result, which is 95.24%, and provided a novel and useful viewpoint for developing multimodal deepfake analysis with the sentiment of speech. However, this study has certain limitations. The experiments were conducted solely on the FakeAVCeleb dataset, which may limit the generalizability of the results. This choice was made due to the dataset's unique structure, which provides audio and visual components—making it highly suitable for emotion- and sentiment-based fusion analysis. Moreover, while we relied on pre-trained models such as wav2vec2, BERT, and Swin Transformer for feature extraction, this decision was made deliberately to isolate and focus on the core contribution of this work: emotional consistency fusion across modalities. In future work, we aim to build our fine-tuned domain-specific models tailored for deepfake detection tasks, allowing for deeper integration of modality-aware learning and potentially improving detection performance and adaptability across datasets. Furthermore, we will conduct a detailed failure case analysis to address issues like speaker mismatches and low-quality audio. We also plan to integrate explainability methods such as SHAP and Grad-CAM to better interpret model decisions and enhance transparency.

REFERENCES

[1] Tableau [n.d.], 'What are the advantages and disadvantages of artificial intelligence?', https: //www.tableau.com/data-insights/ai/advantages-disadvantages. Accessed: November 29, 2024.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Akhtar, Z., Pendyala, T. L. and Athmakuri, V. S. [2024], 'Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve', Forensic Sciences 4(3), 289–377.

[3] Spector, N. [2017], 'So it's fine if you edit your selfies, right? think again', https://www.nbcnews.com/business/consumer/ so-it-s-fine-if-you-edit-your-selfies-not-n766186.

[4] Liu, P., Tao, Q. and Zhou, J. T. [2024], 'Evolving from single-modal to multi-modal facial deepfake detection: A survey', arXiv preprint arXiv:2406.06965 .

[5] Wodajo, D. and Atnafu, S. [2021], 'Deepfake video detection using convolutional vision transformer', arXiv preprint arXiv:2102.11126.

[6] Limer, E. [2021], 'Bank robbers in the middle east reportedly cloned someone's voice to steal $35 million', https://gizmodo.com/ bank-

robbers-in-the-middle-east-reportedly-cloned-someo-1847863805. Accessed: 2024-11-24.

[7] Shah, S. [n.d.], 'Sih31st [kaggle notebook]', https://www.kaggle.com/code/sohamshah03/ sih31st/notebook. Retrieved:2024-11-12.

[8] Awsaf49 [n.d.], 'Asvspoof 2019 tfrecord dataset [data set]', https://www.kaggle.com/ datasets/awsaf49/asvspoof-2019-tfrecord-dataset. Retrieved: 2024-11-12.

[9] DASH Lab, FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. [Online]. Available: https://sites.google.com/view/fakeavcelebdash-lab/. [Accessed: Apr. 28, 2025].

[10] Cozzolino, D., Rössler, A., Thies, J., Nießner, M. and Verdoliva, L. [2021], Id-reveal: Identityaware deepfake video detection, in 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 15108–15117.

[11] Zhong, W., Tang, D., Xu, Z., Wang, R., Duan, N., Zhou, M. and Yin, J. [2020], 'Neural deepfake detection with factual structure of text', arXiv preprint arXiv:2010.07475 .

[12] Hamza, A., Javed, A. R. R., Iqbal, F., Kryvinska, N., Almadhor, A. S., Jalil, Z. and Borghol, R. [2022], 'Deepfake audio detection via mfcc features using machine learning', IEEE Access 10, 134018–134028.

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. [2020], 'Generative adversarial networks', Communications of the ACM 63(11), 139–144.

[14] Chesney, R. and Citron, D. [2019], 'Deepfakes and the new disinformation war: The coming age of post-truth geopolitics', Foreign Affairs 98, 147.

[15] Maras, M.-H. and Alexandrou, A. [2019], 'Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos', The International Journal of Evidence & Proof 23(3), 255–262.

[16] Becker, C., Conduit, R., Chouinard, P. A. and Laycock, R. [2024], 'Can deepfakes be used to study emotion perception? a comparison of dynamic face stimuli', Behavior Research Methods pp. 1–17.

[17] VISO.ai [n.d.], 'Visual emotion recognition with ai', https://viso.ai/deep-learning/ visual-emotion-ai-recognition/. Accessed: 2024-11-24.

[18] Abdallah, A., Al-Sharif, M. H. and Al-Zoubi, A. [2022], 'A systematic review of emotion recognition from video: Opportunities and challenges', Computational Intelligence and Neuroscience. https://doi.org/10.1155/2022/2645381.

[19] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. and Manocha, D. [2020], 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.

[20] Hosler, B., Salvi, D., Murray, A., Antonacci, F., Bestagini, P., Tubaro, S. and Stamm, M. C. [2021], Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotionalinconsistencies, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 1013–1022.

[21] Saravani, S. M., Ray, I. and Ray, I. [2021], Automated identification of social media bots using deepfake text detection, in 'International Conference on Information Systems Security', pp. 111–123.

[22] Li, Y., Li, Q., Cui, L., Bi, W., Wang, Z., Wang, L. and Zhang, Y. [2024], 'Mage: Machinegenerated text detection in the wild', Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics 1, 36–53.

[23] Mcuba, M., Singh, A., Ikuesan, R. A. and Venter, H. [2023], 'The effect of deep learning methods on deepfake audio detection for digital investigation', Procedia Computer Science 219, 211–219.

[24] Pianese, A., Cozzolino, D., Poggi, G. and Verdoliva, L. [2022], Deepfake audio detection by speaker verification, in '2022 IEEE International Workshop on Information Forensics and Security (WIFS)', IEEE, pp. 1–6.

[25] Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A. and Böttinger, K. [2022], 'Does audio deepfake detection generalize?', arXiv preprint arXiv:2203.16263 .

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

[27] Shao, R., Wu, T., Wu, J., Nie, L. and Liu, Z. [2023], Detecting and grounding multi-modal media manipulation, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'. Accessed: 2024-12-18.

[28] Wang, J., Liu, B., Miao, C., Zhao, Z., Zhuang, W., Chu, Q. and Yu, N. [2024], Exploiting modality-specific features for multi-modal manipulation detection and grounding, in 'ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 4935–4939.

[29] Khalid, H., Kim, M., Tariq, S. and Woo, S. S. [2021], 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.

[30] Lewis, J. K., Toubal, I. E., Chen, H., Sandesera, V., Lomnitz, M., Hampel-Arias, Z. and Palaniappan, K. [2020], Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning, in '2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)', IEEE, pp. 1–9.

[31] Raza, M. A. and Malik, K. M. [2023], Multimodaltrace: Deepfake detection using audiovisual representation learning, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 993–1000.

[32] Muppalla, S., Jia, S. and Lyu, S. [2023], Integrating audio-visual features for multimodal deepfake detection, in '2023 IEEE MIT Undergraduate Research Technology Conference (URTC)', pp. 1–5.

[33] Feng, C., Chen, Z. and Owens, A. [2023], 'Self-supervised video forensics by audio-visual anomaly detection', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 10491–10503.

[34] Oorloff, T., Koppisetti, S., Bonettini, N., Solanki, D., Colman, B., Yacoob, Y. and Bharaj, G. [2024], Avff: Audio-visual feature fusion for video deepfake detection, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 27102–27112.

[35] Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F. and Tubaro, S. [2022], Deepfake speech detection through emotion recognition: a semantic approach, in 'ICASSP 2022- 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 8962–8966.

[36] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

[37] Yang, S., Chi, P., Huang, Y., Lim, C., Lee, H., Sung, D. L., Lin, J., & Yi, H. (2021). SUPERB: Speech processing universal PERformance benchmark. Interspeech 2021.

[38] Kirouane, A. (2022). BERT-Emotions-Classifier [Computer software]. Hugging Face. https://huggingface.co/ayoubkirouane/BERT-Emotions-Classifier

[39] Segni, M. W. (2023). swin-tiny-patch4-window7-224-finetuned-face-emotion-v12 [Computer software]. Hugging Face. https://huggingface.co/MahmoudWSegni/swin-tiny-patch4-window7-224-finetuned-face-emotion-v12

[40] 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).

[41] Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). 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 (pp. 5039-5049).

APPENDIX A: GLOSSARY OF TERMS AND ACRONYMS

- Sentiment-Emotion Conflict: This is a discrepancy between a person's tone of voice or facial expression and what they say, or the sentiment or meaning of their words.
- MFCC (Mel-Frequency Cepstral Coefficients): These are characteristics extracted from audio signals that replicate the way sound is perceived by the human ear. They aid in capturing the speaker's voice's emotional cues in our investigation.
- SwinT (Swin Transformer): A smart image-processing model that breaks an image into windows and shifts them to capture visual details more accurately. We use it to analyze facial emotions frame by frame.
- wav2vec2: A speech model that picks up knowledge straight from unprocessed audio. In our situation, it is helpful to identify the sentiment expressed as well as the words themselves.
- BERT: A language model that understands context in text. We use it to analyze the emotional meaning of the transcribed speech.
- FGSM (Fast Gradient Sign Method): A method for determining if the model can still produce accurate predictions by significantly altering the input data. It enables us to assess the robustness of the model.
- Gating Mechanism: When merging several data types, a clever layer that learns to "turn on" or "off" specific properties helps the model concentrate on what really matters.
- Emotion-Sentiment Fusion: This is fusing a person's tone of voice or facial expression with their words to convey their feelings. This combination can highlight minute indications of video tampering.