Speech Emotion Recognition in Multimodal Environments with Transformer: Arabic and English Audio Datasets

Esraa A. Mohamed[1], Abdelrahim Koura[2], Mohammed Kayed[3]
Faculty of Science, Beni-Suef University, Beni-Suef City, Egypt[1]
Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, Egypt[2, 3]

***Abstract*—Speech Emotion Recognition (SER) is a fast-developing area of study with a primary goal of automatically identifying and analyzing the emotional states expressed in speech. Emotions are crucial in human communication as they impact the effectiveness and meaning of linguistic expressions. SER aims to create computational approaches and models to detect and interpret emotions from speech signals. One of the primary applications of SER is evident in the field of Human-Computer Interaction (HCI), where it can be used to develop interactive systems that adapt to the user's emotional state based on their voice. This paper investigates the use of speech data for speech emotion recognition. Additionally, we applied a transformation process to convert the speech data into 2D images. Subsequently, we compared the outcomes of this transformation with the original speech data, aligning the comparison with a dataset containing labeled speech samples in both Arabic and English. Our experiments compare three methods: a transformer-based model, a Vision Transformer (ViT) based model, and a wave2vec-based model. The transformer model is trained from scratch on two significant audio datasets: the Arabic Natural Audio Dataset (ANAD) and the Toronto Emotional Speech Set (TESS), while the vision transformer is evaluated alongside wave2vec as part of transfer learning. The results are impressive. The transformer model achieved remarkable accuracies of 94% and 99% on ANAD and TESS datasets, respectively. Additionally, ViT demonstrates strong capabilities, achieving accuracies of 88% and 98% on the ANAD and TESS datasets, respectively. To assess the transfer learning potential, we also explore the Wave2Vector model with fine-tuning. However, the findings suggest limited success, achieving only a 56% accuracy rate on the ANAD dataset.**

***Keywords*—Speech emotion recognition; transformer encoder; fine-tuning; wav2vec; multimodal emotion recognition**

I. Introduction

Emotions, found across all cultures, play a vital role in interpersonal communication. Research on emotional recognition has evolved since the 1970s, spanning various modalities such as speech, text, video, EEG brain waves, and facial expressions. Additionally, multi-modal approaches combining text and audio data have gained prominence in speech emotion recognition. The objective is to automatically discern an individual's emotional or physical state from their voice. Understanding the speaker's emotional state can aid listeners in deciphering the true intent behind spoken words.

In the current COVID-19 pandemic, where social distancing is crucial, tele-diagnosis or telephone consultation has gained significant prominence. Integrating speech emotion recognition (SER) systems into these applications can have a profound impact on various fields. For example, in telemedicine, SER can play a crucial role in remotely diagnosing patient's condition by analyzing their emotional cues during the conversation. Furthermore, the integration of emotion detection features into speech recognition software can help bridge communication barriers faced by individuals with hearing impairments. Emotions also play a vital role in decision-making and greatly influence the naturalness of human-machine interactions. In the automotive industry, incorporating emotion recognition systems into onboard car systems can help drivers stay alert and prevent accidents caused by stress or fatigue. Additionally, analyzing call center conversations using SER can improve the overall quality of customer service. Moreover, applications such as interactive films, storytelling, and online instruction can benefit from emotion recognition technology to enhance user engagement and overall experience. The wide-ranging applications of speech emotion recognition highlight its potential to revolutionize communication, human-machine interaction, and safety across various domains the recognition and analysis of emotions from speech pose several challenges due to emotions' complex and subjective nature. Unlike visual cues, which can be readily observed and interpreted, emotions conveyed through speech rely on acoustic, prosodic, and linguistic patterns that require sophisticated computational models for accurate recognition. Additionally, the inherent variability in emotional expression across individuals, cultures, and languages further complicates the task of SEA. Over the years, researchers have explored various methodologies for SER, including traditional machine learning algorithms such as support vector machines, Gaussian mixture models, and hidden Markov models. These approaches often rely on handcrafted acoustic and prosodic features to capture relevant information from speech signals. However, they may struggle to capture the intricate nuances and complex emotional patterns.

Recent advancements in deep learning have revolutionized the field of SER by enabling the development of more powerful and flexible models. Deep learning techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers, have demonstrated

remarkable success in various natural language processing tasks. They have the potential to capture high-level representations and learn complex relationships from raw speech data, enabling more accurate and robust emotion recognition.

Transformers are a subset of these deep learning models that have drawn much interest because of their potent ability to capture contextual information and long-range dependencies adequately. Transformers, first developed for applications involving natural language processing, have demonstrated promise in jobs requiring sequential data, such as speech recognition and linguistic translation. The transformer design, based on self-attentional processes and multi-head attention, enables the modelling of links between various components of the speech signal and the capture of global dependencies.

Emotion recognition from a speech is a vital field that employs machine learning to automatically detect and interpret emotional states expressed in spoken language or speech signals. Applications for it can be found in many different fields, including social robots, affective computing, and human-computer interaction. This paper introduces an innovative transformer-based approach to SER, showcasing a detailed analysis of the method's components. This includes the preprocessing of speech data, the feature extraction techniques, and the design and training of the transformer models. The study goes further to evaluate the approach's performance using two audio datasets, comparing it against the vision transformer and a transfer learning model such as wave2vec. The results, benchmarked against existing state-of-the-art methods, underscore the significance of this work in advancing our ability to understand and respond to human emotions conveyed through speech.

The paper is structured as follows. In Section II, we discuss related work in the field. Section III presents our approach, which includes the construction of the transformer. Section IV analyzes and delineates the preprocessing procedures employed in our study, as well as the resulting outcomes of the experiment. Finally, Section V concludes with recommendations for further development.

II. Related Works

Numerous studies have been dedicated to speech emotion recognition, a field of growing prominence. Researchers in this domain employ diverse machine learning algorithms and feature extraction techniques to create robust models for automated emotion recognition from spoken language. This technology finds applications in sentiment analysis, virtual assistants, and affective computing. This section provides a comprehensive literature review, showcasing various approaches to enhance emotion recognition. Traditional classification techniques utilizing distinctive feature vectors form the bedrock of many methodologies. Noteworthy studies combine Support Vector Machine (SVM) classification with fused features such as F0, Energy, and Mel-frequency cepstral coefficients (MFCCs) [1]. Additionally, features like MFCCs and Mel Energy Discrete Cosine Coefficients (MEDC) are adeptly harnessed for emotion classification using SVM [2]. Another approach integrates diverse features from the Berlin emotional database, employing SVM for emotional state classification [3]. Further exploration reveals Gaussian mixture models (GMMs) applied for emotion classification [4]. Innovatively, a hybrid model employing a GMM-based low-level feature extractor and a neural network high-level feature extractor excels in recognizing speaker emotions [5]. Discrete hidden Markov models (HMMs) emerge as a robust classifier, capitalizing on short time log frequency power coefficients (LFPC) to represent speech signals [6]. Spectral features like MFCCs and Mel spectrograms, along with classifiers like Support Vector Machines (SVMs), Multilayer Perceptrons (MLPs), and K-Nearest Neighbors (KNN), further enrich the array of methodologies [7].

Deep learning techniques have proliferated in the realm of speech emotion recognition (SER), offering a plethora of advantages over traditional methods. These approaches boast automated detection of intricate structures and features, eliminating the necessity for manual feature extraction and tuning. They excel in deriving low-level features directly from raw data and adeptly utilize techniques like recurrent neural networks (RNNs) to navigate unlabeled data. A compelling illustration lies in study [8], which deployed a deep recurrent neural network for speech-based emotion recognition. Similarly, in study [9] introduced a pioneering method involving Directional Self-Attention in Bi-directional Long-Short Term Memory (BLSTM-DSA). This journey delves further with [10], presenting a groundbreaking approach termed the Deep Convolutional Neural Network (DCNN) combined with a Bidirectional Long Short-Term Memory with Attention (BLSTMwA) model. Convolutional Neural Networks (CNNs) manifest in other studies [11][12][13], while [14]seamlessly blends RNNs with CNNs. A striking instance is [15] , wherein a Taylor series-based Deep Belief Network (Taylor-DBN) takes center stage. Similarly, [16] harnesses both a 1D CNN LSTM network and a 2D CNN LSTM network. Further exploration leads us to [17], which delves into the potential of the Multi-Layer Perceptron (MLP) deep network architecture.

The paradigm shift arrives with the emergence of deep learning models, particularly those embodying transformer architectures, triggering a revolution in emotion recognition from speech. By tapping into the prowess of self-attention mechanisms and multi-head attention, these transformer-based models adeptly capture long-range dependencies and intricate speech patterns, elevating emotion recognition accuracy. Their forte lies in deciphering contextual relationships within input sequences, empowering them to apprehend nuanced emotional cues and speech pattern variations. Furthermore, the capacity of transformer-based models to accommodate substantial data volumes and exploit parallel processing has solidified their standing as a preferred choice for emotion recognition tasks. Their fusion with deep learning techniques introduces tantalizing prospects for enriching emotion analysis and comprehension across diverse domains. Ultimately, they emerge as an invaluable asset, resonating profoundly with both researchers and practitioners, poised to reshape the landscape of emotion recognition.

Given the transformer's remarkable aptitude for sequence learning tasks, particularly in the realm of natural language

processing, an enhanced transformer-inspired model is developed, finely tailored for the nuances of speech emotion recognition tasks. An innovative deep multimodal transformer network, introduced by study [18], deftly addresses the challenge of asynchronous emotion expressions across multiple modalities. This novel architecture adeptly captures distinctive temporal features and orchestrates emotional evolution over sequences of utterances. This dynamic is further amplified by weight sharing and the fusion of emotional content from audio and text components. The transformative impact continues with the infusion of the Taylor linear attention (TLA) algorithm [19], seamlessly integrated into the transformer architecture by [20]. In a similar vein, [21] introduces the LSTM-Transformer model, ingeniously replacing positional encoding within the Transformer framework with LSTM recurrent processes. This adaptive strategy learns the concealed input feature states and enhances the model's capacity to discern emotion nuances. An inventive deep multimodal transformer network surfaces in [22], laser-focused on unraveling unique temporal features while adroitly managing the asynchronous nature of emotion expression across modalities. The aim is to adeptly model the progression of emotion across the timelines of utterances. Meanwhile, [23] wields advanced Transformers and attention-based fusion mechanisms to fuse the hallmarks of multimodal self-supervised learning, triumphing in the realm of multimodal emotion identification challenges. Embarking on a journey to harness the self-attention prowess and global windowing potential of the transformer model for SER, [24] deftly explores their utility. On a parallel track, [25] forges a groundbreaking frontier by presenting an automatic emotion recognition system (FER) that seamlessly integrates both speech and visual emotion recognizers within a unified framework. To assess SER performance, [26] rigorously examines two transfer-learning strategies. They adroitly employ a pre-trained xlsr-Wav2Vec2.0 transformer for embedding extraction and fine-tuning. Pioneering a new learning paradigm for SER, [27]employs Compact Convolutional Transformers (CCTs) synergized with speaker embeddings. The result is a commendable achievement of real-time results across diverse corpus scenarios.[28] delves into the realm of facial emotion recognition (FER), wielding the ResNet-18 model in conjunction with transformers. The result is superior performance and practical applicability in real-world settings, surpassing existing models on hybrid datasets. Wrapping the discourse is the innovative transfer learning methodology for speech emotion recognition put forth by [29], adroitly leveraging pre-trained wav2vec 2.0 models. By ingeniously combining features with simple neural networks and trainable weights, this approach outshines standard emotion databases as corroborated by the existing literature.

Based on researchers' findings, it has become evident that utilizing the Transformer has yielded remarkable results in the field of speech emotion recognition. These impressive outcomes have prompted the exploration of new avenues, with the application of multi-modal data standing out as an exciting opportunity. Researchers increasingly understand that integrating diverse data sources can significantly enhance results. By merging information from various modalities such as audio, text, and potentially images, additional context is provided for emotion detection and comprehension.

Multimodal data is a type of data that integrates information from various sources, including text, audio, images, and video. This form of data is increasingly prevalent across numerous research domains, encompassing natural language processing, speech and emotion recognition, and computer vision. By harnessing multiple modalities, researchers can attain a more comprehensive grasp of intricate phenomena and enhance the precision of diverse tasks. For instance, multimodal data facilitates the detection of emotions in speech through the analysis of both audio and visual cues. Additionally, in natural language processing, multimodal data aids in extracting more meaningful features from text by incorporating contextual information from images or videos. Exploring multimodal data holds the potential to unlock novel insights and foster the development of more machine learning models. In the context of speech emotion recognition, multimodal data is pivotal. By merging audio and text data, researchers can gain a deeper understanding of the speaker's emotional state. For instance, in [19], [29], and [30], researchers utilized speech, text, and mocap data, including sub-modes such as facial expressions, hand gestures, and head rotations, to accurately identify emotions. Furthermore, [31] introduced a groundbreaking transformer-based model named multimodal transformers for audio-visual emotion recognition, overcoming the limitations of RNN and LSTM in capturing long-term dependencies. Three transformer branches are included in this model: audio-video cross-attention, video self-attention, and audio self-attention. The fusion of multiple modalities has consistently demonstrated its effectiveness in enhancing the accuracy of emotion recognition tasks.

Table I provides a comprehensive overview of the performance of the utilized model in comparison to other studies across a diverse range of datasets .This comparison effectively highlights how the proposed model outperforms previous approaches on diverse datasets, underscoring its remarkable success in achieving superior results within this domain. In the realm of Speech Emotion Recognition (SER), transformer-based approaches have demonstrated remarkable advancements. One instance is seen in "Multimodal Transformer for Speech Emotion Recognition with Shared Weights," which achieves a noteworthy accuracy of 77% on the IEMOCAP dataset [18]. Furthermore [19] have explored emotion datasets such as RAVDESS and Emo-DB, alongside a language-independent dataset. These investigations have showcased the effectiveness of a hybrid LSTM Network and Transformer Encoder, achieving significant SER accuracies of 75.62%, 85.55%, and 72.49%. Building upon this, transformer methodologies continue to make substantial contributions. For instance, the utilization of transformers and an attention-based fusion mechanism results in remarkable progress for emotion recognition on the IEMOCAP and MELD datasets [22].

Furthermore, the capabilities of transformers shine through as we delve deeper into their applications. The transformer model applied to the IEMOCAP dataset not only delivers notable accuracies, with 56.65% for speech, 68.94% for text, 53.14% for mocap, but also impressively reaches 74.59% for multimodal emotion recognition [29]. This multi-dimensional

approach highlights the versatility of transformer-based architectures.

Spearheading this revolution, the Swin-Transformer emerges with its own accolades, achieving an impressive accuracy of 82.55% on the IEMOCAP dataset [36]. Moreover, transformers transcend beyond traditional speech data. ViT, for instance, has proven its potential, achieving an 82.96% accuracy on the CREMA-D dataset by integrating spectrogram image-based techniques [33]. Continuing on this trajectory, ViT demonstrates its competence by securing accuracies of 56.18% and 37.1% on the IEMOCAP and MELD datasets, respectively [34].

TABLE I. PERFORMANCE OF THE PROPOSED MODEL AGAINST OTHER PUBLICATIONS ON DIFFERENT DATASET

Ref. No.	Dataset	Model	Accuracy (%)
[18]	IEMOCAP	Multimodal Transformer for Speech Emotion Recognition with Shared Weights	77
[21]	Emo-DB-URDU	Transformer	74.9 AND 80
[19]	RAVDESS, Emo-DB, a language-independent dataset	Transformer Encoder and hybrid Long Short-Term Memory (LSTM) Network for SER	75.62 85.55 72.49
[29]	IEMOCAP	Transformer	Speech 56.65 Text 68.94 Mocap 53.14 Multimodal 74.59
[22]	IEMOCAP, MELD	Transformers and Attention-based fusion mechanism	
[32]	EMO-DB	CNN-LSTM Mel Spectrogram-Vision Transformer	88.50 85.36 (surpassing existing benchmarks)
[28]	BAVED, EMO-DB, SAVEE, EMOVO	Transformer	95.2, 93.4 85.1 91.7
[33]	CREMA-D	ViT utilizing spectrogram images instead of sound data	82.96
[34]	IEMOCAP, MELD	ViT	56.18 37.1
[35]	IEMOCAP, EMODB, EMOVO, URDU	Multimodal Dual Attention Transformer (MDAT)	75.58 84.50 82.81 94.33
[36]	IEMOCAP	Swin-Transformer	82.55
[37]	IEMOCAP, RAVDESS	Transfer learning method using pre-trained wav2vec 2.0 models.	71.6 64.3
[38]	Tunisian Speech Emotion Recognition dataset (TuniSER)	fine-tuned multilingual wav2vec 2.0 model.	60.6

Incorporating pre-trained wav2vec 2.0 models into the mix, the research landscape evolves. The use of transfer learning yields promising results, as evidenced by accuracy rates of 71.6% and 64.3% on the IEMOCAP and RAVDESS datasets [37]. Finally, fine-tuning a multilingual wav2vec 2.0 model on the Tunisian Speech Emotion Recognition dataset (TuniSER) further underscores the transformer's adaptability across languages, producing an accuracy of 60.6% [38]. These interwoven advancements emphasize the transformative potential of transformer-based techniques in the evolving field of emotion recognition from speech.

III. THE PROPOSED EMOTIONAL RECOGNITION APPROACHES

In this exploration of prominent methodologies in the domain of speech emotion identification, we will spotlight three noteworthy techniques that have significantly impacted the domain. First and foremost, we will delve into the application of the Transformer architecture, showcasing its remarkable achievements in accurately identifying emotional states. Additionally, our examination will extend to the prowess of the Vision Transformer (ViT) within the context of audio analysis, revealing its robust capabilities in deciphering both sound patterns and emotional nuances. Lastly, we will turn our attention to Wave2Vec's role in facilitating knowledge transfer and collaborative learning, underlining its contribution to enhancing the field of speech emotion recognition. Through this comprehensive analysis, , It is our goal to shed light on the transformative potential of these techniques in advancing our understanding of emotions conveyed through speech. In Fig. 1, we present a comprehensive framework meticulously crafted for the purpose of classifying emotions within audio data.

A. Transformer-based Model

In the transformer architecture, attention is implemented as a function that requires a set of key-value pairs and a query vector and generates an output vector. These vectors represent the different components involved in the attention mechanism, including the query, keys, values, and output. A compatibility function determines the weights allocated to each value, and this process is used to generate the output vector. This compatibility function measures the degree of similarity or compatibility between the query and the corresponding key, enabling the model to determine the importance of different

key-value pairs in generating the output representation. It's worth noting that the transformer architecture has been employed in several studies [21], [28], [29], to advance its application and understanding.

By leveraging the attention mechanism, the transformer model can effectively focus on relevant information within the speech data, capturing important linguistic cues and contextual dependencies related to emotions. This allows for more accurate emotion recognition from speech, contributing to advancements in the field and opening up new possibilities for applications such as sentiment analysis, mental health monitoring, and human-computer interaction. In addition to the transformer architecture's self-attention mechanisms, it consists of two key components: Self-Attention (Scaled Dot-Product Attention, SDPA) and Multi-Head Attention (MHA). These components play a crucial role in enabling the transformer model to effectively recognize emotions from speech.

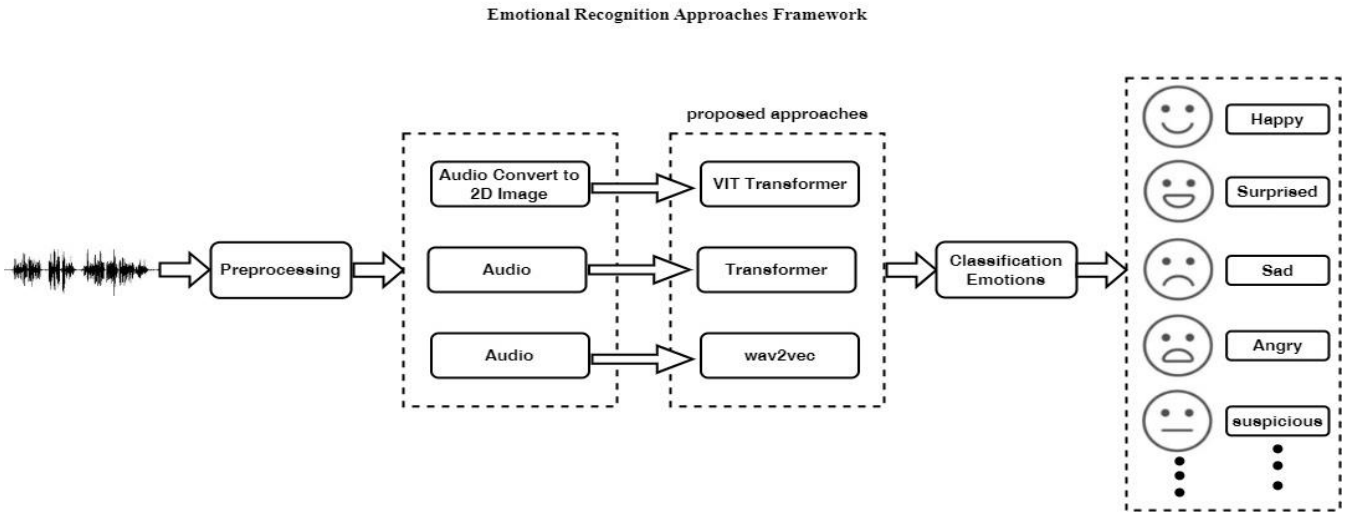

Fig. 1. Emotional recognition approaches framework.

The Scaled Dot Product Attention System (SDPA): is an essential part of the transformer architecture's multi-head attention (MHA) system. It plays a crucial role in capturing the significance and interrelations of different segments in speech input for emotional recognition tasks. SDPA computes attention scores between a query vector (Q) and a set of key vectors (K), representing encoded features of all words in the sample. The dot product between Q and K measures the influence of context words on the central word, revealing dynamics and connections among input tokens. This mechanism enables the model to focus on pertinent aspects of speech for emotional expression, applying the SoftMax function to ensure balanced attention scores. The final attention representation is obtained by multiplying the correlation matrix with the value vector (V), emphasizing significant information while downplaying less consequential portions. Mathematically, the SDPA process can be represented as follows:

$$A=\frac{QK^T}{\sqrt{d}} \quad (1)$$

$$S=soft\ max(A)V \quad (2)$$

$$SoftMax\ (Ai)=\frac{e^{A_i}}{\sum_{i=1}^{N} e^{A_i}} \quad (3)$$

Let A denote the output after scaling, and S represent the output of the attention unit. Q, K, and V are derived from the input feature vector with a shape of (N, d). Therefore, Q, K, and V are vectors of size RN*d, where N represents the length of the input sequence, and d represents the dimension of the input sequence. Typically, in ultralong sequence scenarios, it is observed that N > d, or even N >> d.

Expanding Eq. (2) based on the definition of SoftMax, we have:

$$Si=\frac{\sum_{j=1}^{N} \exp(\frac{q_i^T k_j}{\sqrt{d}})v_j}{\exp(\frac{q_i^T k_j}{\sqrt{d}})v_j} \quad (4)$$

In this equation, Qi, Ki, and Vi are column vectors representing the respective elements of Q, K, and V. Consequently, the mathematical essence of scaled dot product attention (SDPA) can be understood as a weighted average of the value vectors Vi, where weights are established by the exponential term ((qi^T * kj) / √d).

Multihead attention (MHA): is vital for parallel training in the transformer architecture, enabling simultaneous processing by dividing the input vector into multiple feature subspaces. It utilizes the self-attention mechanism, allowing parallel training while extracting essential information. In contrast to single-head average attention weighting, MHA enhances effective resolution, capturing diverse characteristics of speech features in different subspaces. This approach avoids inhibitory effects caused by average pooling on these characteristics MHA is calculated as follows:

$$Qi=XW_{Q_i}$$
$$Ki=XW_{k_i}$$
$$Vi=XW_{v_i}$$
$$Hi=SDPA\ (Qi, Ki, Vi)\ \ \forall i \in [1, n] \quad (5)$$

$$S=concat\ (H1, H2\ldots\ldots, Hn)\ W \quad (6)$$

Here, X represents the input feature sequence, and Qi, Ki, and Vi denote the query, key, and value vectors, respectively. Hi represents the attention scores of each head, and SDPA denotes the self-attention unit for each head. W represents the linear transformation weight. The index i ranges from 1 to n, where n is the number of heads, and i denotes the specific head.

The input feature sequence X is equally divided into n segments along the feature dimension. Each segment undergoes linear transformation, generating groups of (Qi, Ki, and Vi). Subsequently, Hi is individually calculated for each head. The n attention scores are then concatenated sequentially. Finally, the total attention score is obtained by applying linear transformation to the concatenated vectors.

B. Vit Transformer-based Model

The Vision Transformer approach represents a significant advancement in computer vision, leveraging the power of transformer architectures originally developed for natural language processing. ViT revolutionizes image understanding by separating an image into patches that don't overlap, linearly embedding those patches, and processing them using a standard transformer encoder. This methodology allows ViT to capture long-range dependencies within images, enabling it to excel in assignments like object detection and picture categorization. The self-attention mechanism of transformers enables ViT to effectively model contextual relationships among visual elements, contributing to its impressive performance.

In the context of speech emotion recognition, ViT's capabilities have been extended to handle speech data. By converting speech signals into 2D spectrogram images, ViT can efficiently process the visual representations of sound. This conversion enables ViT to recognize emotional cues present in speech, further enhancing its versatility in multimodal applications.

The core equation used in the ViT architecture is the self-attention mechanism, expressed as follows:

$$Attention\ (Q, K, V)=SoftMax(\frac{Qk^T}{\sqrt{d_k}})V \quad (7)$$

where:

- Q represents the query matrix,
- K denotes the key matrix
- V represents the value matrix,
- dk is the dimension of the key matrix.

The self-attention mechanism empowers ViT to assess the significance of diverse elements in a sequence, capturing intricate patterns essential for tasks like emotion recognition. ViT's capacity to learn complex patterns from raw image data, without relying on handcrafted features, has spurred its extensive use in various computer vision domains. This innovative approach has catalyzed progress in multimodal learning, integrating ViT with other modalities, like text and audio, to deepen the comprehension of complex data structures.

C. Wav2vec Transfer Model

Wav2Vec is a deep learning model that has been developed for speech processing and speech recognition tasks. Specifically, Wav2Vec is designed to convert raw audio waveforms (hence the "Wav" in its name) into meaningful representations that can be used by downstream speech recognition systems. Wav2Vec is particularly notable its capability of learning directly from raw audio data without requiring manual feature extraction. It employs a self-supervised learning approach, where the model learns by predicting future audio samples based on past samples. This helps the model to record high-level characteristics and patterns within speech, making it well-suited for speech recognition tasks.

Fine-Tuning Explained: Fine-tuning is a process that capitalizes on the knowledge and features a model has gained from being trained on a large dataset. Instead of training a model from scratch, which can be computationally expensive and time-consuming, fine-tuning leverages the existing knowledge encoded in a pre-trained model. By modifying specific layers or weights of the model and training it further on a smaller, task-specific dataset, the model can learn to perform the new task more effectively. Fine-tuning the pre-trained Wav2Vec model for emotion classification enhances its ability to discern emotional cues from speech, making it a valuable tool for a variety of applications, including sentiment analysis, virtual assistants, and affective computing systems.

Fine-Tuning Loss Function: During fine-tuning, a common loss function used for classification tasks like emotion classification is the categorical cross-entropy loss. It calculates the difference between the true class labels and the expected class probabilities.

$$Loss= -\sum_i y_i\ log(\hat{y}_i) \quad (8)$$

where:

- y_i is the true probability of class.
- $\hat{y}_i$ is the predicted probability of class ii.

This loss function penalizes large differences between predicted and true probabilities, encouraging the model to update its parameters to improve classification accuracy.

IV. Expermintal Setup

In this section, we will explain three sub-sections: Data Set, Preprocessing, and Results of Experiments.

A. Dataset

Databases are essential for speech emotion recognition, as the classification process relies heavily on labeled data. The accuracy of the recognition process is directly impacted by the quality of the data used. Incomplete, poor-quality, or flawed data can lead to incorrect predictions. The effectiveness of the classification is also influenced by factors such as language, the number of emotions, and the data collection method. Thus, it is crucial to carefully.

Design and collect the data. For example, to recognize emotions through speech, data sets in multiple languages, including English, German, Swedish, Turkish, French, Mandarin, Italian, Japanese, and Arabic have been employed. Ensuring high-quality data sets is vital for accurate and reliable speech emotion recognition.

The Arabic Natural Audio Dataset (ANAD) and the Toronto Emotional Speech Set (TESS) were used in this study to assess how well the suggested speech emotion recognition technique worked. The ANAD dataset is a publicly available dataset comprised of Arabic audio files obtained from online Arabic talk shows. Specifically, eight videos of live calls between a host and an external person were downloaded and segmented into turns involving callers and receivers. To classify videos, 18 listeners assessed emotions like happiness, anger, and surprise. After removing silence, laughs, and noise, the chunks were automatically divided into one-second speech units. The resulting corpus consisted of 1384 records, as depicted in Fig. 2, which illustrates the number of audio files for each emotion in the dataset. The dataset size is 587MB. The usage of ANAD provides a valuable opportunity to assess the proposed method's efficacy in recognizing emotions in natural Arabic speech. The TESS dataset, on the other hand, contains English audio files representing seven emotions: neutrality, pleasant surprise, anger, disgust, fear, and happiness. There are 2800 audio files in this collection, as shown in Fig. 3, which illustrates the number of audio files for each emotion in the dataset. These recordings were made by two females, ages 26 and 64. The dataset size for TESS is 449MB.By utilizing these datasets; the study aims to evaluate the performance and accuracy of the proposed speech emotion recognition method in natural Arabic speech (ANAD) and English speech (TESS), covering a range of emotions. The availability of ANAD and TESS datasets allows for a comprehensive assessment of the proposed method's capabilities in recognizing emotions across different languages and contexts. The selection of datasets was based on their credibility and widespread use in the field of Speech Emotion Recognition (SER). These well-established datasets allow for an effective comparison of the proposed model's performance with other studies that use the same datasets. Tables II and III illustrate how many audio segments exist for each expression in each dataset.

TABLE II. TABLE SHOWS THE NUMBER OF AUDIO CLIPS FOR EACH EMOTION IN ANAD DATASET

DATA SET	TESS
Emotions	***Number of audio files***
Angry	400
Happy	400
Surprise	400
Disgust	400
Fear	400
Sad	400
Neutral	400

TABLE III. TABLE SHOWS THE NUMBER OF AUDIO CLIPS FOR EACH EMOTION IN TESS DATASET

DATA SET	ANAN
Emotions	***Number of audio files***
Happy	505
Angry	741
Surprised	137

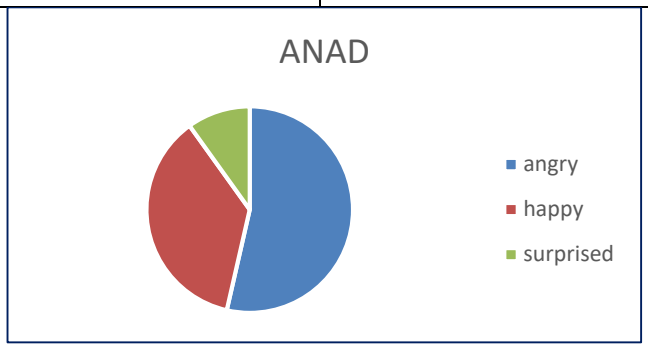

Fig. 2. The data distribution of emotion in ANAD.

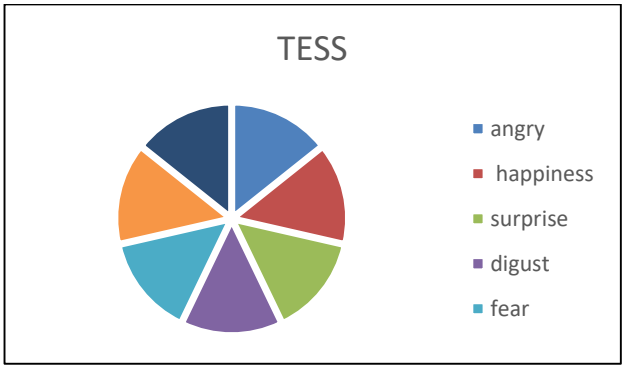

Fig. 3. The data distribution of emotions in TESS.

B. Preprocessing

Preprocessing plays a vital role in preparing raw data for analysis and model training. It involves a series of data transformation steps to clean, normalize, and enhance the dataset's quality. Standard techniques include data cleaning to remove missing values or outliers, feature scaling to bring features within a consistent range, and feature engineering to extract relevant information. Proper preprocessing ensures that the data is in a suitable format for the specific analysis or model, reducing noise and improving performance. It also helps address potential biases or inconsistencies, ultimately contributing to more accurate and reliable research findings.

The "Arabic Natural Audio Dataset" was used in this study. It consists of eight videos downloaded from online Arabic talk shows, capturing live calls between an anchor and an individual outside the studio. The videos were divided into turns of callers and receivers, and emotions in each video were labelled by 18 listeners (happy, angry, or surprised). After removing silence, laughs, and noise, the audio was automatically divided into 1-second speech units, resulting in a corpus of 1384 records.

To extract features from the audio, we collected twenty-five acoustic features or low-level descriptors (LLDs). These features included intensity, zero-crossing rates, Mel-frequency cepstral coefficients (MFCC 1-12), fundamental frequency (F0), F0 envelope, probability of voicing, and LSP frequency 0-7. Each feature underwent nineteen statistical functions, such as maximum, minimum, range, absolute positions of maximum and minimum, mean arithmetic, various linear regression functions, standard deviation, kurtosis, skewness, quartiles 1, 2, 3, and inter-quartile ranges 1-2, 2-3, 1-3. Additionally, we computed the delta coefficient for each LLD to estimate the first derivative, resulting in a total of 950features.

When it comes to recognizing emotions from speech, audio files can be transformed into 2D spectrogram images. These images show the frequency content of the audio signal over time and allow the audio data to be treated as an image. This makes it possible to use computer vision-based models like transformers to analyze the spectrograms. By training a transformer model on these spectrogram images, the model can learn the patterns in the audio, both over time and in frequency. This enables the model to identify emotions based on the distinct features present in the spectrograms. This approach has been successful in recognizing emotions from speech data, thanks to the power of transformers to capture long-range dependencies and achieve high accuracy. To prepare these images for training, the ImageDataGenerator was used to preprocess and flow batches of grayscale images and corresponding emotion labels from a DataFrame. The images were resized to a target size of 128x128 pixels and assumed to be grayscale, indicated by the color_mode="grayscale" parameter.

In the wave2vector experiment, all audio files were converted to a standard sampling rate of 16000 Hz.

C. Experimental Results

When creating machine learning models, separating data into training, validation, and testing sets is crucial. Three subsets of the original dataset have been identified: the training set, validation set, and testing set. Each set has a specific role in developing and evaluating the model. By dividing the data in this way, we can assess the model's performance in various scenarios. The allocation of data to each set is customized for each experiment, ensuring a fair evaluation of the model's abilities.

In the first experiment, we employed the ANAD dataset and applied the TRANSFORMER approach to audio data, partitioning it into 80% for training, 20% for testing, and reserving an additional 15% for validation within the training set.

In the second experiment, we used the TESS dataset and employed the VIT TRANSFORMER approach to convert audio data into 2D images. The data was divided into 80% for training, 20% for testing, with an extra 15% set aside for validation.

For the third experiment, we leveraged the ANAD dataset with audio data, utilizing the Wave2Vec approach. Data allocation involved dedicating 68% to training from the original dataset, allocating 17% to testing, and reserving 15% for validation.

Experiment 1: In this experiment, we used the following audio processing parameters: sampling_rate = 30100, duration = 1, hop_length = 300, fmin = 20, n_mels = 128, time_steps = 128, and epochs = 80 and 40.

The researchers used the Arabic Natural Audio Dataset (ANAD) , which was previously employed in [31] Novel emotion recognition for Arabic speech using deep feed-forward neural network (DFFNN) achieves 98.56% accuracy with PCA and 98.33% with combined features from ANAD dataset. In [39] evaluate three speaker traits—gender, emotion, and dialect—from Arabic speech, employing multi-task learning (MTL). The dataset, assembled from six publicly available datasets, including the ANAD dataset, underwent exploration with three networks—LSTM, CNN, and FCNN—across different features. Multi-task learning consistently demonstrated superior performance compared to single task learning (STL). Results for emotion classification are as follows: For LSTM STL achieved 50.4%, and MTL 57.05%, CNN: STL 51.18%, and MTL 51.25% and FCNN: STL 66.53%, and MTL 70.16%. Results show improvement over previous studies. In the first experiment, the same ANAD dataset was used. However, the data underwent preprocessing and was converted into a numerical 2D array before being fed into the Transformer model. As a result of this approach, the Transformer model achieved a high level of performance, reaching 94% accuracy in its predictions. This suggests that representing the data as a 2D numerical array and utilizing the Transformer model was effective in extracting valuable patterns and features from the dataset. In addition to the previous experiment where the data was represented as a numerical 2D array and fed into the Transformer model, there was another aspect to the study. In this alternative approach, the same dataset (ANAD) was entered into the Transformer model as 2D images. Interestingly, this variation yielded a slightly lower accuracy of 88% compared to the 94% accuracy achieved with the numerical 2D array representation. This suggests that the numerical format may have been more suitable for this specific dataset and the task at hand. The researchers concluded that representing the data in a specific format, such as a numerical 2D array, can significantly impact the model's performance. It is crucial to explore different data representations and preprocessing techniques to determine the most suitable approach for the given task and dataset.

Experiment 2: During the audio processing experiment, the following parameter values were used: sampling rate of 50000, duration of 1 second, hop length of 300, minimum frequency (fmin) of 20, number of Mel filters (n_mels) set to 128, time steps at 128, and 60 epochs. [40] utilized the Toronto Emotional Speech Set (TESS) which was previously used in a study comparing CNN-based emotion recognition using spectrograms and Mel-spectrograms and found Mel-spectrograms to be more suitable for Speech Emotion Recognition (SER). The study used four datasets, including TESS, which has six emotion classes. The most accurate model obtained an accuracy of 57.42% on four datasets, including TESS. In[41] combines RAVDESS and TESS datasets for emotion classification from speech, extracting 180

features using various techniques. Gradient Boosting excels with 84.96% accuracy on the merged dataset. The datasets RAVDESS and TESS datasets were integrated using CNN, yielding a 97.1% accuracy in [42]. RAVDESS, TESS and SAVEE datasets were integrated using neural network yielding a testing accuracy of about 89.26% in [43]. In the second experiment, the same TESS dataset was used, but it underwent preprocessing to convert the data into a numerical 2D array before being fed into the Transformer model. This approach achieved an impressive accuracy of 99%, highlighting the effectiveness of representing data as a 2D numerical array and utilizing the Transformer model to extract essential patterns and features from the dataset.

The researchers also tried a different method by using the ANAD dataset as 2D images with the Transformer model. However, this approach resulted in slightly lower accuracy of 98% compared to the 99% accuracy achieved with the numerical 2D array representation. These findings indicate that converting audio data into 2D images had a significant impact on the model's performance. Therefore, the numerical 2D array representation is more effective for this dataset and task.

The study underscores the importance of researching different data representations and preprocessing techniques to achieve optimal performance in machine learning models. It highlights that there is no one-size-fits-all approach, and the choice of data representation should align with the dataset's characteristics and the specific task at hand.

Ultimately, this research contributes valuable insights into the influence of data representation on deep learning model performance and knowledge extraction from audio datasets. It may pave the way for the application of such techniques in various fields, including machine empathy, emotion analysis, and voice recognition.

Experiment 3: In the third experiment, we leveraged the ANAD dataset with audio data, utilizing the Wave2Vec approach. Data allocation involved dedicating 68% to training from the original dataset, allocating 17% to testing, and reserving 15% for validation. The primary objective of this experiment was to conduct emotion classification through fine-tuning the Wave2Vec model. Following the training and evaluation, the model achieved an accuracy of approximately 56%. This implies that the model attained a correct classification rate of 56% on the testset. Tables IV and V display the findings achieved by the three models across both the ANAD and TESS datasets.

TABLE IV. TABLE SHOWING APPROCHES RESULT ON ANAD DATA SET

DATA SET	**ANAD**		
APPROCH	***TRASFORMER***	***VIT TRANSFORMER***	***WAVE2VECTOR***
ACCURACY	94%	88%	56%

TABLE V. TABLE SHOWING APPROCHES RESULT ON TESS DATA SET

DATA SET	**TESS**	
APPROCH	***TRASFORMER***	***VIT TRANSFORMER***
ACCURACY	99%	98%

Fig. 4 to Fig. 7 display accuracy and loss curves of a machine learning model. These figures reveal performance trends over training epochs, guiding model adjustments. Furthermore, the accuracy curves for both training and validation of the TESS dataset uses the impressive accuracy of 99% and 98%, respectively. These accuracies surpass those observed in the validation and training accuracy curves. This improvement could potentially be attributed to the imbalanced distribution of the data within the ANAD dataset.

In Fig. 4 and 5, illustrating the ANAD dataset's training and validation accuracy, it is clear that the model's performance improved considerably. Initially, the model began with an accuracy close to zero, but gradually, it showed a steady enhancement, ultimately reaching accuracies of 94% for the Transformer model and 88% for the ViT Transformer model. This progressive improvement reflects the model's growing comprehension of the dataset and its overall performance enhancement. This transition from near-zero accuracy to high accuracy underscores the model's learning process and its ability to successfully capture intricate patterns within the data.

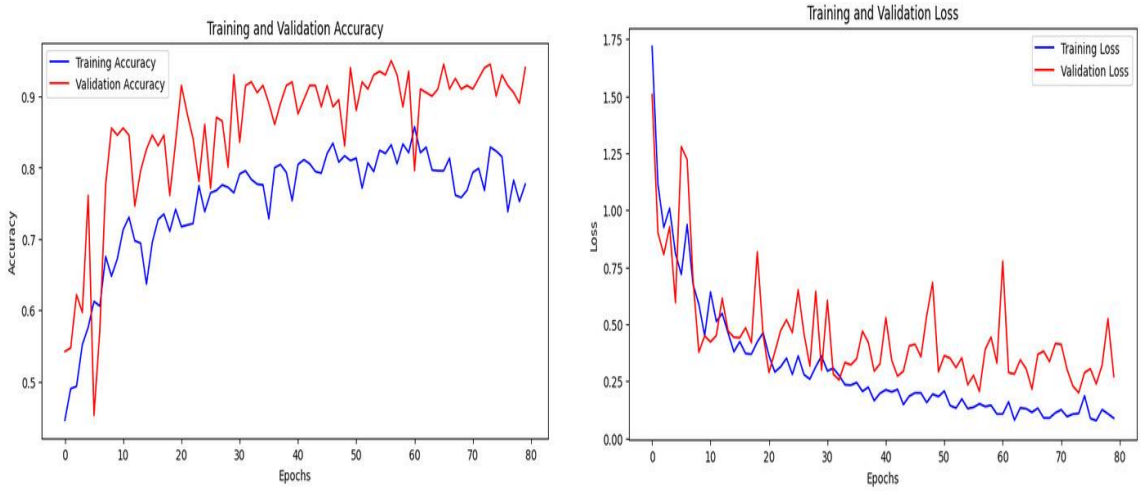

Fig. 4. ANAD dataset accuracy (left) and loss (right) curves for transformer model.

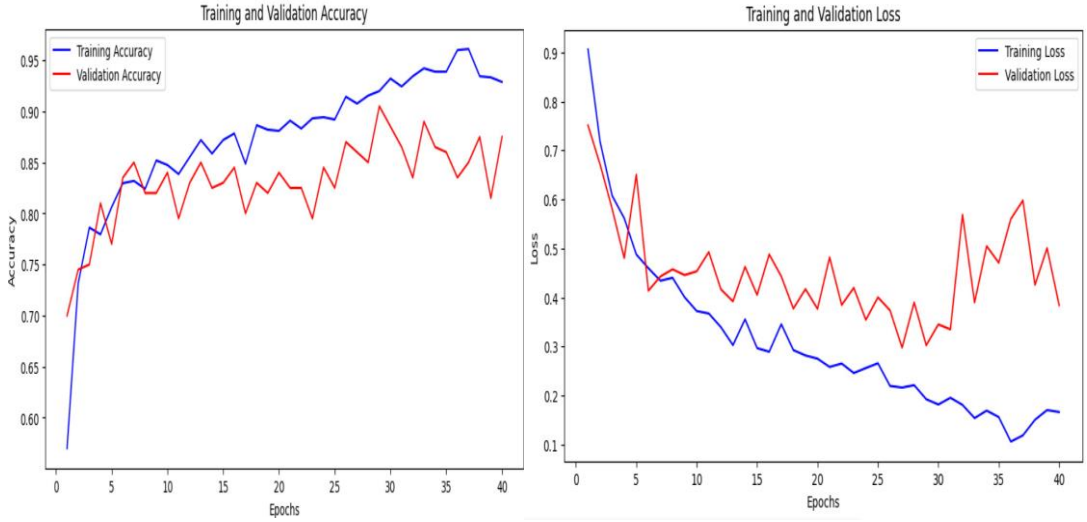

Fig. 5. ANAD dataset accuracy (left) and loss (right) curves for ViT transformer model.

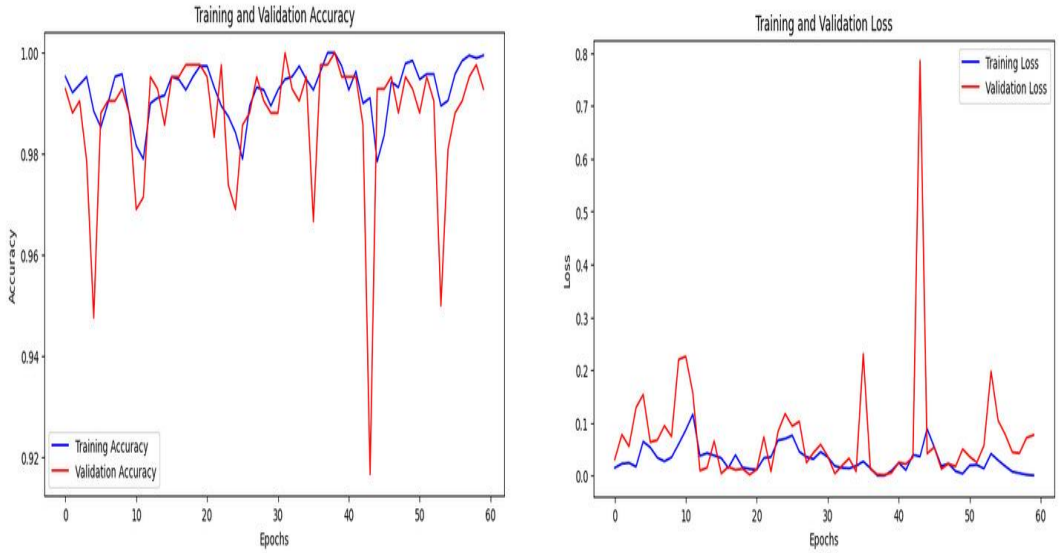

Fig. 6. TESS dataset accuracy (left) and loss (right) curves for transformer model.

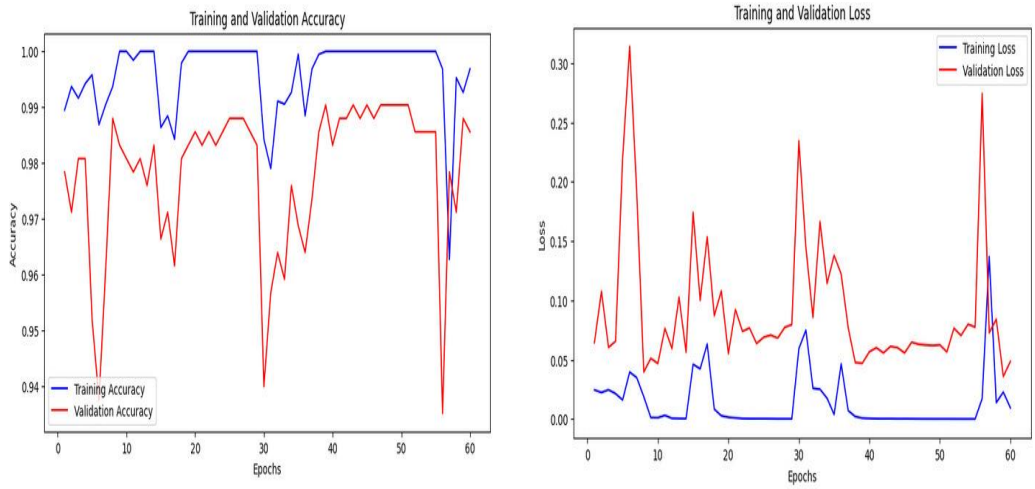

Fig. 7. TESS dataset accuracy (left) and loss (right) curves for ViT transformer model.

V. Conclusion and Future Work

In conclusion, this research highlights the significant impact of sound on human well-being, recognized since ancient civilizations and still relevant in our modern world. It emphasizes the fast-growing field of speech emotion recognition, which has found diverse applications in enhancing human-computer interaction, aiding mental health diagnosis, and facilitating human-robot interaction. The core focus of this study is the classification of emotions from speech. The proposed three approaches, utilizing transformer-based deep learning models, demonstrates its efficacy in accurately identifying and categorizing emotions from both audio signals and transformed 2D images. The experimental evaluations on the Arabic Natural Audio Dataset (ANAD) and the Toronto Emotional Speech Set (TESS) have produced highly promising results. For the audio-based emotion classification model, ANAD achieved an impressive 94% accuracy, while TESS achieved an equally remarkable 99% accuracy. On the other hand, the image-based emotion classification model attained 88% accuracy for ANAD and 98% accuracy for TESS. These high accuracy rates show how reliable and successful the suggested method is in identifying the emotions expressed in speech. Additionally, the research incorporates the fine-tuning of wav2vec for emotion classification from the ANAD dataset, leading to a respectable 56% accuracy. While slightly lower than the other models, this result still showcases the practical implementation of fine-tuning in achieving reasonable accuracy rates in emotion classification from speech data.

The research underscores the potential of both approaches: direct audio usage and transforming audio into 2D images, yielding comparable results. Despite the vision-based model showing advantages with more data, the matrix input approach ultimately proved superior. The study's use of three diverse approaches, particularly transformer-based models, was crucial for success in emotion recognition from speech. Transformer models consistently excel in natural language processing and audio data extraction. Furthermore, diverse datasets like ANAD and TESS, enriched with varied voices and emotional expressions, significantly contributed to achieving remarkable results, enhancing model effectiveness.

In light of the promising findings and practical implications of this research, several avenues for future work can be explored. Firstly, it is recommended to further investigate the performance of the proposed three approach using larger and more diverse datasets. Expanding the dataset size can potentially improve the accuracy and robustness of the emotion classification models. Additionally, extending the application of the models to datasets that contain non-language-specific vocal expressions can be an interesting direction. By analyzing vocal expressions unrelated to a specific language, the models can be tested for their ability to capture universal emotional cues, thus enhancing their generalizability. Furthermore, it would be valuable to explore the transferability of the trained models to different domains and applications. Applying the models to datasets that are unrelated to the ones used in training, such as real-world scenarios or specific professional environments, can shed light on their adaptability and effectiveness in practical settings. In terms of methodology, incorporating multimodal approaches by Compiling speech data with additional modalities, such as physiological signs or facial expressions, can yield a more comprehensive understanding of emotions. This integration of multiple modalities can potentially enhance the accuracy and richness of emotion classification systems. Moreover, fine-tuning wav2vec in future research can be instrumental in achieving even better results than the current accuracy. Fine-tuning offers opportunities to fine-tune pre-trained models to specific datasets, leading to improved performance and more accurate emotion classification from speech. Lastly, considering the ethical implications of emotion classification from speech is crucial. Future work should address privacy concerns and ensure the responsible and transparent use of such technologies. Developing guidelines and frameworks for the ethical implementation and deployment of these models will be essential to build trust and ensure their positive impact on society.

In summary, future research should focus on expanding the dataset size, exploring non-language-specific vocal expressions, testing the models on different domains, incorporating multimodal approaches, and addressing ethical considerations. By addressing these areas and leveraging fine-tuning techniques, the proposed three approaches can be further improved and applied to a wider range of practical applications, advancing the domains of emotion recognition and human-computer interaction.

References

[1] T. Seehapoch and S. Wongthanavasu, “Speech emotion recognition using support vector machines,” in 2013 5th international conference on Knowledge and smart technology (KST), IEEE, 2013, pp. 86–91.

[2] Y. Chavhan, M. L. Dhore, and P. Yesaware, “Speech emotion recognition using support vector machine,” Int. J. Comput. Appl., vol. 1, no. 20, pp. 6–9, 2010.

[3] P. Shen, Z. Changjun, and X. Chen, “Automatic speech emotion recognition using support vector machine,” in Proceedings of 2011 international conference on electronic & mechanical engineering and information technology, IEEE, 2011, pp. 621–625.

[4] X. Cheng and Q. Duan, “Speech emotion recognition using gaussian mixture model,” in 2012 International Conference on Computer Application and System Modeling, Atlantis Press, 2012, pp. 1222–1225.

[5] I. J. Tashev, Z.-Q. Wang, and K. Godin, “Speech emotion recognition based on Gaussian mixture models and deep neural networks,” in 2017 information theory and applications workshop (ITA), IEEE, 2017, pp. 1–4.

[6] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, 2003.

[7] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, “Arabic speech emotion recognition from saudi dialect corpus,” IEEE Access, vol. 9, pp. 127081–127085, 2021.

[8] V. Chernykh and P. Prikhodko, “Emotion recognition from speech with recurrent neural networks,” arXiv Prepr. arXiv1701.08071, 2017.

[9] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, “Speech emotion recognition using recurrent neural networks with directional self-attention,” Expert Syst. Appl., vol. 173, p. 114683, 2021.

[10] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, and G. Dai, “Pre-trained deep convolution neural network model with attention for speech emotion recognition,” Front. Physiol., vol. 12, p. 643202, 2021.

[11] S. Parthasarathy and I. Tashev, “Convolutional neural network techniques for speech emotion recognition,” in 2018 16th international

workshop on acoustic signal enhancement (IWAENC), IEEE, 2018, pp. 121–125.

[12] M. D. Pawar and R. D. Kokate, “Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients,” Multimed. Tools Appl., vol. 80, pp. 15563–15587, 2021.

[13] D. Issa, M. F. Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” Biomed. Signal Process. Control, vol. 59, p. 101894, 2020.

[14] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and recurrent neural networks,” in 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), IEEE, 2016, pp. 1–4.

[15] A. Valiyavalappil Haridas, R. Marimuthu, V. G. Sivakumar, and B. Chakraborty, “Emotion recognition of speech signal using Taylor series and deep belief network based classification,” Evol. Intell., pp. 1–14, 2020.

[16] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” Biomed. Signal Process. Control, vol. 47, pp. 312–323, 2019.

[17] B. Tris Atmaja and M. Akagi, “Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition,” arXiv e-prints, p. arXiv-2004, 2020.

[18] Y. Wang, G. Shen, Y. Xu, J. Li, and Z. Zhao, “Learning Mutual Correlation in Multimodal Transformer for Speech Emotion Recognition.,” in Interspeech, 2021, pp. 4518–4522.

[19] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, “Hybrid LSTM-transformer model for emotion recognition from speech audio files,” IEEE Access, vol. 10, pp. 36018–36027, 2022.

[20] S. Siriwardhana, A. Reis, R. Weerasekera, and S. Nanayakkara, “Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition,” arXiv Prepr. arXiv2008.06682, 2020.

[21] D. Jing, T. Manting, and Z. Li, “Transformer-like model with linear attention for speech emotion recognition.,” J. Southeast Univ. (English Ed., vol. 37, no. 2, 2021.

[22] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, “Multimodal emotion recognition with transformer-based self supervised feature fusion,” IEEE Access, vol. 8, pp. 176274–176285, 2020.

[23] L. Tarantino, P. N. Garner, and A. Lazaridis, “Self-Attention for Speech Emotion Recognition.,” in Interspeech, 2019, pp. 2578–2582.

[24] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez, “A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset,” Appl. Sci., vol. 12, no. 1, p. 327, 2021.

[25] A. Arezzo and S. Berretti, “Speaker vgg cct: Cross-corpus speech emotion recognition with speaker embedding and vision transformers,” in Proceedings of the 4th ACM International Conference on Multimedia in Asia, 2022, pp. 1–7.

[26] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, “ViTFER: facial emotion recognition with vision transformers,” Appl. Syst. Innov., vol. 5, no. 4, p. 80, 2022.

[27] Z. Zhao, Y. Wang, and Y. Wang, “Multi-level fusion of wav2vec 2.0 and BERT for multimodal emotion recognition,” arXiv Prepr. arXiv2207.04697, 2022.

[28] B. B. Al-onazi, M. A. Nauman, R. Jahangir, M. M. Malik, E. H. Alkhammash, and A. M. Elshewey, “Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion,” Appl. Sci., vol. 12, no. 18, p. 9188, 2022.

[29] R. A. Patamia, W. Jin, K. N. Acheampong, K. Sarpong, and E. K. Tenagyei, “Transformer based multimodal speech emotion recognition with improved neural networks,” in 2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML), IEEE, 2021, pp. 195–203.

[30] V. John and Y. Kawanishi, “Audio and video-based emotion recognition using multimodal transformers,” in 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 2582–2588.

[31] E. R. Abdelmaksoud, “Arabic Automatic Speech Recognition Based on Emotion Detection,” Egypt. J. Lang. Eng., vol. 8, no. 1, pp. 17–26, 2021.

[32] C. S. A. Kumar, A. Das Maharana, S. M. Krishnan, S. S. S. Hanuma, G. J. Lal, and V. Ravi, “Speech Emotion Recognition Using CNN-LSTM and Vision Transformer,” in International Conference on Innovations in Bio-Inspired Computing and Applications, Springer, 2022, pp. 86–97.

[33] J.-Y. Kim and S.-H. Lee, “CoordViT: A Novel Method of Improve Vision Transformer-Based Speech Emotion Recognition using Coordinate Information Concatenate,” in 2023 International Conference on Electronics, Information, and Communication (ICEIC), IEEE, 2023, pp. 1–4.

[34] X. Huang, Q. Zheng, Y. Zhang, D. Cheng, Y. Liu, and C. Dong, “Speech emotion analysis based on vision transformer,” in 2022 2nd Conference on High Performance Computing and Communication Engineering (HPCCE 2022), SPIE, 2023, pp. 400–405.

[35] S. A. M. Zaidi, S. Latif, and J. Qadi, “Cross-Language Speech Emotion Recognition Using Multimodal Dual Attention Transformers,” arXiv Prepr. arXiv2306.13804, 2023.

[36] Z. Liao and S. Shen, “Speech Emotion Recognition Based on Swin-Transformer,” in Journal of Physics: Conference Series, IOP Publishing, 2023, p. 12056.

[37] L. Pepino, P. Riera, and L. Ferrer, “Emotion recognition from speech using wav2vec 2.0 embeddings,” arXiv Prepr. arXiv2104.03502, 2021.

[38] A. Messaoudi, H. Haddad, M. B. Hmida, and M. Graiet, “TuniSER: Toward a Tunisian Speech Emotion Recognition System,” in Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022), 2022, pp. 234–241.

[39] W. Farhan, M. E. Za’ter, Q. A. Obaidah, H. al Bataineh, Z. Sober, and H. T. Al-Natsheh, “SPARTA: Speaker Profiling for ARabic TAlk,” in 2021 28th Conference of Open Innovations Association (FRUCT), IEEE, 2021, pp. 103–110.

[40] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, “Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets,” Electronics, vol. 11, no. 22, p. 3831, 2022.

[41] A. S. Nasim, R. H. Chowdory, A. Dey, and A. Das, “Recognizing Speech Emotion Based on Acoustic Features Using Machine Learning,” in 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), IEEE, 2021, pp. 1–7.

[42] R. R. Choudhary, G. Meena, and K. K. Mohbey, “Speech emotion based sentiment recognition using deep neural networks,” in Journal of Physics: Conference Series, IOP Publishing, 2022, p. 12003.

[43] B. Salian, O. Narvade, R. Tambewagh, and S. Bharne, “Speech Emotion Recognition using Time Distributed CNN and LSTM,” in ITM Web of Conferences, EDP Sciences, 2021, p. 3006.