IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 February 2025, accepted 2 April 2025, date of publication 10 April 2025, date of current version 18 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3559339

RESEARCH ARTICLE

BiMER: Design and Implementation of a Bimodal Emotion Recognition System Enhanced by Data Augmentation Techniques

EMRAH DIKBIYIK[1,2], ONDER DEMIR[3], AND BUKET DOGAN[3]

[1]Department of Computer Technologies, Vocational School of Technical Sciences, Istanbul University-Cerrahpaşa, 34500 Istanbul, Türkiye
[2]Institute of Pure and Applied Sciences, Marmara University, 34722 Istanbul, Türkiye
[3]Department of Computer Engineering, Faculty of Technology, Marmara University, 34854 Istanbul, Türkiye

Corresponding author: Emrah Dikbiyik (emrahdikbiyik@iuc.edu.tr)

This work was supported in part by the Marmara University Scientific Research Coordination Unit under Grant FDK-2021-10294.

ABSTRACT In today's world, accurately understanding and interpreting emotions in human-computer interaction is important. In this context, this study has adopted a detailed approach to the emotion recognition problem on both speech and text data using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. First, the problem of datasets with limited number of records and unbalanced distribution across classes was addressed. For this purpose, a dataset obtained from records created as improvised in the IEMOCAP dataset was used and data augmentation methods were applied for both speech and text data. Using datasets that were balanced by applying data augmentation, single-mode emotion recognition experiments were performed with models developed for Speech Emotion Recognition (SER) and Textual Emotion Recognition (TER). Subsequently, the features obtained from these two single modalities were combined with the intermediate fusion method to provide more comprehensive emotion recognition and accuracy, and the Bimodal Emotion Recognition (BiMER) system was developed. The ResNet50-CRNN+AT model, which we obtained the highest accuracy from the three different models developed for SER, creates the speech mode of BiMER, while the Bidirectional Encoder Representations from Transformers (BERT) model used for TER creates the text mode of BiMER. In this way, BiMER was supported with data augmentation methods and the robustness and generalization ability of the model were improved, reaching 88.33% accuracy. Finally, the developed BiMER system was implemented as a real-time web application using the Flask framework, and the capacity of this application to recognize emotions interactively through the user interface was tested.

INDEX TERMS Bimodal emotion recognition, intermediate fusion, data augmentation, real-time emotion recognition, IEMOCAP.

I. INTRODUCTION

Emotions play a significant role in our daily lives. People's body movements, tone of voice, facial expressions, and the words they use while speaking provide clues about their emotions. In the rapidly evolving artificial intelligence environment, the ability to accurately recognize and interpret these emotions plays an important role in bridging the gap between applications using information technologies and human communication. In a world where digital communication channels and multimedia environments are becoming increasingly dominant, the ability to correctly interpret emotional expressions in speech and text will enhance user experience [1]. Furthermore, it will support application areas related to education [2], healthcare [3], [4], security [5], and marketing [6], and will open new avenues for enhancing human-computer interaction.

Approaches to emotion recognition can be examined in four groups as studies conducted on text, studies conducted on speech, studies conducted on images of facial and body

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

movements, and techniques that use them together [7]. On the other hand, there are also studies that use physiological signals obtained from the human body through sensors for emotion recognition [8]. The process of determining emotional states from vocal features is referred to as Speech Emotion Recognition (SER), whereas analyzing emotions from text is known as Textual Emotion Recognition (TER). Traditional approaches to emotion recognition have mainly focused on a single mode, primarily analyzing speech or text. However, human communication is generally multimodal in nature. Multimodal Emotion Recognition (MER) integrates information from multiple modalities to convey all emotional states for this purpose. MER approaches are methods that combine analyses of text, speech and images to more accurately determine emotional intensity [9].

Each emotion carries unique information, reflecting different psychological and situational states. The emotions we shape in our inner world emerge at various levels of intensity to convey any message about the situation we are in [10]. Different models presented regarding the theory of emotion offer different perspectives in understanding the complexity and richness of human emotions. Emotion theories are generally examined under two main approaches: discrete emotional model and dimensional emotional model [11]. The discrete emotional model treats emotions as distinct and separate categories, aiming to recognize and classify specific emotional reactions. Basic emotions are usually grouped under certain universally accepted labels. One of the most important advocates of this model is psychologist Paul Ekman. The model developed by Paul Ekman offers a universal understanding of emotions by defining basic emotions and includes the six basic emotions of anger, disgust, fear, happiness, sadness, and surprise [12], [13]. Ekman's model classifies emotions into clearly defined and distinct categories and is therefore considered a discrete emotional model. On the other hand, Robert Plutchik's model presents a ''wheel of emotions'' representing eight basic emotions (acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise) and their interactions with each other [14]. Dimensional models, conversely, assess emotions on a continuous spectrum and usually measure emotions along several axes (dimensions). These models can better reflect the complexity and gradients of emotions. The Valence, Arousal, Dominance (VAD) model presented by Russell and Mehrabian examines emotions along three main dimensions [15]. This model allows emotions to be analyzed not only with rigid categories but also by considering the intensity and mix of each emotion [16].

In the developing field of human-computer interaction, understanding emotional cues and recognizing emotions through various forms of communication is very important. Within the scope of the study, the emotion recognition problem was addressed using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [17] and focused on basic emotions (anger, happiness, sadness, neutral) compatible with Ekman's emotion model. The fact that the Ekman model is based on universal emotion theory allows consistent classification of emotional expressions across different modes, especially text and audio, when working with multimodal datasets. Another issue that the study focuses on is the limited number of records in the datasets and the unbalanced distribution between classes. In this context, various data augmentation techniques were applied to both speech and text data in the dataset used, enriching the datasets and making them more balanced. Data augmentation plays a critical role in improving performance in emotion recognition tasks by allowing emotion recognition models to generalize to wider and more diverse situations.

The main motivation of this study is to provide a more holistic and accurate interpretation of emotions. For this purpose, the study focuses on two main modalities: Speech Emotion Recognition (SER) and Textual Emotion Recognition (TER). As a result of the study, a Bimodal Emotion Recognition (BiMER) system that uses speech and text data and integrates intermediate fusion technology has been developed, and this system has been implemented as a real-time web application using Flask framework.

The remaining parts of this article are organized as follows: Section II summarizes 'Literature Review'; Section III, titled 'Proposed Methods' details the models developed for SER, TER and BiMER, as well as the augmentation methods used; Section IV describes the implementation of the Bimodal Emotion Recognition model as a web-based application; Section V presents the experimental results of the developed models; while Section VI provides the conclusions of the study.

II. LITERATURE REVIEW

In this section, various studies that utilize data augmentation for SER and TER systems, as well as studies on Bimodal Emotion Recognition systems, are reviewed. The reviewed studies are explained in a way that includes models developed for emotion recognition and their achievements.

The literature review revealed that various methods are used for data augmentation in audio and text; moreover, it has been observed that models employing deep learning architectures are preferred during the emotion identification phase.

In 2021, Xu et al. published a study that aimed to create diversity in voice data using the Vocal Tract Length Perturbation (VTLP) method. VTLP was used to increase the generalization ability of the model by slightly changing the characteristics of the voice. In their study, they increased the size of the training dataset by producing seven different copies in addition to the original data. They used the multi-layer area attention mechanism to increase the performance in emotion recognition. In the experiments conducted on the IEMOCAP data set, they reached the values of Weighted Accuracy (WA) 79.34% and Unweighted Accuracy (UA) 77.54% [18]. Braunschweiler et al. used the CNN-RNN bidirectional LSTM (CNNRNNATT) model with attention mechanism for emotion recognition in their studies. This model classifies the emotional content of audio signals using

deep learning approach. For this purpose, it uses log-Mel filterbank features. Data augmentation was performed with Speed Perturbation and Volume Perturbation methods. In the experiments, they reached 77.4% UA value in IEMOCAP script dataset and 78.6% UA value in IEMOCAP improvised dataset [19]. Latif et al. developed a model (MTL-AUG) to improve emotion recognition performance by combining Multitask Learning (MTL) and data augmentation methods. Unlike other studies, they tried data augmentation in MTL scenarios to learn generalized representation to increase robustness in SER. They performed data augmentation with Speed Perturbation, Mixup and SpecAugment techniques and achieved 68.7% accuracy on IEMOCAP dataset [20]. Atmaja and Sasou investigated the effects of data augmentation on emotion recognition from speech in their work. They applied glottal source extraction, silence removal, impulse response application and noise addition methods for data augmentation. They performed classification with support vector machine (SVM) using voice features obtained using Wav2vec 2.0 and achieved an unweighted average recall (UAR) value of 76.39% on the IEMOCAP dataset [21]. Tu et al. In their study published in 2023, they used the Multi-Head Attention Convolutional Recurrent Neural Network model. In addition, LightGBM was used for feature selection and dimension reduction. Log-Mel Spectrograms are the main feature extracted from speech. A method called Mix-wav was used for data augmentation. This method creates new data samples by mixing different speech recordings. It was especially performed by combining sounds belonging to the same emotion category. They achieved 66.44% accuracy on the IEMOCAP dataset and 93.47% accuracy on the Chinese Hierarchical Speech Emotion Dataset of Broadcasting (CHSE-DB) dataset [22]. Dang et al. used Transformer-based network model for emotion recognition in their study. This model is designed to classify emotional states from speech using attention mechanisms and neural network architectures. Mel-spectrogram images of speech are used and EMix method is used as a new data augmentation method. This method creates new data samples by linear combinations of selective data pairs from emotional datasets. They reached 77.63% WA value in IEMOCAP dataset [23]. Qu et al. proposed a model called EmoAug for emotion recognition. This model is based on enriching speech emotions by transferring speech styles and changing prosody features. EmoAug uses HuBERT representations to capture semantic information. They developed a SER model with enriched data with this method. In their applications on the IEMOCAP dataset, they reached 72.66% WA and 73.75% UA values with the HuBERT Large + EmoAug method [24].

While reviewing studies on emotion recognition and data augmentation in text, research involving sentiment analysis was also considered.

In their study, Abonizio et al. investigated the effects of text data augmentation on sentiment analysis and examined the advantages and disadvantages of text data augmentation methods with different classification algorithms. Experiments were conducted on seven different emotion analysis datasets, including IEMOCAP. New text samples were generated using methods such as Easy Data Augmentation (EDA), Back-Translation (BT), PREDATOR, and BART for data augmentation. For classification purposes, models such as LSTM, CNN, BERT, and ERNIE were utilized. The LSTM model achieved the highest F1-score on the IEMOCAP dataset [25]. Imran et al. have investigated data augmentation techniques to improve emotion recognition performance in software engineering communications. For this purpose, Word Insertion (using BART), Word Substitution (using BART), Word Deletion, Sentence Shuffling methods were used in data augmentation. For emotion classification, software engineering specific emotion classification tools such as ESEM-E, EMTk and SEntiMoji were used. A new dataset consisting of GitHub comments and labelled according to different emotional categories was created. After data augmentation, an average of 9.3% improvement in the performance of the three classification tools was observed [26]. Gong et al. used a Transformer architecture-based model in their study to perform sentiment analysis on text data, aiming to increase the accuracy of this model with text augmentation techniques. The methods used to augment the text data include Easy Data Augmentation (EDA) techniques, back translation, and Word2vec based semantic similarity augmentation. They used AG News Corpus and Stanford Sentiment Treebank (SST) datasets. In the AG News dataset, 80.18% accuracy was obtained with 100 samples, and in the SST dataset, 78.21% accuracy was obtained with 100 samples [27]. In their 2023 study, Mohammad et al. proposed the Text Augmentation-Based Model for Emotion Recognition Using Transformers, named TA-MERT. This model captures both forward and backward contextual information with Transformer-based encoders, particularly through Bidirectional Encoder (BE) representations. It processes text-based inputs using Contextual Word Embeddings to encode their contextual information. The authors employed data augmentation techniques such as Back Translation, Easy Data Augmentation Methods, NLPAlbumentation Methods, and NLPAug Library methods. In experiments conducted on the MELD dataset, the model achieved a weighted F1 score of 62.60% and an accuracy of 64.36% [28]. Messaoudi et al. in their study, they performed the emotion recognition task by estimating the dimensions of valence (positivity or negativity), arousal (activation or intensity) and dominance (control or power) from text data. Long Short-Term Memory (LSTM) was used to estimate emotional states from text data. Global Vectors for Word Representation (GloVe) was used to encode semantic information for text data. BERT model was used to capture contextual information in the text. Experiments were conducted on the IEMOCAP dataset, and no data augmentation method was used in the study. The results obtained are Concordance Correlation Coefficient (CCC) scores for valence: 0.724, for arousal 0.423 and for

dominance 0.488, Average CCC Score: 0.545 [29]. Onan and Balbal have introduced data augmentation techniques customized for the Turkish language, in a study that examines data augmentation methods suitable for the structural features of different languages. They employed various methods such as Swap Synonym, Masked Language Model Based Suggestion, Spelling Error, Swap Named Entity, and Sentence Order Swap for data augmentation. They evaluated their proposed approaches on the TRSAv1 dataset and compared them with established data augmentation techniques [30].

When examining the literature on applications that perform bimodal emotion recognition using both speech and text, it becomes evident that there are different methodologies employed.

Priyasad et al. in their work published in 2020, they proposed an attention-based multimodal emotion recognition system that combines acoustic and textual data. In the proposed model, they used a SincNet filter layer to learn special filter banks tuned for emotion recognition from speech. After the sequence vector for text processing was passed through a common embedding layer, they used two parallel branches, one using Bi-RNNs with DCNNs and the other using only DCNNs, to increase the effectiveness of the learned features. They achieved a weighted accuracy of 80.51% in their model applied on the IEMOCAP dataset [31]. Sing et al. proposed a multimodal hierarchical DNN (Deep Neural Network) architecture for emotion recognition, integrating both speech and text. Their model incorporates 16 spectral features, including 13 MFCCs and 3 HSF spectral centers. To process textual data, word vector representations are calculated using ELMo. During textual data preprocessing, these word embeddings are combined with preprocessed audio features into a single vector. In their work, they achieved a 75.4% accuracy rate on the IEMOCAP dataset [32]. In their work, Padi et al. combined the scores of Resnet and BERT based models using a late fusion strategy to further improve the emotion recognition performance. They evaluated the effectiveness of their proposed model on the IEMOCAP dataset. After combining the speech and text-based emotion recognition systems, they achieved approximately 6% absolute improvement compared to the best performing single-mode system [33]. Zhang et al. presented AIA-Net, an Adaptive Interactive Attention Network developed for emotion recognition from text and speech. This network uses a multimodal approach with text as the primary modality and audio as the auxiliary modality and aims to improve emotion recognition through adaptive interaction between these modalities. In their study with IEMOCAP dataset, they obtained f1 score values of 79.28%, 88.09%, 91.66%, 89.63 for four different emotion classes (neutral, sadness, anger, happiness), respectively [34]. William and Zahra present an approach to emotion recognition by integrating textual and audio data in their work. They use BERT for text analysis. They process audio data by combining CNN (Convolutional Neural Network) and Bi-LSTM (Bidirectional Long Short-Term Memory) enhanced with Local Feature Learning Block (LFLB) for the audio model. The models' predictions are combined using a Stacking Ensemble Technique that increases accuracy by leveraging the strengths of both modalities. In their work, they achieve a combined model accuracy of 65.4% in the text model and 60.6% in the audio model, while achieving 75.181% in the bimodal application [35]. Li et al. presented a structure called Emoformer to identify emotional tendencies using multimodal information. This structure is designed to extract emotional vectors from three different modalities, namely text, audio and visual, and to create an emotion capsule by combining them with a sentence vector. Experiments were conducted on IEMOCAP and MELD datasets. They used BERT to extract text feature vectors, and OpenSMILE for audio feature extraction using IS13 ComParE config file with 6373 features. No data augmentation method was mentioned in the study. The Emoformer was tested on the IEMOCAP dataset, achieving an average F1 score of 69.49% in text-only mode and 71.39% in experiments using both text and audio mode [36]. Dutta and Ganapathy proposed ''Hierarchical cross-attention model (HCAM) for multimodal emotion recognition'' combining recurrent and co-attention neural networks. This model takes audio data processed using a learnable wav2vec technique and text data annotated using a BERT model. The model applied on IEMOCAP dataset for angry, happy (excited and happy categories merged), sad, neutral classes achieved 85.9% weighted F1 score [37]. In their work, Tafti and BabaAli presented a system that uses pre-trained transformer models to process both voice and text data. The aim is to minimize trainable parameters by fine-tuning these pre-trained models. In their study developed using the IEMOCAP dataset, they obtained unweighted average recall (UAR) of 78.42%, a weighted average recall (WAR) of 77.75% [38]. In their study, Guder et al. presented a dimensional approach to recognizing emotions from speech and text. Instead of classifying emotions, they aimed to recognize them by representing them through dimensions such as valence, arousal, and dominance. For acoustic features, they utilized the pre-trained VGGish model to create audio embeddings. The text transcription of the speech was processed using the WhisperX model, followed by sentence embedding with the MiniLM L3 model. These two sets of features were combined, and the emotional dimensions (valence, arousal, and dominance) were predicted using an LSTM (Long Short-Term Memory) based model. In their work using the IEMOCAP dataset, they achieved a Concordance Correlation Coefficient of 0.5915 for arousal, 0.1431 for valence, and 0.5899 for dominance [39].

III. PROPOSED METHODS

In the proposed methodology, a bimodal emotion recognition application is developed that processes and analyzes both speech and text data to identify emotional states. This approach leverages advanced data augmentation techniques to improve the robustness and accuracy of the emotion recognition process. A schematic representation of the research methodology is provided in Figure 1.

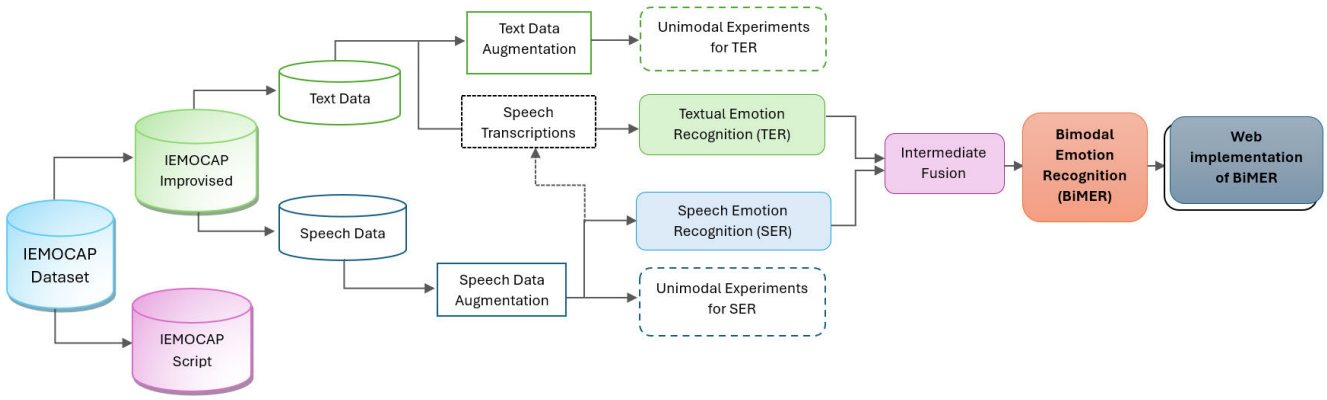


FIGURE 1. Schematic presentation of research.

In the study, IEMOCAP (Interactive Emotional Dyadic Motion Capture) [17] dataset, which is widely used in multimodal emotion recognition processes, was used as a dataset. One of the problems addressed in the study was data augmentation in emotion recognition for datasets with limited data and data imbalance between classes. In order to eliminate the imbalance in emotion formations in the IEMOCAP improvised dataset, which is close to natural speech data, balanced datasets were created by using data augmentation methods in both speech recordings and textual transformations of speeches. The balanced dataset with data augmentation methods applied to speech recordings was used in the Speech Emotion Recognition (SER) application. For the SER model, Mel spectrogram features were extracted from speech data, treating SER as an image classification problem using these spectrogram representations. For this purpose, three different models have been developed for SER.

The first model is the ResNet-CNN model developed with 2D Convolutional (2DCNN) layers and Residual Blocks. The second model is the ResNet CRNN+AT model developed by adding GRU (Gate Recurrent Unit), Attention Mechanism and Transformer Encoder Layer to the first model. The third model is the ResNet50-CRNN+AT model, where the ResNet50 architecture was initially used to obtain features from mel spectrogram images. As a result of the experiments conducted on these architectures, the highest accuracy was achieved with the ResNet50-CRNN+AT model and this model was used for the SER modality in BiMER. The balanced dataset, created using data augmentation methods applied to text data derived from speech transcriptions, was employed in experiments conducted with the Textual Emotion Recognition (TER) application, which operates as a single-modal system. In this application, BERT (Bidirectional Encoder Representations from Transformers) model was used to ensure deep and detailed learning of text features and emotion classification was performed on the text. Thus, the performance increase provided by data augmentation methods in both SER and TER applications was observed. The features extracted from these two singular modalities have been combined using the intermediate fusion method to develop a more comprehensive emotion recognition system, enhancing accuracy through the Bimodal Emotion Recognition (BiMER) system. The BiMER application has achieved higher accuracy compared to single-modal emotion recognition applications conducted on speech and text alone. In the final phase of the study, the BiMER system has been made available for real-time use through a web server created with the Flask framework

A. DATASET

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset was used in the application carried out within the scope of the study. IEMOCAP is a multimodal dataset proposed in 2008, where recordings are created by more than one actor. It consists of five different sessions containing approximately 12 hours of audiovisual data, including video, audio, facial motion capture, and text transcriptions. It consists of dual dialogues in which the actors perform specially selected improvisations or written scenarios to elicit emotional expressions. IEMOCAP includes nine emotion classes (anger, happiness, excitement, sadness, disappointment, fear, surprise, other, and neutral status). There are more than three raters for each emotion expression. Only if more than half of the raters have the same opinion, the expression is marked with the relevant label, otherwise it is labeled as 'other' [17].

With these features and robustness, IEMOCAP has become one of the frequently preferred datasets in emotion recognition and multimodal emotion recognition studies over audio [35]. However, the interpretation and labelling of the emotion expression by the raters caused a large imbalance in the emotion formations in the dataset. For example, since excitement and happiness have a certain degree of similarity and there are very few happy expressions, researchers sometimes use emotions labelled with excitement instead of happiness in their studies or increase the amount of data by combining excitement and happiness [40]. On the other hand, the improvised recordings that are close to the natural

speech data included in the IEMOCAP dataset are obtained by actors acting out scenarios designed to reflect real emotional experiences. Since this type of data is closer to real-life conditions and natural human interactions, it contains the real emotional reactions of the actors. In this respect, the IEMOCAP improvised dataset can be distinguished from datasets such as Emo-DB (Berlin Database of Emotional Speech) [41], RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) [42], SAVEE (Surrey Audio-Visual Expressed Emotion) [43], which are recorded in more controlled environments and consist of repetitions of certain sentences into theatrical expressions of different emotions.

Due to its features and its extensive use in many studies, and with the goal of developing a real-time web application by the end of this study, the IEMOCAP dataset was used with four improvised emotion classes (neutral, angry, sad, happiness). Table 1 lists the emotional distributions for a total of 2280 improvised recordings in the dataset.

As seen in this table, there is an imbalance among the classes. For example, while there are 1099 recordings labelled as neutral, the numbers of recordings for the angry and happy classes are comparatively lower. The data augmentation methods used to address this imbalance among the classes are explained in the relevant sections later in the study.

TABLE 1. Number of recordings in the Neutral, Angry, Sad, and Happy emotion classes in the IEMOCAP dataset.

Emotion Class	Improvised conversations	Scripted conversations	Total
Neutral	1099	609	1708
Angry	289	814	933
Sad	608	476	1084
Happy	284	311	595
Total	**2280**	**2210**	**4490**

B. SPEECH EMOTION RECOGNITION (SER)

Emotion recognition is not a simple classification problem. The expression of emotions may vary from culture to culture, and even from person to person within the same culture. However, the same speech may contain different emotions depending on the context of the speech. Emotion analysis over speech is basically a system that identifies the emotional state of a person from his/her own voice. Emotion recognition over speech is actually a pattern recognition application, and the stages found in a pattern recognition system can also be applied to emotion recognition systems over speech. In this section, the features used in the speech mode of the developed BiMER application, the data augmentation methods used, and the models developed for SER are presented.

1) MEL SPECTROGRAMS

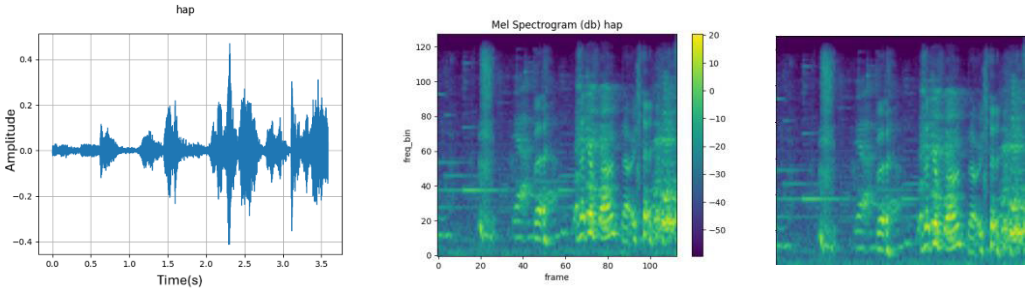
One of the important stages in speech emotion recognition applications is to extract appropriate features that characterize emotions, and many methods have been used for this purpose [44]. The selection of the right features affects the classification performance and will increase the emotion recognition rate. Different features have been used in the systems developed to date, but there is no distinctive and definitively accepted set of features [45].

In this study, mel spectrograms generated from speech recordings in the IEMOCAP dataset were used. Because the Mel frequency scale provides a rough model of human frequency perception, it is widely used to represent audio signals and has extensive applications in the field of audio analysis [46]. Mel spectrograms use the Mel scale, a transformation technique that converts audio data to the frequency domain and mimics the way the human ear perceives frequencies. It has also been considered as important features for emotion recognition from audio [47]. Creating a spectrogram in the Mel scale involves creating a spectrogram and performing the transformation in the Mel scale. There are various versions of the Mel scale, and in this study, the HTK (Hidden Markov Toolkit) version was used. The number of Mel filter banks was determined as 128, the FFT (Fast Fourier Transform) size was determined as 1024, and the jump length between STFT (Short-Time Fourier Transform) windows was determined as 512. Within the scope of the study, the problem was transformed into an image processing problem by obtaining Mel spectrogram images of the raw audio files recorded in an improvised manner in the IEMOCAP improvised dataset, as specified in the Improvised Conversations column of Table 1.

Audio recordings were sampled as 16 Khz and silences at the beginning and end of the speech recordings were removed. Figure 2 shows the raw waveform of a recording labelled as happy as an example and the created mel spectrogram image. The "viridis" colour scale was used while creating the mel spectrograms and the recorded mel spectrogram image dimensions are 496px wide and 369px high.

2) DATA AUGMENTATION FOR SER

Data augmentation is a preprocessing technique used to increase both the quantity and diversity of datasets. This process is essential for addressing issues such as data scarcity or imbalance and for improving model performance. It is particularly useful in situations where the number of sample data in a class is limited or missing. Data augmentation is a commonly used method in applications developed with deep learning techniques, such as image and audio processing [48]. Different methods have been used for data augmentation in audio classification and speech recognition studies. Some of these methods include time stretching, pitch shifting and noise addition on the raw audio signal, and Generative Adversarial Networks (GAN) [49]. In this study, to increase the number of samples for the emotions with low



(a) (b) (c)

FIGURE 2. Raw audio waveform of the emotion happy (a), mel spectrogram images of an example audio recording (b), frameless mel spectrogram image created for the dataset (c).

representation—angry, sad, and happy—in the IEMOCAP improvised dataset, new recordings were obtained by applying *speed change, room impulse response, background noise, pitch shift, and specAugment* processes to the raw audio files. While performing these operations, the functions under the torcaudio library were used.

Speed Change: Speed change is done by changing the playback speed of the audio signal. This process changes the sampling rate to shorten or lengthen the duration of the sound, but the pitch is not affected. This method is often used to increase the tolerance of speech recognition systems to different speech rates. In this study, new recordings were obtained by changing the speeds of the original speech recordings by 0.75 and 1.25.

Room Impulse Response: Room Impulse Response (RIR) is achieved by adding artificial echo to audio signals. This makes audio recordings suitable for different acoustic environments, particularly helping to increase the accuracy of remote speech recognition systems.

Background Noise: Background noise is achieved by adding various background sounds (e.g., traffic, crowds, nature sounds) to audio signals. Background noise is used to make automatic speech recognition systems work better in noisy environments.

Pitch Shift: Pitch shifting is done by changing the fundamental frequency of the audio signal. This method changes the characteristics of the voice by making the speaker's voice pitch higher or lower. Pitch shifting is used to improve the adaptation of speech recognition and speech analysis systems to different pitch characteristics.

Figure 3 shows the waveform and mel spectrogram image of the newly created voice after background noise is added to a sample speech recording of the angry (ang) class.

SpecAugment: In 2019, Park et al. proposed a simple data augmentation method named SpecAugment (Spectrogram Augmentation) that can be directly applied to the feature input of a neural network [50]. This method is a modern data augmentation technique used in various fields including speech translation [51], speaker verification [52], speech emotion recognition [53], and brain-computer interface (BCI) applications developed using Electroencephalography (EEG) signals [54].

It has been observed that this method increases the success rate when applied to mel spectrogram images in the application of emotion recognition from speech [53], [55]. SpecAugment uses three main techniques to augment speech data: time warping, frequency masking and time masking. These processes help the model to become more robust against different sound and time variations by changing the randomly selected spectrogram regions. In this study, frequency masking was applied. In frequency masking, a random region with a certain bandwidth is selected along the frequency axis of the spectrogram and this region is replaced with the masking value (usually 0) at all time steps. The frequency components of the sound usually play an important role in capturing emotional tones and intonations. For example, happiness is usually associated with high tones, while anger is associated with sharp tones. Therefore, the frequency masking method may be more effective in helping the model learn these types of features. During augmentation, one of the masking values (15, 20, or 25) was randomly selected and applied to the spectrogram image to enhance feature diversity.

Figure 4 shows the new mel spectrogram images resulting from the data augmentation methods applied to a mel spectrogram image of the angry class. The reason for using the improvised recordings of the IEMOCAP dataset is that, as mentioned in the previous sections, it is more suitable for the natural flow of daily speech. These operations applied to the audio files for data augmentation were also chosen because there are possible effects that can be included in speech applications in daily life.

3) DATASET PARTITIONING AND UTILIZATION OF AUGMENTED DATA

The dataset is divided into two as train set and validation set to be used for model training and validation. In the first stage, a validation set was created from the original records with 120 records in each emotion class. After this stage, augmented records were added to the classes with less data, and under sampling was done for the neutral class with more records, and a train set was created with 800 records in each

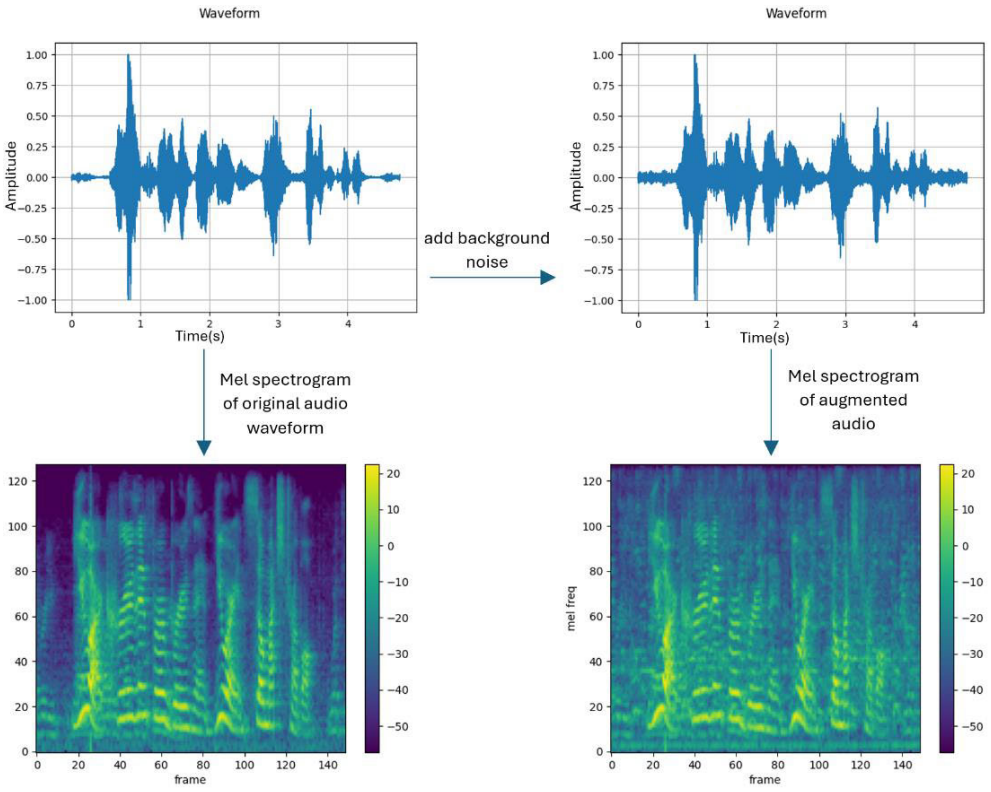


FIGURE 3. New waveform and Mel spectrogram images after adding background noise to a sample audio recording from the angry class.

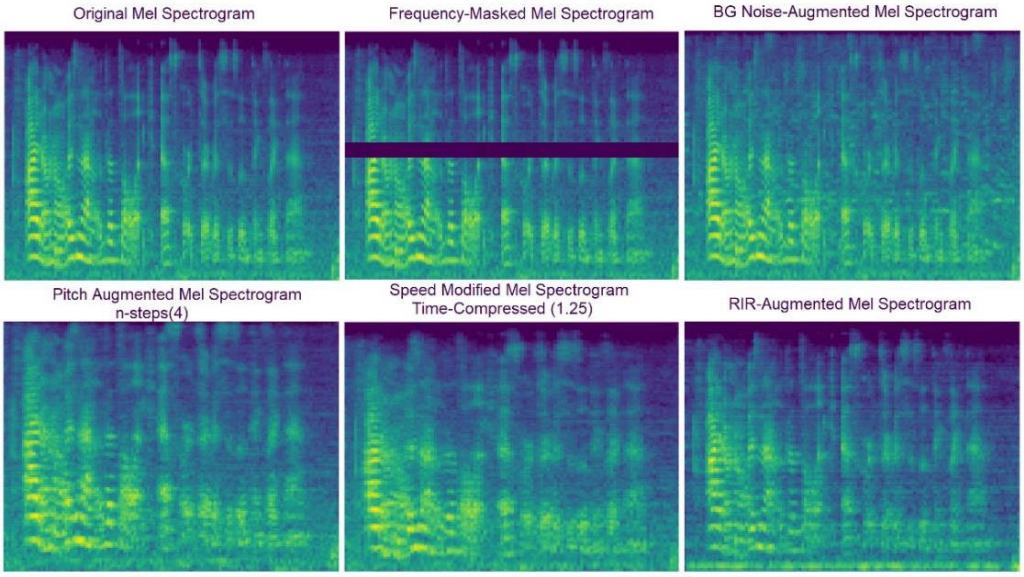


FIGURE 4. Original mel spectrogram image of a sample record of the Angry class and augmented mel spectrograms.

class. The final record numbers for each emotion class are given in Table 2. When adding to the train set from the five different augmented methods used, the distribution is made in such a way that an equal number of records are added from each method. Adding augmented data with various techniques helps the model learn generalized representations compared to adding augmented data with a single method and provides better model performance [20]. The train set

was created by randomly selecting records in each application while being used in the training process of the model.

TABLE 2. Number of records generated through data Augmentation and the total number of records used in the balanced dataset.

Emotion Classes	Number of data in Original Dataset (improvised)	Validation Set	Train Set with Augmented Data
Angry	289	120	800
Sad	608	120	800
Happy	284	120	800
Neutral	1099	120	800
Total	**2280**	**480**	**3200**

4) DEEP LEARNING MODELS FOR SER

In this section, we detail three models developed for emotion recognition from speech. Our first model, the ResNet-CNN, is a 2D Convolutional Neural Network that utilizes residual blocks. The second model, the ResNet CRNN+AT, extends the first model by incorporating a Gate Recurrent Unit (GRU), an Attention Mechanism, and a Transformer Encoder Layer. The third model, ResNET50-CRNN+AT, performs feature extraction based on ResNet50, and like the second model, it also includes a 2D CNN, GRU, Attention Mechanism, and Transformer Layer for emotion classification. The results obtained from these models in the SER application are presented in the 'Experimental Results of SER' section.

a: ResNET-CNN MODEL

The fundamental component of this model is the residual learning blocks, introduced in ResNet, a variant of CNN architecture. A residual block transfers part of its input directly to subsequent layers, a mechanism known as a shortcut connection. This structure facilitates efficient learning in deeper network layers and significantly mitigates the problem of vanishing gradients. Figure 5 illustrates a simple Residual Block architecture. In this setup, the identity function directly transmits the input X to the output of the residual block unchanged, acting as a shortcut that bypasses intermediate connections. The residual function path modifies X to produce the residual function F(x), containing two weighted layers and a ReLU activation function in the depicted simple case. The initial output of this path, F(x), represents the change or deviation from the identity. The final output of the block, F(x)+x, indicates that the output is the sum of the original input x (via the shortcut) and the learned residual F(x). In this model, H(x) specifies the desired underlying mapping. By defining F(x) = H(x) - x, the residual block aims to learn the difference H(x) - x, rather than H(x) directly, thereby reformulating the original mapping as F(x) + x. This methodology underpins the basic operating principle of the residual block [56].

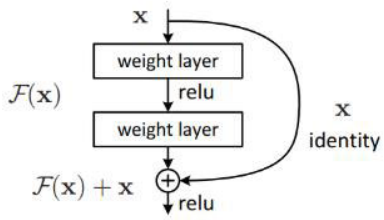


FIGURE 5. Basic residual block architecture [56].

Figure 6 shows the residual block architecture of the model, which consists of two consecutive 2D CNN layers developed for SER. Here, X represents the tensor entering the residual block. The input tensor undergoes a 2D CNN process specified by channel number and stride value. Following this layer, batch normalization is applied, followed by the Leaky ReLU activation function. The feature map from the first layer then passes through a second 2D convolution (Conv 2D) layer with the same channel number. After this second layer, batch normalization and Leaky ReLU activation are again applied. If the stride value differs from 1 or if the input and output channel numbers differ, the input tensor is processed through a 1×1 convolution and batch normalization to adjust its dimensions appropriately. This step establishes the shortcut connection to be added to the block's output. In the addition phase, the output of the second convolution layer and the output of the shortcut connection are added. This addition process allows the model to learn only the residual, providing a solution to the vanishing gradient problem that is often encountered in deep networks. The output from this addition is then passed through the Leaky ReLU activation function to produce the final output.

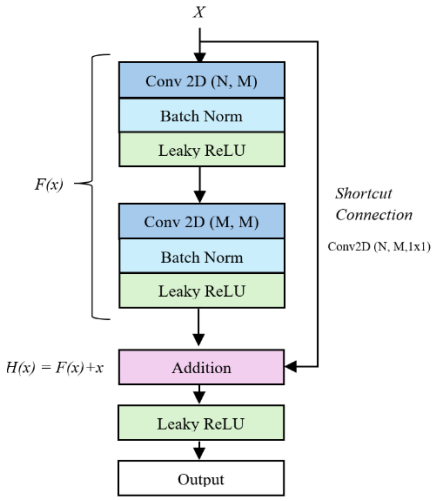


FIGURE 6. Residual block architecture with 2 consecutive 2D CNN layers for the SER model.

The diagram of the first model ResNet-CNN developed for speech emotion recognition application is shown in Figure 7. The deep learning model consists of six Residual Block

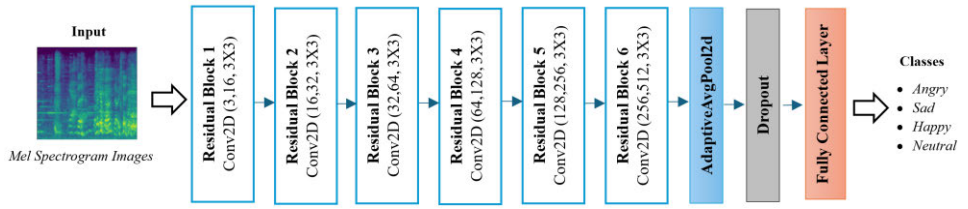


FIGURE 7. Diagram of ResNet-CNN model.

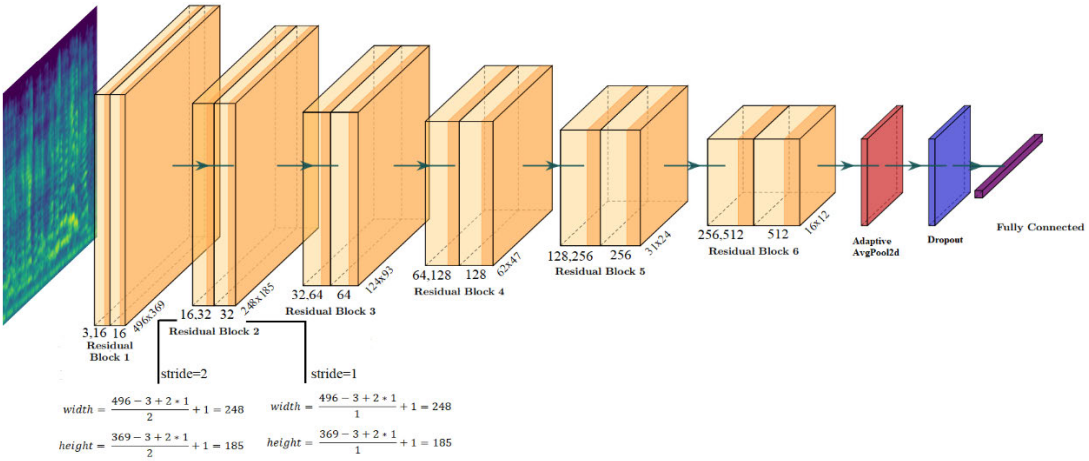


FIGURE 8. Layers of ResNet-CNN model.

structures, each containing a total of 12 convolutional layers. The model takes 496 × 369 Mel spectrogram images as input. Each residual block contains two convolution layers that transform the input data into feature maps. Each convolution layer has an input and output channel number and uses a 3 × 3 Kernel. In the first convolution layer in the blocks, stride=2 is used to halve the size of the input feature map. This process effectively reduces the input size of the model and allows it to learn higher level abstractions in deeper network layers. In the second convolution layer, stride=1 is defined to preserve the size of the feature map and enable processing of more detailed features. Batch Normalization and Leaky ReLU activation function are applied after each convolution layer. Then, the Pooling Layer reduces the feature maps to a fixed-size vector, which is provided as input to the fully connected layer.

The feature vectors are processed in the final layer, the Fully Connected Layer, which is used for classification. This layer is designed to predict emotional states and produces a score for each emotion class at the output.

The input (N) and output (M) channel numbers of the 2DCNN model created with residual blocks are given in detail for each block in Figure 8. For example, in the values sent to the first residual block, it is defined as N=3 to M=16, kernel size = 3 × 3, padding = 1. It performs the input size reduction process with the stride = 2 value in the first 2DCNN (Conv 2D) layer in the residual block. In the second 2DCNN layer in the block, it is defined as stride = 1 and will be transferred to the second residual block with N=16 and M=16 channels. The padding value is used as padding = 1 in all convolution layers. In the first 2DCNN layer in the second residual block, the convolution layers are activated as N=16 to M=32. In the last sixth block, a 512-dimensional output is processed in the Adaptive Average Pooling layer (1 × 1). AvgPool2d reduces the feature map size of each channel to 1 × 1 independently and produces an output of 512 × 1×1. Then, the tensor is reshaped to enter the fully connected layer with the flattening process (view). Dropout is applied at a rate of 0.5.

Finally, the fully connected layer calculates the final scores for classification.

b: ResNET-CNN+AT MODEL

The second model created for speech emotion recognition application is the ResNet CRNN+AT model, which was developed by adding Attention Mechanism and Transformer Encoder Layer along with GRU, a type of Recurrent Neural Network (RNN), to the first model. The working principle of the residual blocks, which are the main components of the model, is explained in the first model. The architectural diagram of the second model is shown in Figure 9. In order to rearrange the feature maps obtained from the residual network used in the first model and make them processable with GRU, reshape and permute operations are applied as seen in Figure 10. Permute operation reorders the dimensions

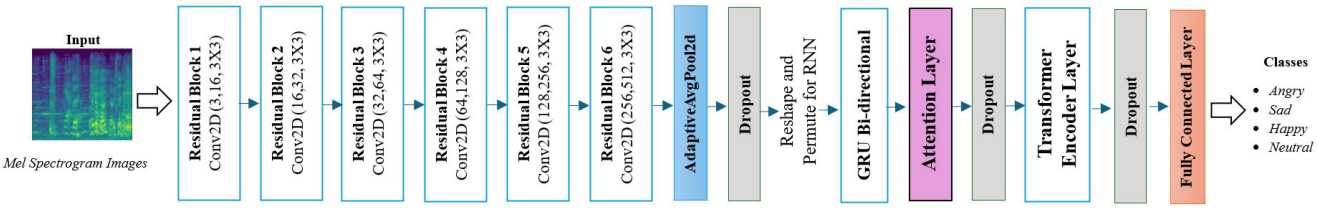


FIGURE 9. Diagram of ResNet CRNN+AT model.

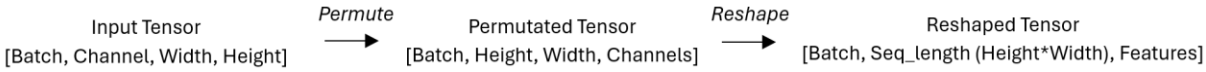


FIGURE 10. Permute and reshape to tensor.

of the tensor. Reshape operation reshapes the dimensions of the tensor so that the amount of data and the total number of elements remain the same. It converts the multidimensional feature map (batch_size, channels, height, width) into a single-dimensional sequence (batch_size, sequence_length, feature_size) and is used to put the feature maps into a format suitable for the GRU or Transformer layer.

Gate Recurrent Unit (GRU) is a type of Recurrent Neural Network developed as an alternative to Long Short-Term Memory (LSTM) units proposed by Chung et al in 2014. Similar to LSTM, it is designed to carry information over time but has a simpler structure. By merging the 'forget' and 'input' gates of LSTM into a single 'update' gate, GRU reduces both model complexity and computational cost. Unlike LSTM, GRU uses only two gates: update gate and reset gate. These gates decide which information the network will carry from the past to the future [57].

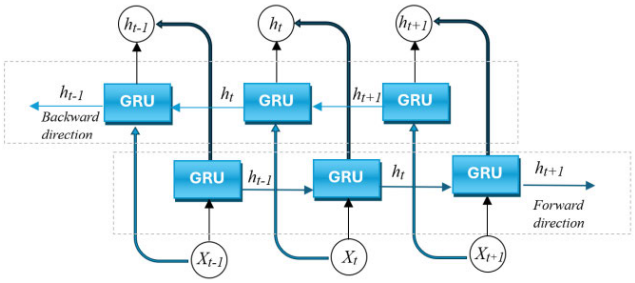


FIGURE 11. The architecture of bidirectional GRU.

Bidirectional GRU (BiGRU) provides a better understanding of the context by processing incoming data both from the past to the future and from the future to the past. This is especially important in sequential data such as time series data or text because it evaluates a broader context by using information from both directions. Bidirectional GRU contains two separate GRU layers, as seen in Figure 11, one of these layers processes data in the forward direction and the other in the backward direction. The resulting outputs are usually concatenated or summed.

The BiGRU layer in the model takes 512-dimensional features from the residual network as input and generates 256-dimensional hidden layers bidirectionally. This process allows the model to learn the connections between past and future information. By bidirectionally processing the time series both forward and backward, it captures the context from past and future states to understand the full emotional context of the conversation.

Following the GRU layer, a soft attention mechanism dynamically assigns importance to different features. For this purpose, the attention layer takes the 512-dimensional input and calculates an attention score to model the dependencies between the features. This mechanism allows the model to focus on the most important features.

The Transformer Encoder layer further improves the model's ability to process long-range dependencies in speech data. It takes 512-dimensional features marked by the attention mechanism and applies self-attention mechanisms using 8 heads. This layer captures long-range connections between features, allowing for a better understanding of emotional nuances. Dropout layers aim to increase the generalization ability of the model by skipping random units during training to prevent overfitting. The dropout rate is applied as 0.3 in dropout layers. The fully connected layer is the last layer of the model and converts 512-dimensional features into a 4-class output (Angry, Sad, Happy, Neutral). This layer enables classification using the softmax activation function.

c: ResNET50-CNN+AT MODEL

The latest model developed for emotion recognition from speech has the ResNet50-CRNN+AT architecture. At the beginning of the model, the ResNet-50 layer was used to extract features from the Mel spectrogram images it received as input. ResNet-50, as its name suggests, is a ResNet model consisting of 50 layers, and the architectural diagram of ResNet-50 is shown in Figure 12. ResNet-50 uses residual blocks to cope with the vanishing gradient problem encountered in deep networks. The ''identify'' in the diagram represents the identification blocks (shortcut connection)

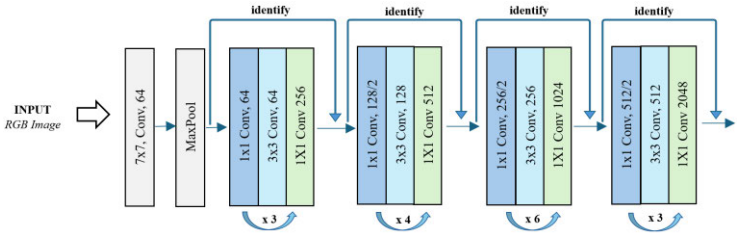


FIGURE 12. ResNet-50 architecture.

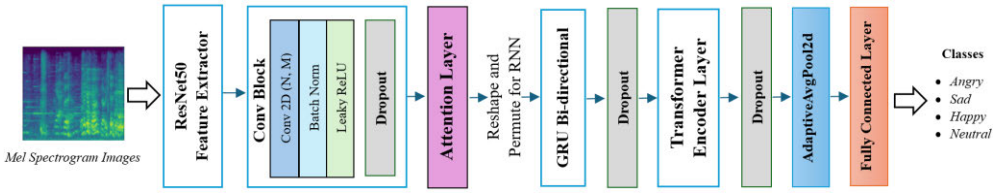


FIGURE 13. Diagram of ResNet50-CRNN+AT model.

used to indicate the use of previous layers in subsequent layers. There are 64 filters with a kernel size of 7×7 in the first layer of the model, followed by a maxpooling layer of size 3×3. The subsequent residual blocks consist of blocks containing 1×1, 3×3 and again 1×1 convolutions. The first layer group consists of 3 consecutive identical blocks, the second layer group consists of 4 identical blocks, the third layer block consists of 6 identical blocks, and finally the fourth layer group consists of 3 identical blocks [58]. This structural arrangement increases the efficiency of the network while reducing the computational load. Since this model has more layers and parameters, the feature maps of the mel spectrogram images were first obtained with the pretrained ResNet-50 model, as seen in Figure 13.

In the ResNet50 model, the size of the feature maps coming out of the last layers of the deep learning network usually contains 2048 channels. Then, by using an additional conv block, the features are reduced from 2048 channels to 256 channels. This layer contains Conv2D, batch normalization and ReLU activation functions. It processes the feature maps it receives as input and produces lower-dimensional feature maps as output.

Then, the attention layer takes the 256-dimensional input from the Conv Block layer, processes the feature maps, and calculates an attention score to model the dependencies between the features. For this purpose, a type of attention mechanism, channel-wise spatial attention, is used. This mechanism calculates attention weights in the spatial plane (height and width) by averaging over the feature maps of each channel in the input (i.e., mean(dim=1)). This attention map is normalized using the Softmax function, resulting in an attention distribution covering the entire height and width. The original input tensor x is multiplied by the element-wise attention weights, effectively scaling the contribution of each channel according to its importance. After the Attention layer, reshape and permute operations are applied to rearrange the feature maps appropriately and make them processable with GRU. The information from the GRU layer is passed to a Transformer encoder layer. Finally, the outputs from the Transformer layer are reduced in size by average pooling and sent to the fully connected layer. This layer performs the final classification and produces predictions for four emotion classes (angry, sad, happy, neutral).

C. TEXTUAL EMOTION RECOGNITION (TER)

Emotion recognition from text is the process of automatically identifying and classifying emotional content in texts, which lies at the intersection of natural language processing and artificial intelligence. Research in this area aims to enable computers to understand and interpret subtle emotional nuances in human-written communication. Different methods are used in emotion recognition studies from text. Following the classical approaches of keyword-based and rule-based, machine learning and especially deep learning applications have gained popularity in the field of emotion recognition from text in the last decade. These methods can be listed as Word-Based Approaches, Rule-Based Approaches, Machine Learning-Based Approaches, Deep Learning-Based Approaches, Hybrid Applications [59].

In this section, the BERT model used in the textual emotion recognition modality in the Bimodal emotion recognition system is explained. This model was also used for emotion recognition from text in a single-mode manner in the experiments. Textual data augmentation methods used to eliminate the imbalance between the classes in the data set in the single-mode text emotion recognition application are also included in this section.

1) BERT MODEL FOR TER

In the study's text emotion recognition modality, Bidirectional Encoder Representations from Transformers (BERT) architecture was used. BERT is a transformer model created

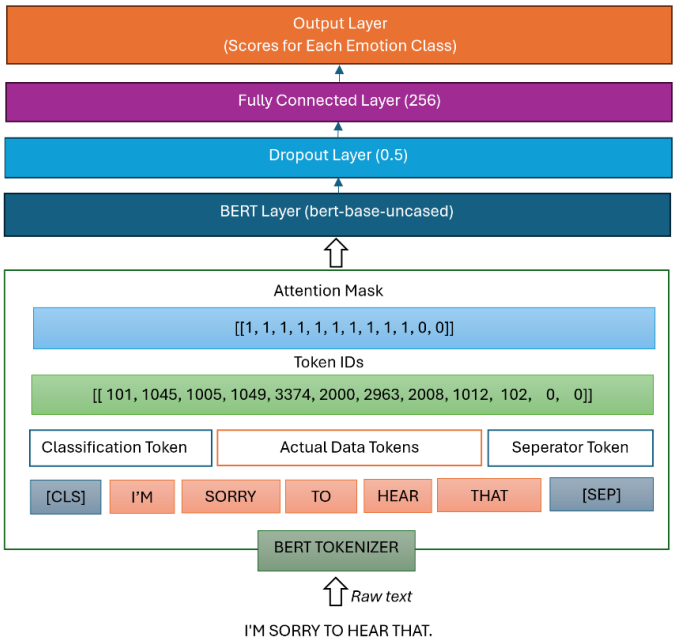


FIGURE 14. BERT Model for Textual Emotion Recognition.

and published by Jacob Devlin et al. in 2018. The developed framework has two steps: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data from different tasks. For fine-tuning, the BERT model is first initialized with pre-trained parameters and all parameters are fine-tuned using labelled data from downstream tasks [60]. BERT is based on Transformer [61], a network that can model long contextual information and produce word embeddings conditioned on the phrase the word is in. The BERT model has allowed unsupervised pre-training, which can be shown as an integral part of many language understanding systems with recent studies, to successfully handle natural language processing tasks with deep bidirectional architectures [62]. With the successful results it has achieved, it has recently become a frequently used model in scientific studies [60].

The developed model aims to classify the given text into emotional classes by taking advantage of BERT's powerful language processing capabilities. The model's diagram is shown in Figure 14 and, for example, it takes a sentence like "I'M SORRY TO HEAR THAT." as input. Using BERT Tokenizer, the text is converted into Token IDs and related attention masks. In this process, the text is enriched with special tokens such as [CLS] and [SEP] and filled with [PAD] tokens if necessary to reach a fixed length (for example, a maximum of 512 tokens). [CLS] is the starting token used for classification; [SEP] is the separator token indicating the end of the sentence. In the next stage, each token is matched with a numeric ID defined as Token ID, and the Attention mask determines which tokens the model should focus on. Dropout is applied in the dropout layer to prevent overfitting of the model. The output obtained after dropout is inserted into a fully connected layer that is sized to perform emotion classification. In the final stage, scores are generated for each emotion class and these scores are used to determine which emotion best fits the text. Within the scope of the study, the results of the unimodal experiments conducted with emotion recognition from text only are explained in the "Experimental Results of TER" section.

2) DATA AUGMENTATION FOR TER

Textual data augmentation allows artificially increasing the variety and quantity of existing text data without collecting new data. This technique is especially important in natural language processing and machine learning applications, especially when training data is scarce or unbalanced. There are different methods to perform text data augmentation. Some of these methods can be listed as follows [63], [64], [65].

Character level methods: Methods that can be used at the character level are performed by adding, deleting or changing random characters in the text. (Ex: cheese → chese)

Random addition, Random replacement, Random deletion, Synonym Replacement: Wei et al. applied the textual data augmentation methods they call Easy Data Augmentation (EDA) to 50% of the training set and achieved the same accuracy rate as the training set with all the data. These methods can be listed as operations to be performed at the word level within the sentence.

Rule-based methods: In these methods, new sentences are produced based on certain grammatical rules and dictionary knowledge, such as creating new forms by changing the inflectional suffixes of words.

Back-translation: In the back translation method, the original text is first translated into another language and then translated back into the original language, allowing new

Algorithm 1: Implementing Text Augmentation methods

```
Input: DataFrame D
Output: Modified DataFrame D
Begin
    for each row in DataFrame D do
        original_text ← row["text"]
        if not isValidText(original_text) then
                continue
        end if
        new_text ← original_text
        similarity ← 0
        similarity_threshold ← 0.9
        max_attempts ← 10
        attempts ← 0
        while (similarity < similarity_threshold) and (similarity ≠ 1.0) do
            attempts ← attempts + 1
            if attempts > max_attempts then
                print("After ", max_attempts, " steps, the original text '", original_text, "' was preserved.")
                break
            end if
            rnd ← random()
            if rnd < 0.5 then
                new_text ← back_translation(original_text)
            else
                new_text ← synonym_replacement_bert(original_text)
            end if
            embeddings ← sentence_transformers_similarity(original_text, new_text)
            similarity ← cosine_similarity(embeddings[0], embeddings[1])
        end while
    end for
End
```

isValidText ()
short_words ? {'PLEASE', 'GOD', 'WHAT', '{BREATHING}', 'YEAH',... }
length(word_list==1) ?

FIGURE 15. Implementation of text augmentation methods.

variations to be produced without any change in the meaning of the sentence.

Model based approaches: In model-based approaches, the data augmentation process is driven by the machine learning models themselves. Pre-trained models such as BERT or specially trained models are used to make changes to the data while preserving the semantic meaning. Contextual Word Replacement or Contextual Word Insertion techniques can be listed in this category. Using pre-trained models such as BERT [60] and GPT-3 [73], it is ensured that the context of the words in a sentence is understood, and words appropriate to the context are added instead of these words or replaced with words appropriate to the context. This method will allow the sentence to create variety without disrupting its general meaning.

Generative models: Generative models use GANs to generate synthetic texts. Models that specialize in language processing tasks, such as GPT-3 and T5 [74], can generate entirely new texts or rephrase sentences using paraphrasing.

In emotion recognition systems, it is important to preserve the semantic integrity of the texts while augmenting textual data on original texts. Augmented texts should have a structure that will not disrupt the existing emotion transfer. In this context, we performed text augmentation using two methods in the developed study. The first of these methods is BERT Synonym Replacement as a Model-based approach method. BERT analyses the given sentence at the word level and determines the meaning of each word in context. As a result of the analysis, possible synonyms are determined for the word to be changed.

This process is carried out thanks to the word vectors and grammar knowledge that BERT has previously learned. One of the determined synonyms is replaced with the word in the original sentence and a new sentence is created. As the second method, we used the backtranslation method. For back translation, the original English text was translated into English→French→German→Turkish→English languages, respectively, using Google Translate API, and a different variation of the text was obtained.

Examples of sentences using these two methods are provided in Table 3.

With its multimodal structure, the IEMOCAP dataset includes not only the speech recordings but also the transcriptions of these speech recordings. In this context, the transcriptions of the sentences belonging to the speech recordings for the 5 Sessions included in the dataset are also included in the dataset. As we mentioned in the previous sections, we used only the records created as improvised in

TABLE 3. Sample sentences demonstrating text augmentation techniques.

BERT Synonym Replacement *Text Augmented Method*	
Original Text	DON'T TALK TO ME LIKE I'M A CHILD.
New Text	DO N'T TALK TO ME if I 'M A CHILD .
Cosine Similarity	0.994
Back Translation *Text Augmented Method*	
Original Text	YOU'RE KIDDING ME THIS IS A JOKE.
New Text	YOU MAKE FUN OF ME, IT'S A JOKE.
Cosine Similarity	0.994

the IEMOCAP dataset within the scope of the study. However, there is a situation where some transcriptions of the improvised records are repeated within the same classes. For example, there are two records where the sentence "WHAT? ARE YOU KIDDING ME?" is tagged with the emotion of angry. As another example, there are four records where the sentence "I DON'T KNOW." is tagged with the emotion of sad. In this way, after the preprocessing of the dataset in such a way that only one of the repeated sentences in each emotion class is used, the resulting emotion numbers are as indicated in Table 4.

After this preprocessing stage, data augmentation was performed using BERT Synonym Replacement and Back Translation methods for the emotion classes ang, hap, sad, initially had fewer sentences. Under sampling was performed for the neutral class and the sentence numbers were balanced across the classes. In the final case, a balanced dataset was obtained with 625 sentences in each class.

TABLE 4. Class distributions of IEMOCAP dataset used for TER application.

Emotion Classes	Numbers of sentences	After removing repeated sentences	After applied Text Augmentation
Angry	289	286	625
Happy	284	273	625
Sad	608	545	625
Neutral	1099	1008	625
Total	**2280**	**2112**	**2500**

D. BIMODAL EMOTION RECOGNITION ON SPEECH AND TEXT

Multimodal fusion methods allow for the creation of a more comprehensive model by combining information from different data types (e.g. text, audio, images). They are used especially in the fields of artificial intelligence and machine learning to develop more accurate and effective models by integrating various data types. Combining different data types will allow for more comprehensive and in-depth analyses. With these analyses, the multimodal approach will enable the emergence of more accurate emotion recognition systems for emotion recognition systems. For example, in the IEMOCAP dataset, the sentence "THANK YOU." in the speech recordings is labelled in both sad, neutral, and happy emotion classes. Here, it is seen that the meaning that adds emotion to the text is the features of the sound formed in the speech recordings. One of the difficulties of multimodal machine learning is in the methods for combining different methods. Different fusion strategies are used for this purpose.

Multimodal fusion strategies can be generally examined in three categories as early fusion, intermediate fusion, and late fusion [66], [67]. Intermediate Fusion; is a method of combining information from different data sources at the intermediate level (intermediate layers or feature level). Intermediate fusion combines the advantages of early fusion and late fusion strategies. It allows the production of a more meaningful new representation from separate representations by combining the features that distinguish each data type. While allowing each mode to preserve its own features, it benefits from the combination of these features and allows each modality to be processed on its own and combined at certain levels [66]. The general structure of intermediate fusion is given in Figure 16.

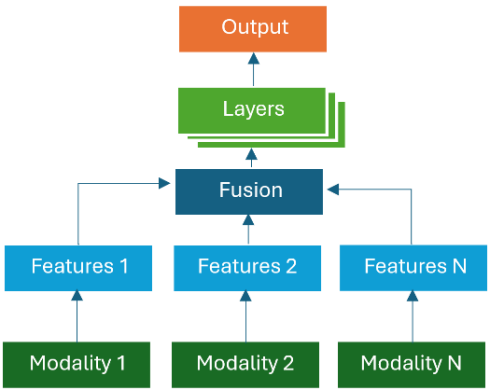


FIGURE 16. Intermediate fusion.

In this study, intermediate fusion is used as the fusion method. Figure 17 shows the architecture of the bimodal emotion recognition application BiMER. The diagram shows a bimodal approach used to recognize emotional states from raw speech signals and simultaneous text data. The model processes bimodal inputs with two main modules: speech (Speech Modality) and text (Textual Modality).

In the BiMER application, the data set we created for SER using data augmentation methods on mel spectrogram images was used as the data set.

Speech Modality: As explained in Section III: Proposed Methods, under subsection B. Speech Emotion Recognition, three different models were developed for SER. Among these models, the ResNet50-CRNN+AT model, which we achieved the highest accuracy in emotion recognition, was

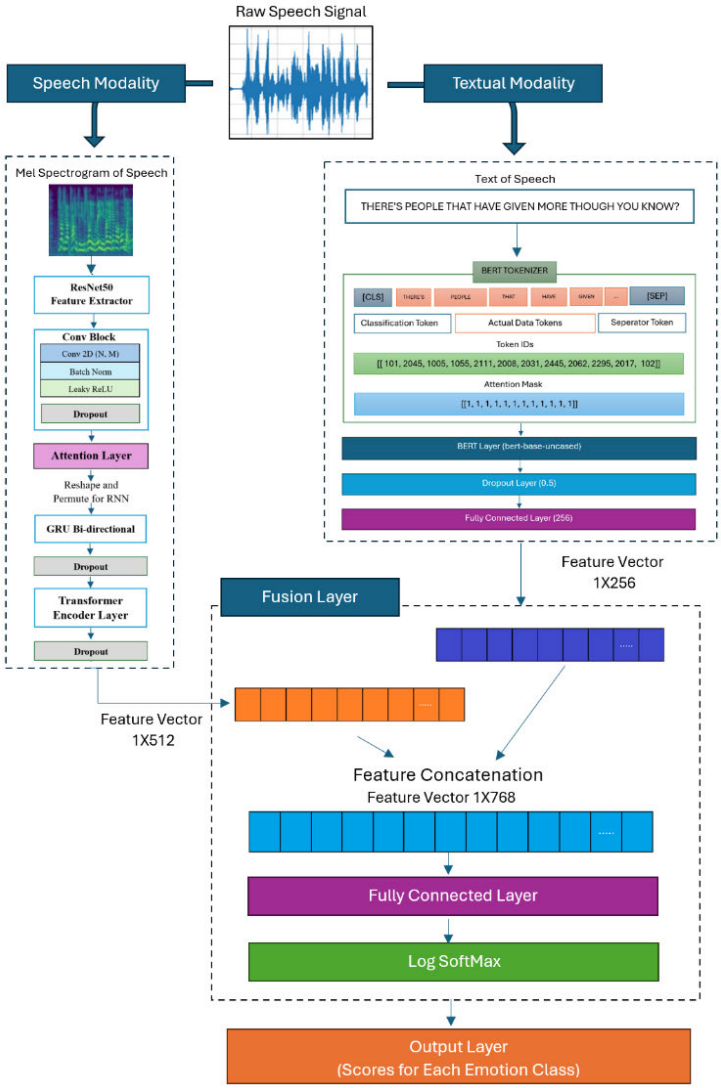


FIGURE 17. BiMER architecture with intermediate fusion.

used for BiMER's speech modality. The features obtained from this model are reduced to a single vector with global average pooling (AdaptiveAvgPool2d) and transmitted to a dropout layer. Then, the resulting feature vectors are transferred to the Fusion Layer.

Textual Modality: The details of the model used in the emotion recognition module from text are explained in Section III, under subsection C. Textual Emotion Recognition. Transcriptions of speech recordings are tokenized using the pre-trained BERT model and after which the corresponding token IDs and attention masks are generated. The BERT model takes these inputs and produces high-level features for each token. These features are then processed by a dropout layer and a fully connected layer. The resulting feature vectors are transferred to the fusion layer to be combined with the feature vectors of the speech recording.

Fusion Layer: The feature vectors obtained from the audio module (s = (16, 512) and the feature vectors obtained from the text module (t = (16, 256) are combined with the feature concatenation process c = [s;t]. This combination of features creates a larger feature vector c = (16, 768). Here, 16 represents the batch size, i.e. the number of samples processed at a time. In Figure 17, it is given as 1 × 768 for 1 sample. This vector serves as input to the next fully connected layer used to classify emotions. As a result, scores are calculated for neutral, angry, happy and sad emotion classes using the log softmax activation function.

The results of the implemented bimodal emotion recognition application are explained comparatively in the "Experimental Results of BiMER" section.

IV. WEB BASED IMPLEMENTATION OF BIMER

In this section, a web-based implementation of a bi-modal emotion recognition system that performs emotion recognition on speech and text data is described.

The system is designed to use Google Speech-to-Text API to process speech data directly and to convert it to text. The user interface of the web-based application, which has two

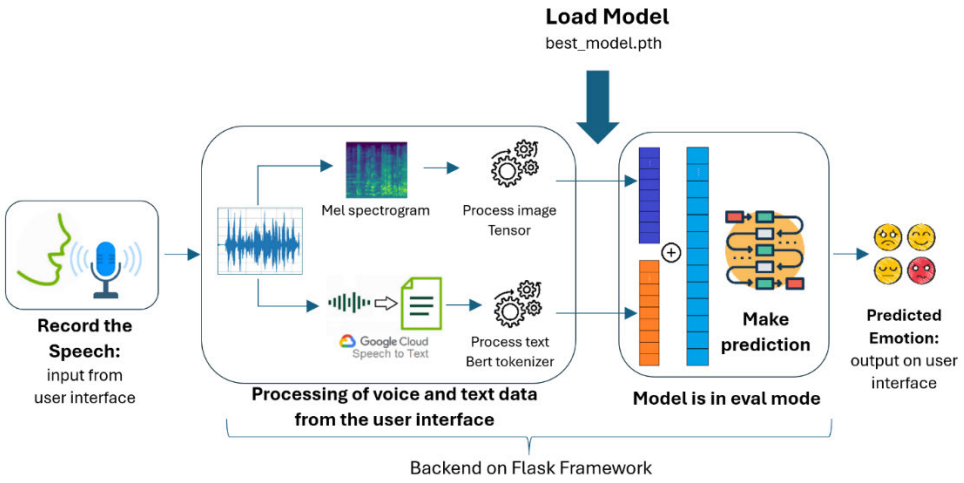


FIGURE 18. System diagram of bimodal emotion recognition implementation.

options that users can use by uploading a speech file or recording audio with a microphone, is shown in Figure 19. On the server side, a web server is set up with the help of Flask framework. This server processes the speech data received from users and the text transcriptions of these speeches to perform emotion recognition. For this purpose, at the end of the model training, which reached 88.3% accuracy, the weights of the model were saved to a file in.pth format. Before using the model, the trained weights are loaded from the. pth file. After the weights of the model are loaded using PyTorch's torch.load() function, the model is put into prediction mode (model.eval()).

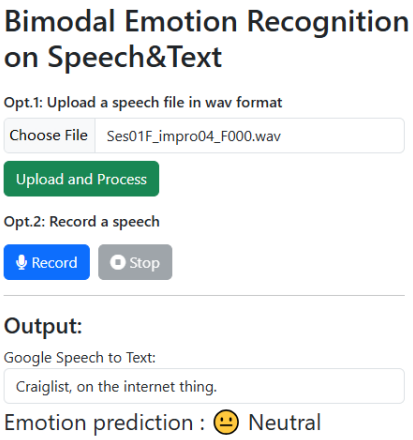


FIGURE 19. User Interface of BiMER.

Figure 18 shows the diagram showing the general architecture of the system. After users upload their speech recordings to the server via the web interface, the uploaded speech files are converted into mel spectrogram images, while text transcriptions of the speech are obtained using the Google Speech-to-Text API. This text data is then tokenized.

Pre-processed mel spectrogram images and text data are fed into a pre-trained bimodal emotion recognition model. As described in the section "Bimodal Emotion Recognition on Speech and Text", the model extracts features from speech and text data and combines these features to estimate emotions. The system presents the estimated emotions to the users in a visual format. This visualization is provided with the help of HTML and JavaScript using Flask's render_template feature.

V. EXPERIMENTAL RESULTS AND DISCUSSION

The results of the experimental studies on Speech Emotion Recognition (SER), Textual Emotion Recognition (TER) and Bimodal Emotion Recognition (BiMER) models are presented in this section. The models were also implemented on two different datasets where data augmentation was balanced, and class distributions were unbalanced to evaluate the effect of data augmentation on the performance of emotion recognition models. We used the Pytorch library to develop the models. The GPU model on which the application was implemented is NVIDIA GeForce RTX 3070-8GB. While evaluating the experimental results, recall and F1-score values were used as performance metrics of the model, and the success rate of the model was presented on a class basis. Additionally, to ensure a fair evaluation based on dataset distribution, Unweighted Accuracy (UA) was used for balanced datasets, while Weighted Accuracy (WA) was used for imbalanced datasets. WA was preferred for unbalanced datasets as it accounts for class distribution differences, preventing the evaluation from being biased toward majority classes. Conversely, UA was used for balanced datasets to equally reflect the performance across all classes [70].

A. EXPERIMENTAL RESULTS OF SER

In this section, the results obtained from three models developed for speech emotion recognition are presented.

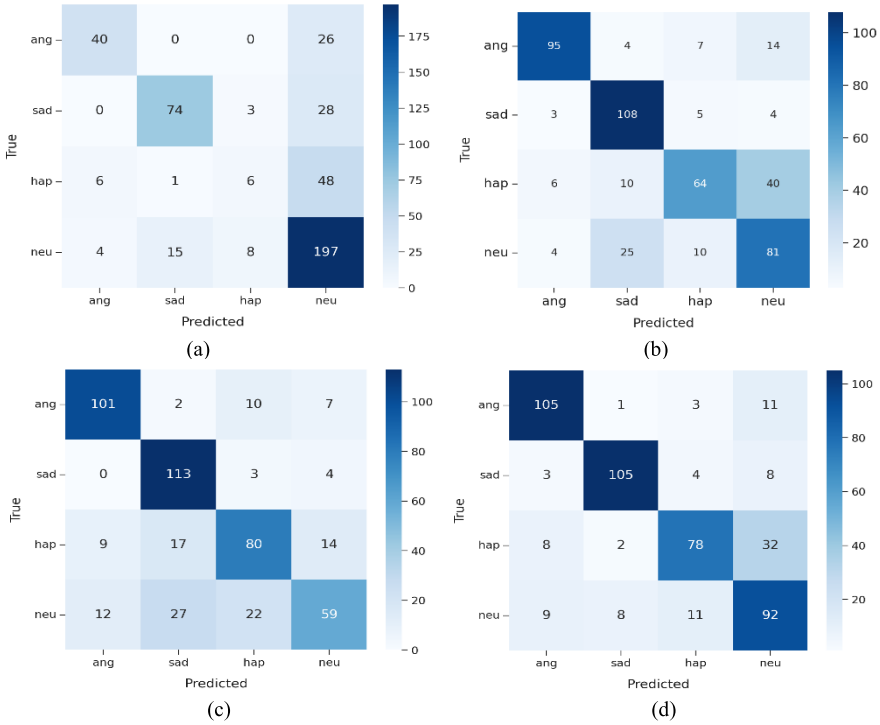


FIGURE 20. Confusion Matrices of SER: ResNet50-CRNN+AT on Unbalanced Dataset (a), ResNet-CNN (b), ResNet-CRNN+AT (c), ResNet50-CRNN+AT (d).

TABLE 5. Performance Metrics for SER models (*Weighted Accuracy (WA)Unweighted Accuracy (UA)*).

Class Names	Unbalanced Dataset		Augmented Dataset					
	ResNet50-CRNN+AT (a)		ResNet-CNN (b)		ResNet-CRNN+AT (c)		ResNet50-CRNN+AT (d)	
	Recall %	F1-score %	Recall %	F1-score %	Recall %	F1-score %	Recall %	F1-score %
Angry	60.61	68.97	79.16	83.33	84.17	83.47	87.50	85.71
Happiness	9.84	15.38	53.33	62.13	66.67	68.09	65.00	72.22
Neutral	87.95	75.33	67.50	62.55	49.17	57.84	76.67	69.96
Sad	70.48	75.90	90.00	80.90	94.17	81.00	87.50	88.98
Accuracy	WA: 69.52%		UA: 72.50%		UA: 73.54%		UA: 79.17%	

TABLE 6. Performance Metrics for TER (*Weighted Accuracy (WA)Unweighted Accuracy (UA)*).

Class Names	Unbalanced Dataset		Augmented Dataset	
	Recall %	f1-score %	Recall %	f1-score %
Angry	45.59	52.99	89.15	88.80
Happiness	31.67	44.71	83.49	80.18
Neutral	80.41	73.07	63.08	70.09
Sad	69.31	64.52	82.58	77.86
Accuracy	WA: 65.25%		UA: 79.57%	

The performance metrics of experimental studies conducted with the three models developed for SER; ResNet-CNN, ResNet-CRNN+AT and ResNet50-CRNN+AT are given in Table 5. The parameters used while training the ResNet-CNN and ResNet-CRNN+AT models were set as follows; learning rate: 10-5, weight decay: 1e-3, optimizer:

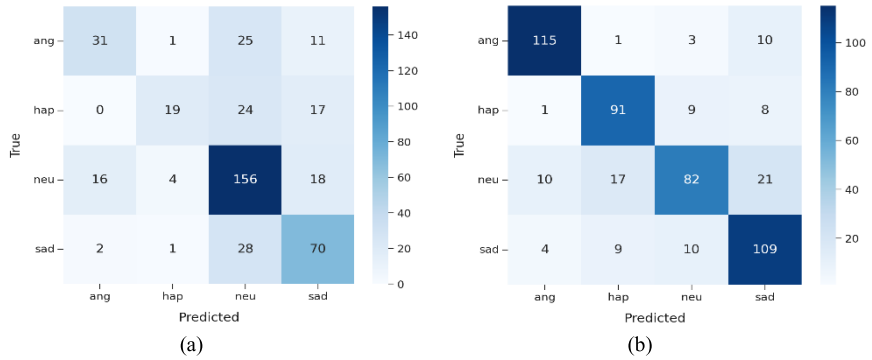


FIGURE 21. Confusion Matrix of TER on Unbalanced Dataset (a), Confusion Matrix of TER on Augmented Dataset (b).

TABLE 7. Comparative performance metrics for SER, TER and BiMER (*Weighted Accuracy (WA)Unweighted Accuracy (UA)*).

	Unimodal for SER ResNet50-CRNN+AT		Unimodal for TER BERT		BiMER	
Class labels	Recall %	f1 score %	Recall %	f1 score %	Recall %	f1 score %
Angry	87.50	85.71	89.15	88.80	88.33	92.98
Happiness	65.00	72.22	83.49	80.18	89.17	92.24
Neutral	76.67	69.96	63.08	70.09	88.33	80.00
Sad	87.50	88.98	82.58	77.86	87.50	89.36
Accuracy	UA: 79.17%		UA: 79.57%		UA: 88.33%	

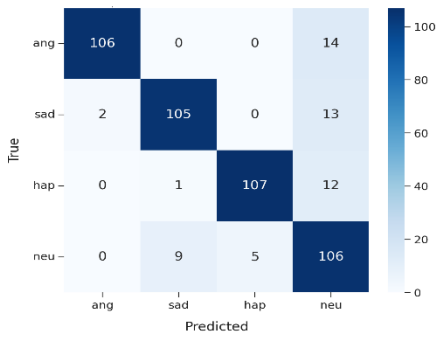


FIGURE 22. Confusion matrix of BiMER.

AdamW, batch size: 64, loss function: cross-entropy, activation function: LeakyReLU. In the training of the last model, ResNet50-CRNN+AT model, the batch size was set as 16. In the training of the models, the number of epochs ranged from 15 to 20, ensuring adequate learning without overfitting the data. The ResNet-CNN model achieved an unweighted accuracy (UA) of 72.50%, the ResNet-CRNN+AT model reached 73.54%, and the ResNet50-CRNN+AT model attained an unweighted accuracy of 79.17%. Three models were tested on the unbalanced dataset, and since the highest accuracy (WA) of 69.52% was achieved with ResNet50-CRNN+AT, the results of this model are given in Table 5.

When the results in Table 5 are examined, the ResNet50-CRNN+AT model achieved better results in all emotion classes compared to the other two models and reached the highest UA value (79.17%). This model showed great success especially in the Neutral and Sad classes and significantly increased the overall performance. Although ResNet-CRNN+AT provided a significant improvement in the Happiness class, it showed low performance in the Neutral class. Although ResNet-CNN had the lowest UA value, it showed strong results in the Angry and Sad classes.

In Figure 20, the confusion matrix graphs of these models are shown. In these graphs, 'ang' represents the angry class, 'neu' represents the neutral class, 'sad' represents the sad class, and 'hap' represents the happy class. When the results are examined, the best results are obtained for the 'neu' class, which is the most represented class in the imbalanced data set. The 'hap' class, which has the fewest examples, showed significantly lower performance. This situation highlights the difficulties created by imbalanced class distributions. When the confusion matrix of the ResNet-CNN model in Figure 20 (b) is examined, the correct prediction value is the highest for the 'sad' class. No improvement is observed in the performance metrics for the under sampled 'neu' emotion. This can be considered as a natural result of the under-sampling application. When the confusion matrix of

TABLE 8. Comparison results of classification performance of different approaches in state-of-the-art.

References, Year	IEMOCAP Dataset Usage	Emotions	Speech Modality Model	Text Modality Model	Fusion Method	Results
[68], 2020	impro+script	Happy, neutral, angry, sad	CNN, BiLSTM, Atteniton mechansim	ALBERT	Intermediate Fusion	WA : 71.06% UA : 72.05%
[31], 2020	impro+script	Happy, neutral, angry, sad	DCNN	Bi-RNN, DCNN	Intermediate Fusion	WA 80.51% UA : 79.22%
[32], 2021	impro+script	Happy, neutral, angry, sad	Hierarchical DNN (Final classification)	BiLSTM+ CNN	Feature Level Concatenation	Accuracy: 81.2% Average UR: 79.7%
[35], 2022	impro+script	Happy, neutral, angry, sad	CNN + BiLSTM	BERT	Multi-scale fusion	Accuracy: 75.181%
[69], 2022	impro+script	Happy, neutral, angry, sad	BiLSTM, MSA	GRU	Interactional Attention	WA: 82.4% UA: 80.6%
[37], 2023	impro+script	Happy (happy+excited), neutral, angry, sad	Wav2Vec 2.0, 1D CNN	Roberta,	Co-attention fusion	Weighted F1 score: 85.9%
[38], 2024	impro+script	Happy (happy+excited), neutral, angry, sad	Wav2Vec 2.0	BERT	Feature-level fusion	UAR: 78.42% WAR:77.75%
[70], 2024	Seperate experiments for imrpovised and scripted	Happy, neutral, angry, sad	CNN, CwGHP	BERT	Intermediate Fusion	improvised WA: 88.14% scripted WA : 82.34
[71], 2024	impro+script	Happy, neutral, angry, sad	Time frequency embedding, attention mechansim	Relative Entropy Alignment	Multi-level attention fusion	WA: 75.68% UA : 75.46%
[72], 2024	impro+script	Happy, neutral, angry, sad	BiGRU, Multihead Self Attention	BiGRU, Multihead Self Attention	Intermediate Fusion	WA: 76.52% UA: 77.13%
Proposed BiMER	improvised	Happy, neutral, angry, sad	ResNet50-CRNN+AT	BERT	Intermediate Fusion	**UA: 88.33%**

the ResNet-CRNN+AT model in Figure 20 (c) is examined, a significant increase is obtained for the angry and sad classes. However, it showed low performance for the neutral class. When the confusion matrix of the ResNet50-CRNN+AT model in Figure 20 (d) is examined, it is seen that there is a significant increase in the correct prediction rate in all classes.

B. EXPERIMENTAL RESULTS OF TER

This section includes the results obtained by applying the BERT model developed for TER to unbalanced datasets and augmented datasets. The datasets were divided into 80:20 ratio train set and validation set. The augmented dataset was set to have original texts in the validation set. In the experiments, the training of the model was completed after 6 epochs. The parameters used during the training of the model were set as learning rate: 2e-5, batch size: 64, optimizer: Adam and loss function: cross-entropy loss function. We observed a consistent improvement in validation accuracy and validation loss during the training in the balanced dataset and 6 epochs. In the unbalanced dataset, the start and end metrics were lower compared to the balanced dataset, which highlights the difficulties created by class imbalance. The model examined in our article reached 65.25% weighted accuracy rate on the dataset where the class distributions were unbalanced, while it reached 79.57% unweighted accuracy rate on the balanced dataset where augmented data was included. Table 6 shows performance metrics to evaluate model performance, and Figure 21 shows the confusion matrix. As a result of data augmentation, it is seen that there is a significant increase in the number of correct predictions for the ang, hap and sad classes. For the under sampled neutral class, although there is a decrease compared to the unbalanced dataset, consistency has occurred between the general classes.

C. EXPERIMENTAL RESULTS OF BiMER

In this section, we present the experimental results of our Bimodal Emotion Recognition application over speech and text. While the BiMER model is applied, an augmented dataset consisting of mel spectrogram images is used for the SER modality (ResNet50-CRNN+AT), and for the text modality (BERT), the text transformations corresponding to these speeches constitute the input of the model. In the final case, the training set consists of 800 records for each emotion class, while the validation set consists of 120 records for each emotion class. The validation set consists of only the original records. Learning rate: 10-5, weight-decay: 1e-3, optimizer: AdamW and loss function: cross-entropy loss function parameters were applied for the training of the

model. Due to hardware limitations, the batch size was determined as 8 when training this model.

When the performance metrics in Table 7 and the confusion matrix in Figure 22 are examined, it is revealed that BiMER is quite effective in detecting and correctly classifying emotions such as anger, sadness and happiness, but it has difficulty recognizing neutral situations compared to other classes. The model reached higher values in F1 scores in all classes compared to single-mode applications. It exhibited the highest performance with an accuracy of 88.33%, which shows that combining different data sources can significantly increase emotion recognition accuracy.

The comparative results of our application developed for BiMER with other models are given in Table 8. Some studies have presented general model evaluations specifying UA or WA values, while others have focused on performance metrics such as class-specific F1-scores or recall values. In the comprehensive analysis of bimodal emotion recognition systems, our proposed BiMER model, leveraging the ResNet50-CRNN+AT architecture coupled with BERT for text analysis, demonstrates enhanced performance when benchmarked against contemporary models. The comparative results depicted in the table show that BiMER achieves an impressive accuracy rate of 88.33%, which is notably higher than other approaches listed.

VI. CONCLUSION

This work has presented comprehensive research in the field of emotion recognition through both unimodal and bimodal approaches by integrating advances in speech and text processing technologies with deep learning frameworks. Significant improvements in model performances by applying data augmentation techniques have demonstrated their critical role in addressing the challenges of class imbalance and overfitting in machine learning.

In unimodal emotion recognition, separate models are developed for speech emotion recognition and text emotion recognition. Initially, the speech emotion recognition model exhibited an accuracy of 69.52%, which significantly increased to 79.17% after the application of data augmentation techniques. This improvement highlights the effectiveness of augmentation in enriching the training dataset and thus improving the model's ability to generalize from audio cues. Similarly, the accuracy of the text emotion recognition model increased from 65.25% to 79.57% through data augmentation techniques.. This increase in performance demonstrates its role in diversifying linguistic contexts and increasing the model's robustness to various textual emotion expressions.

The bimodal emotion recognition model built on unimodal foundations achieved a significant accuracy of 88.33% by combining features from both speech and text modes with the intermediate fusion method. This accuracy rate reveals that the proposed model has higher performance compared to other studies conducted and examined in the field of emotion recognition using the IEMOCAP dataset. This model captures a richer emotional context than can be distinguished from either model alone by leveraging the complementary strengths of its unimodal counterparts. It provides more comprehensive and accurate results by improving emotion recognition performance. Combining modalities will not only increase accuracy but also provide greater consistency and reliability across different datasets and scenarios. These results demonstrate the potential benefits of using multimodal approaches in emotion recognition applications.

In the final phase of the study, transforming the BiMER system into a real-time web application using the Flask framework represents a significant step towards practical, user-centric solutions. The web application seamlessly integrates Google's Speech to Text API to facilitate real-time emotion analysis from voice recordings provided by users, extracting text content from speech. This dual analysis capability enriches human-computer interaction and enhances the application's utility in real-world scenarios across various fields such as customer service, education, and healthcare. In future studies, the integration of additional modalities such as facial expressions, body movements or physiological signals can further increase the accuracy and applicability of the emotion recognition system. In addition, exploring more dynamic emotion states that take into account changes over time can provide improvements in scenarios where emotional states develop, such as conversations or therapeutic sessions. Integrating dimensional emotional models can further enhance the BiMER system. These models, which evaluate emotions across dimensions such as valence, arousal, and dominance, offer a more granular understanding of emotional states. By adopting a dimensional approach, future versions of BiMER can provide more detailed and dynamic emotional assessments. A limitation of this study is that the developed models have not been tested on other datasets containing improvised or real-life speech recordings beyond the IEMOCAP dataset. Again, how language features affect performance can be examined with datasets created or to be created in different languages other than English. We aim to expand the task by addressing these disadvantages in future studies.

In conclusion, this study contributes to the development of high-accuracy models in the field of emotion recognition with datasets that have limited and real-life scenarios, and to future studies in the field of multimodal human-computer interaction by using these in a real-world application.

REFERENCES

[1] G. V. Singh, S. Ghosh, M. Firdaus, A. Ekbal, and P. Bhattacharyya, "Unmasking offensive content: A multimodal approach with emotional understanding," *Multimedia Tools Appl.*, vol. 2025, pp. 1–24, Jan. 2025, doi: 10.1007/s11042-025-20603-w.

[2] G. Zhao, Y. Zhang, and J. Chu, "A multimodal teacher speech emotion recognition method in the smart classroom," *Internet Things*, vol. 25, Apr. 2024, Art. no. 101069, doi: 10.1016/j.iot.2024.101069.

[3] S. Ghosh, G. V. Singh, A. Ekbal, and P. Bhattacharyya, "COMMA-DEER: COmmon-sense aware multimodal multitask approach for detection of emotion and emotional reasoning in conversations," in *Proc. 29th Int. Conf. Comput. Linguistics*, Oct. 2022,pp. 6978–6990. [Online]. Available: https://aclanthology.org/2022.coling-1.608/

[4] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "Deep cascaded multitask framework for detection of temporal orientation, sentiment and emotion from suicide notes," *Sci. Rep.*, vol. 12, no. 1, p. 4457, Mar. 2022, doi: 10.1038/s41598-022-08438-z.

[5] P. Ghadekar, M. Ranjan Pradhan, D. Swain, and B. Acharya, "EmoSecure: Enhancing smart home security with FisherFace emotion recognition and biometric access control," *IEEE Access*, vol. 12, pp. 93133–93144, 2024, doi: 10.1109/ACCESS.2024.3423783.

[6] A. K. Bar and A. K. Chaudhuri, "Emotica.AI–A customer feedback system using AI," *Int. Res. J. Adv. Sci. Hub*, vol. 5, no. 3, pp. 103–110, Mar. 2023, doi: 10.47392/irjash.2023.019.

[7] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, Sep. 2017, doi: 10.1016/j.imavis.2017.08.003.

[8] R. Francese, M. Risi, and G. Tortora, "A user-centered approach for detecting emotions with low-cost sensors," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 35885–35907, Nov. 2020, doi: 10.1007/s11042-020-09576-0.

[9] A. Zadeh, "Micro-opinion sentiment intensity analysis and summarization in online videos," in *Proc. ACM Int. Conf. Multimodal Interact.*, Nov. 2015, pp. 587–591, doi: 10.1145/2818346.2823317.

[10] E. H. Houssein, A. Hammad, and A. A. Ali, "Human emotion recognition from EEG-based brain–computer interface using machine learning: A comprehensive review," *Neural Comput. Appl.*, vol. 34, no. 15, pp. 12527–12557, May 2022, doi: 10.1007/s00521-022-07292-4.

[11] Y. Cai, X. Li, and J. Li, "Emotion recognition using different sensors, emotion models, methods and datasets: A comprehensive review," *Sensors*, vol. 23, no. 5, p. 2455, Feb. 2023, doi: 10.3390/s23052455.

[12] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Proc. Nebraska Symp. Motivat.*, vol. 1971, Jan. 1972, pp. 207–282.

[13] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, vol. 476, T. Dalgleish and M. Power, Eds., Sussex, U.K.: Wiley, 1999.

[14] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *Amer. Sci.*, vol. 89, no. 4, pp. 344–350, 2001.

[15] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Personality*, vol. 11, no. 3, pp. 273–294, Sep. 1977.

[16] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "VAD-assisted multitask transformer framework for emotion recognition and intensity prediction on suicide notes," *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103234, doi: 10.1016/j.ipm.2022.103234.

[17] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Nov. 2008, doi: 10.1007/s10579-008-9076-6.

[18] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," 2021, *arXiv:2102.01813*.

[19] N. Braunschweiler, R. Doddipatla, S. Keizer, and S. Stoyanchev, "A study on cross-corpus speech emotion recognition and data augmentation," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 24–30, doi: 10.1109/ASRU51503.2021.9687987.

[20] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Multitask learning from augmented auxiliary data for improving speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3164–3176, Oct. 2022, doi: 10.1109/TAFFC.2022.3221749.

[21] B. T. Atmaja and A. Sasou, "Effects of data augmentations on speech emotion recognition," *Sensors*, vol. 22, no. 16, p. 5941, Aug. 2022, doi: 10.3390/s22165941.

[22] Z. Tu, B. Liu, W. Zhao, R. Yan, and Y. Zou, "A feature fusion model with data augmentation for speech emotion recognition," *Appl. Sci.*, vol. 13, no. 7, p. 4124, Mar. 2023, doi: 10.3390/app13074124.

[23] A. Dang, T. H. Vu, L. Dinh Nguyen, and J.-C. Wang, "EMIX: A data augmentation method for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096789.

[24] L. Qu, W. Wang, C. Weber, P. Yue, T. Li, and S. Wermter, "Improving speech emotion recognition with unsupervised speaking style transfer," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 10101–10105, doi: 10.1109/ICASSP48485.2024.10446186.

[25] H. Q. Abonizio, E. C. Paraiso, and S. Barbon, "Toward text data augmentation for sentiment analysis," *IEEE Trans. Artif. Intell.*, vol. 3, no. 5, pp. 657–668, Oct. 2022, doi: 10.1109/TAI.2021.3114390.

[26] M. M. Imran, Y. Jain, P. Chatterjee, and K. Damevski, "Data augmentation for improving emotion recognition in software engineering communication," in *Proc. 37th IEEE/ACM Int. Conf. Automated Softw. Eng.*, Oct. 2022, pp. 1–13, doi: 10.1145/3551349.3556925.

[27] X. Gong, W. Ying, S. Zhong, and S. Gong, "Text sentiment analysis based on transformer and augmentation," *Frontiers Psychol.*, vol. 13, May 2022, Art. no. 906061, doi: 10.3389/fpsyg.2022.906061.

[28] F. Mohammad, M. Khan, S. Nawaz Khan Marwat, N. Jan, N. Gohar, M. Bilal, and A. Al-Rasheed, "Text augmentation-based model for emotion recognition using transformers," *Comput., Mater. Continua*, vol. 76, no. 3, pp. 3523–3547, Jan. 2023, doi: 10.32604/cmc.2023.040202.

[29] A. Messaoudi, H. Boughrara, and Z. Lachiri, "Modeling continuous emotions in text data using IEMOCAP database," in *Proc. IEEE 7th Int. Conf. Adv. Technol., Signal Image Process. (ATSIP)*, vol. 1, Jul. 2024, pp. 397–402, doi: 10.1109/ATSIP62566.2024.10638843.

[30] A. Onan and K. Filiz Balbal, "Improving Turkish text sentiment classification through task-specific and universal transformations: An ensemble data augmentation approach," *IEEE Access*, vol. 12, pp. 4413–4458, 2024, doi: 10.1109/ACCESS.2024.3349971.

[31] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 3227–3231, doi: 10.1109/ICASSP40776.2020.9054441.

[32] P. Singh, R. Srivastava, K. P. S. Rana, and V. Kumar, "A multimodal hierarchical approach to speech emotion recognition from audio and text," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107316, doi: 10.1016/j.knosys.2021.107316.

[33] S. Padi, D. Manocha, and R. D. Sriram, "Multi-window data augmentation approach for speech emotion recognition," 2020, *arXiv:2010.09895*.

[34] T. Zhang, S. Li, B. Chen, H. Yuan, and C. L. P. Chen, "AIA-Net: Adaptive interactive attention network for text–audio emotion recognition," *IEEE Trans. Cybern.*, vol. 53, no. 12, pp. 7659–7671, Dec. 2023, doi: 10.1109/TCYB.2022.3195739.

[35] S. William and A. Zahra, "Bimodal emotion recognition using text and speech with deep learning and stacking ensemble technique," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 19, pp. 5521–5530, Oct. 2022.

[36] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," 2022, *arXiv:2203.13504*.

[37] S. Dutta and S. Ganapathy, "HCAM-hierarchical cross attention model for multi-modal emotion recognition," 2023, *arXiv:2304.06910*.

[38] Z. D. Tafti and B. BabaAli, "Audio-textual emotion recognition using pre-trained models: Investigating various representations and fusion techniques," *Res. Square*, Sep. 2024, doi: 10.21203/rs.3.rs-4963739/v1.

[39] L. Guder, J. P. Aires, F. Meneguzzi, and D. Griebler, "Dimensional speech emotion recognition from bimodal features," in *Proc. Simpósio Brasileiro De Computação Aplicada Saúde*, Jun. 2024, pp. 579–590, doi: 10.5753/sbcas.2024.2779.

[40] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021, doi: 10.1109/ACCESS.2021.3067460.

[41] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.

[42] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391, doi: 10.1371/journal.pone.0196391.

[43] S. Haq, "Audio visual expressed emotion classification," Ph.D. Dissertation, Fac. Eng. Phys. Sci., Univ. Surrey, Guildford, U.K., 2011.

[44] E. DDikbıyık, Ö. Demir, and B. Doğan, "Derin öğrenme yöntemleri İle konuşmadan duygu tanıma üzerine bir literatür araştırması a literature review on speech emotion recognition using deep learning techniques," *Gazi Üniversitesi Fen Bilimleri Dergisi C, Tasarım Ve Teknoloji*, vol. 10, no. 4, pp. 765–791, Dec. 2022, doi: 10.29109/gujsc.1111884.

[45] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[46] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Amer.*, vol. 8, no. 3, pp. 185–190, Jan. 1937, doi: 10.1121/1.1915893.
[47] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowl.-Based Syst.*, vol. 244, May 2022, Art. no. 108580, doi: 10.1016/j.knosys.2022.108580.
[48] E. Kamalloo, M. Rezagholizadeh, and A. Ghodsi, "When chosen wisely, more data is what you need: A universal sample-efficient strategy for data augmentation," 2022, *arXiv:2203.09391*.
[49] S. Wei, S. Zou, F. Liao, and W. Lang, "A comparison on data augmentation methods based on deep learning for audio classification," *J. Phys., Conf. Ser.*, vol. 1453, no. 1, Jan. 2020, Art. no. 012085, doi: 10.1088/1742-6596/1453/1/012085.
[50] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2613–2617, doi: 10.21437/INTERSPEECH.2019-2680.
[51] P. Bahar, A. Zeyer, R. Schlüter, and H. Ney, "On using SpecAugment for end-to-end speech translation," 2019, *arXiv:1911.08876*.
[52] M. Yusuf Faisal and S. Suyanto, "SpecAugment impact on automatic speaker verification system," in *Proc. Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRITI)*, Dec. 2019, pp. 305–308, doi: 10.1109/ISRITI48646.2019.9034603.
[53] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech emotion recognition based on parallel CNN-attention networks with multi-fold data augmentation," *Electronics*, vol. 11, no. 23, p. 3935, Nov. 2022, doi: 10.3390/electronics11233935.
[54] P. R. A. S. Bassi, W. Rampazzo, and R. Attux, "Transfer learning and SpecAugment applied to SSVEP based BCI classification," *Biomed. Signal Process. Control*, vol. 67, May 2021, Art. no. 102542, doi: 10.1016/j.bspc.2021.102542.
[55] S. Padi, S. O. Sadjadi, R. D. Sriram, and D. Manocha, "Improved speech emotion recognition using transfer learning and spectrogram augmentation," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 645–652, doi: 10.1145/3462244.3481003.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[57] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.
[58] I. Z. Mukti and D. Biswas, "Transfer learning based plant diseases detection using ResNet50," in *Proc. 4th Int. Conf. Electr. Inf. Commun. Technol. (EICT)*, Khulna, Bangladesh, Dec. 2019, pp. 1–6, doi: 10.1109/EICT48899.2019.9068805.
[59] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowl. Inf. Syst.*, vol. 62, no. 8, pp. 2937–2987, Mar. 2020, doi: 10.1007/s10115-020-01449-0.
[60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North*, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/n19-1423.
[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.
[62] C.-G. Kim, Y.-J. Hwang, and C. Kamyod, "A study of profanity effect in sentiment analysis on natural language processing using ANN," *J. Web Eng.*, vol. 21, no. 3, pp. 751–766, May 2022, doi: 10.13052/jwe1540-9589.2139.
[63] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," 2021, *arXiv:2105.03075*.
[64] G. H. de Rosa and J. P. Papa, "A survey on text generation using generative adversarial networks," *Pattern Recognit.*, vol. 119, Nov. 2021, Art. no. 108098, doi: 10.1016/j.patcog.2021.108098.
[65] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *J. Big Data*, vol. 8, no. 1, p. 101, Jul. 2021, doi: 10.1186/s40537-021-00492-0.
[66] Y. Li, M. El Habib Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec, "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Comput. Biol. Med.*, vol. 177, Jul. 2024, Art. no. 108635, doi: 10.1016/j.compbiomed.2024.108635.
[67] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Mach. Vis. Appl.*, vol. 32, no. 6, p. 121, Sep. 2021, doi: 10.1007/s00138-021-01249-8.
[68] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in *Proc. Interspeech*, Oct. 2020, pp. 374–378, doi: 10.21437/INTERSPEECH.2020-3156.
[69] Y. Tang, Y. Hu, L. He, and H. Huang, "A bimodal network based on audio–text-interactional-attention with arcface loss for speech emotion recognition," *Speech Commun.*, vol. 143, pp. 21–32, Sep. 2022, doi: 10.1016/j.specom.2022.07.004.
[70] K. Chauhan, K. K. Sharma, and T. Varma, "Multimodal emotion recognition using contextualized audio information and ground transcripts on multiple datasets," *Arabian J. Sci. Eng.*, vol. 49, no. 9, pp. 11871–11881, Nov. 2023, doi: 10.1007/s13369-023-08395-3.
[71] J. Lei, J. Wang, and Y. Wang, "Multi-level attention fusion network assisted by relative entropy alignment for multimodal speech emotion recognition," *Appl. Intell.*, vol. 54, nos. 17–18, pp. 8478–8490, Jun. 2024, doi: 10.1007/s10489-024-05630-8.
[72] Y. Shang and T. Fu, "Multimodal fusion: A study on speech-text emotion recognition with the integration of deep learning," *Intell. Syst. Appl.*, vol. 24, Dec. 2024, Art. no. 200436, doi: 10.1016/j.iswa.2024.200436.
[73] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 1877–1901.
[74] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, Jan. 2019.

EMRAH DIKBIYIK received the B.Sc. and M.S. degrees in computer and control education from Marmara University, in 2009 and 2013, respectively, where he is currently pursuing the Ph.D. degree in electrical-electronics engineering. Since 2020, he has been a Lecturer with the Department of Computer Technologies, Istanbul University-Cerrahpaşa. His research interests include artificial intelligence, machine learning, deep learning, speech analysis and processing techniques, and software engineering.

ONDER DEMIR received the M.S. and Ph.D. degrees in electronics and computer education from Marmara University, in 2006 and 2013, respectively. From 2003 to 2013, he was a Research Assistant and a Lecturer. He has been an Associate Professor with the Department of Computer Engineering, Marmara University. His research interests include digital image processing, biomedical image processing, cyber security, and algorithms.

BUKET DOGAN received the B.Sc. degree from the Department of Electronics and Computer Education, Faculty of Technical Education, Marmara University, Istanbul, Türkiye, in 1999, and the M.S. and Ph.D. degrees from Marmara University, in 2001 and 2006, respectively. She is currently an Associate Professor with the Department of Computer Engineering, Marmara University. Her research interests include data mining, artificial intelligence, machine learning, and software project management.