SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

A Lightweight CNN-BiLSTM-Attention Framework for Speaker-Independent Speech Emotion Recognition Using CREMA_D Dataset

Amjed Kadhum Aljadiri[1]*, Mohammed S. H. Al-Tamimi[2]*

[1]Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq.
E-mail : Amjad.Kadhum2501m@sc.uobaghdad.edu.iq
[2]Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq. E-mail: mohammed.s@sc.uobaghdad.edu.iq

Abstract
With applications in affective computing and human-computer interaction HCI, Speech Emotion Recognition (SER) seeks to deduce a speaker's emotional state from the acoustic information in speech. However, due to the variability of speakers and the possibility of speaker overlap in the training and testing sets, it is difficult to measure the performance of deep learning-based SER systems. In this work, an efficient hybrid CNN-BiLSTM-Attention approach for the speech emotion recognition problem is proposed which uses the CREMA-D dataset. The proposed framework uses a low dimensional set of 42 handcrafted acoustic features which include MFCCs, delta MFCCs, delta MFCCs, pitch, zero crossing rate and RMS energy. The CNN component is learned to address local acoustic structure, and the BiLSTM models long-term temporal structure, along with the attention mechanism emphasizing emotionally informative speech frames. A strict speaker independent protocol is used to create a more realistic evaluation of model generalization, where the test set speakers are not used in training. The model is tested within six emotions: Angry, Disgust, Fear, Happy, Neutral, and Sad. The proposed framework obtains an accuracy of 71.72% at a test and a UAR of 71.95%. The result of per-emotion analysis shows that angry has the highest F1-score, which is 83.06%, while Neutral had the second highest F1-score, which is 79.09%, and Disgust is the most difficult class with F1 score of 59.20%. The results show that the proposed lightweight architecture could effectively realize the balanced emotion recognition and a more stringent assessment for generalization to speakers' unseen during training.

Keywords: (Speech Emotion Recognition SER, Deep Learning DL, speaker independent, Feature Extraction, CREMA_D dataset)

1- Introduction

Speech is a signal that humans generate in order to engage in communication and interaction (1). Due to its ability to independently identify human emotional states from voice data, (SER) has grown in importance as a field of study in affective computing. Precise emotion identification facilitates more intuitive and intelligent (HCI), and underpins various applications: such as healthcare monitoring, intelligent virtual assistants, e-learning, customer service, and driver assistance systems. Recent advancements in deep learning have significantly enhanced speech emotion recognition (SER) performance by allowing models to acquire discriminative emotional representations directly from speech signals, rather than depending exclusively on manually created acoustic features. As a result, deep neural networks have emerged as the prevailing method for contemporary speech emotion recognition systems (2). With the development of speech recognition technology, (SER) has been growing in popularity in the rapidly developing field of speech recognition. It has to do with the study of human-computer interactions (HCIs). This is because it is important to examine how emotions are conveyed in speech in order to raise conversational intelligence. There are also robotics and HCI systems that might be explored. By engaging with people, and grasping the meanings of it, it is also possible to deliver better services and to build a better service, by taking into account the emotions that are conveyed in speech to make human-machine interaction more intelligent, organic, and customized. Speech emotion recognition has become more important in a variety of application domains such as home automation, customer service, healthcare, finance, and others. Uses and even enjoyment. In all these applications, communication is effective interaction between human and computer/machine/robot needs understanding of human the intentions of the users, which could be inferred from speech emotions (3). Many techniques have been proposed in recent years to extract features from audio signals in order to handle issues like signal complexity and noise. Many features are employed such as zero cross rating, pitch, Mel-Frequency Cepstral Coefficients (MFCC), and Root Mean Square Energy (RMSE). Additionally, it has been discovered that

sophisticated neural models such as CNN, (CNN-LSTM), and Bi-Directional Long Time Short Time For audio signal processing, memory (Bi-LSTM) are helpful. When it comes to automatically learning temporal correlations and extracting valuable information from audio sources, these models excel. However, despite its potential, SER research has difficulties because of a shortage of high-quality Nevertheless, despite its potential, SER research faces difficulties because of poor quality. The subjectivity of feelings, the intricacy of data and feature extraction techniques. Although reasonably high accuracy has been achieved thru the extraction of features such as spectral and qualitative information, their extraction requires specialist knowledge. In order to achieve this, this study presents a lightweight hybrid deep learning architecture for SER that makes use of manually created acoustic features obtained from the CREMA-D emotional speech database. MFCCs, delta MFCCs, delta-delta MFCCs, pitch, zero-crossing rate, and root mean square (RMS) energy are among the frame-level acoustic parameters that are modeled as variable-length sequences in the utterances. A hybrid CNN1D-BiLSTM network with an attention mechanism receives the extracted sequences. Prior to the final classification step, the attention layer locates the speech frames that are instructive for emotions, the convolutional layers extract local spectral patterns, and the BiLSTM layers learn long-term contextual information. Since the test set's speaker is typically excluded from the training set, a strict speaker-independent approach is employed to test the suggested framework with a realistic evaluation of model generalization when it is used in practical applications, this speaker-independent assessment method is more likely to accurately reflect system performance. This study has made the following contributions:

1- A compact frame-level acoustic representation: A 42-dimensional handcrafted acoustic representation is used which includes MFCCs, first-order MFCC derivatives, second-order MFCC derivatives, pitch, zero crossing rate, and RMS energy. This representation embodies complementary characteristics in the spectra, temporal, prosodic and energy dimensions with a relatively low dimensional representation of the input.

2- A lightweight hybrid CNN-BiLSTM-Attention framework: A hybrid architecture is designed to take advantage of complementary temporal features of emotional speech. CNN layers extract local acoustic patterns from the frame-level feature sequence; BiLSTM models bidirectional temporal dependencies; the attention mechanism underscores the frames with the greatest emotional informativeness for utterance-level classification.

3- Strict speaker-independent evaluation: The proposed framework is evaluated with the speaker independent data partitioning strategy, where the speakers are disjointed in the sets for training, validation and test. This offers a more stringent test of generalization to a different speaker not included in the training set, and prevents speaker leakage of its data.

4-In-depth performance evaluation: The proposed system is evaluated using complementary metrics such as Accuracy, UAR, Precision, Recall, F1-score, AUC and confusion-matrix analysis which enables the overall and emotion-specific performance to be explored.

5-Comparative evaluation with existing CREMA-D studies: Performance of the proposed framework is compared with existing deep learning-based SER approaches on CREMA-D, paying special consideration to differences in the evaluation protocols, particularly speaker-independent versus random utterance-level partitioning.

2- Related work

Overview of the field. With the introduction of deep learning techniques, Speech Emotion Recognition (SER) systems have seen a great transformation. Most of the early SER methods involves manually engineered acoustic features like MFCC, pitch, and spectral descriptors along with traditional machine learning classifiers like SVM, Random Forest, GMM and HMM. Later, research gradually turns to the deep neural architectures like CNN, LSTM/BiLSTM, attention-based models, and Transformers, motivated by the expectation of learning feature representations which are good at discriminating among emotions directly from speech, without the need for extra handcrafted features. The rest of this section summarizes these architectures in the order of closeness to the framework proposed in this study, and focuses on a subset of studies that are evaluated on the CREMA-D dataset, and, subsequently discussed quantitatively in Table 6.

CNN and CNN1D based approaches: One of the first deep models that are used in SER is convolutional architectures due to their capability to learn local spectral patterns directly from the acoustic feature sequences. The 1D CNN and Feature-Selected Random Forest (RF) models are trained using data augmentation on four public datasets (RAVDESS, CREMA, TESS, and, SAVEE) with 1D CNN model reaching 69% accuracy, whereas the RF model achieved 61% (4), the sample size is not exactly defined in the reported results, and no confusion matrix is provided, but data augmentation is used prior to train/test split. (5) designs a CNN with a Convolutional Block Attention Module (CBAM) and tests it on CREMA-D with an 80/10/10 random split that resulted in a weighted accuracy of 68.30% and unweights accuracy of 68.22%, but does not report precision, recall, and F1 score and has a potential risk of speaker overlap between partitions. (3) introduced a meta-learning optimized CNN which obtained an 83.28% classification rate on CREMA-D in the case of random train-

test split, without class-wise analysis, and feature pre-processing before train-test split. In total, the accuracies from these CNN-based studies are moderate (61-83%) on CREMA-D, but the utterances are partitioned randomly per level and not guaranteed to be speaker independent between training and test sets.

Recurrent (LSTM/BiLSTM) and embedding based approaches: Emotion is spread across individual frames of an utterance, not just limited to a specific frame, recurrent architectures, such as Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) networks are commonly used to model emotion over time with bidirectional processing enabling a network to benefit from both past and future context, which also improves accuracy in recognition (6). Along these lines (7) applied transfer learning, combining speaker embeddings (d-vector, x-vector, r-vector) with the multi-head attention mechanism and an AM-Softmax loss, with an unweighted accuracy of 80.25%, and a weighted accuracy of 79.81%, on CREMA-D with a random 80/20 split (no speaker independent partitioning).

Hybrid CNN-LSTM/BiLSTM approaches: Combining convolutional and recurrent layers allows a model to capture local spectral structure and long-range temporal dependencies jointly. The proposed DCRF-BiLSTM model yielded the highest accuracy of 95.10% on CREMA-D with a random 80/20 split and the confusion matrix reported is not consistent, while there is no mention of the risk of speaker overlap between partitions. Directly comparing a CNN-LSTM model with an attention-enhanced CNN-LSTM model, (6) shows that the attention network could greatly improve the performance of the CNN-LSTM model on the benchmark emotional speech datasets, and that it enhances the model's generalization ability; this encourages the incorporation of attention into CNN-recurrent hybrids. In this model, CNN layers are used for local feature extraction, BiLSTM layers is used to model the temporal dependency, and a Transformer encoder is employed to model the global context with self-attention. The study uses 40 MFCC coefficients as acoustic features and applies a Partial-Disordered Speech Augmentation Strategy (PDSAS), which increases the number of samples from 7,442 to 59,536 before applying five-fold cross-validation. The proposed framework are able to attain an accuracy of 93.2% and an F1-score of 92.5%, respectively, outperforming the CNN, BiLSTM, CNN–BiLSTM, and CNN–Transformer baseline models (8). Increasing the number of samples from 7,442 to 59,536 resulted in the same speakers appearing in both the training and testing sets, this is considered a weakness of the study and poses a risk of speaker overlap. The evaluation protocol does not clearly establish strict speaker-independent separation and augmentation isolation between training and evaluation folds.

Attention-enhanced hybrid architectures: most closely related to this work. The combined CNN and recurrent layers with attention mechanisms are the closest architecture family to the one proposed here, as there are several recent studies that could be found. In another study, an Improved Jaya Optimization Algorithm (IJOA) is used to select features, followed by a Convolutional Peephole LSTM and attention mechanism to get 99.12% accuracy (9) but the data-split strategy and speaker independence are not well described. To simultaneously capture the spatial-temporal features of facial expressions, (10) introduces a hybrid CNN-BiLSTM deep learning model that incorporates multiple attention mechanisms and shows that spatial-temporal emotion representations generated by this model are more discriminative than those generated by deep learning models based on either modality alone. (11) proposed a CNN with channel attention and spatial attention (CBAM-DenseNet121), and achieved 71.26% accuracy, 71.01% UAR and 71.25% F1-score on CREMA-D when it is split into training/validation/test sets, but it is not explicitly reported whether the partitioning is speaker-independent or not, and there is an overfitting effect as the accuracy on the validation set is higher than the test set. Combining all of these studies, it could be concluded that the attention mechanism always helps HYBRID CNN-RNN SER models to achieve better performance, validating the design of the present work; however, no strict speaker-independent evaluation protocol is imposed.

Raw-waveform and Transformer-based approaches: One line of work avoids the need for handcrafted feature extraction altogether. (12) directly uses raw audio waveforms as input to a CNN based SER architecture and achieves 69.72% accuracy on CREMA-D in an 80/20 random split of the test set, eliminating the need to avoid speaker overlap. Transformer bases architectures and self-supervised speech representation models like HuBERT and Wav2Vec2 have recently been shown to obtain state-of-the-art performance in SER, but are generally very resource-intensive, requiring significantly more hardware resources, training time, and number of trainable parameters, making them impractical for deployment in embedded devices and real-time systems.

Evaluation method and research gap: Based on the recent survey and meta-analysis studies, the major areas of research in current SER are enhancement of recognition accuracy, enhancement of model generalization, simplicity of computation, and designing realistic evaluation protocols (13, 14). These reviews underscore the

ongoing relevance of lightweight hybrid models in comparison to transformer-based models, which are known to be highly costly in terms of computation. The studies review above, and summarize quantitatively in Table 5, have a common methodological disadvantage in that they generally use random splitting of utterances, in which utterances from the same speaker could be in both the training and test sets. This partitioning could also introduce speaker-specific information into the learning process, leading to optimistic performance estimates as compared to a realistic deployment scenario. Owing to this, recent benchmarking efforts strongly suggest a strictly speaker-independent evaluation protocol where the speakers for evaluation are not included in the training of the model (15). Based on these findings, the current study uses a compact 42-dimensional handcrafted acoustic representation in conjunction with a lightweight hybrid CNN1D-BiLSTM network supplemented with an attention mechanism. The suggested framework stresses computing efficiency while maintaining the ability to capture long-range temporal relationships and local spectrum patterns, in contrast to the transformer-based approaches mentioned above. Additionally, the suggested system is assessed using a stringent speaker-independent methodology, which offers a more realistic evaluation of the model's generalization to unknown speakers than most of the CREMA-D research covered in this area (Table 6).

3- Proposed Method

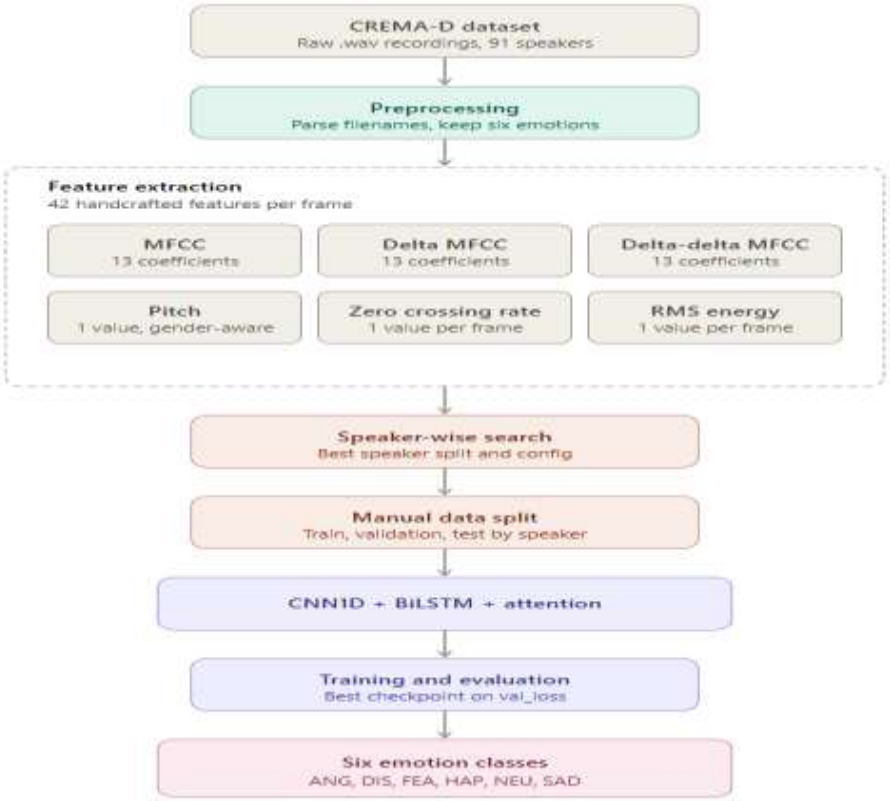

Figure 1: The proposed system

3.1 Dataset

Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), one of the most popular benchmark datasets for emotion identification research, is utilized to create and evaluate the proposed Speech Emotion recognition (SER) framework (16). In this regard, CREMA-D is a set of spoken emotional utterances recorded under controlled recording conditions by professional actors that could be used for training and testing speech-based affective computing systems, offering good-quality utterances of emotion. The entire CREMA-D corpus comprises of recordings from 91 actors covering a wide range of genders and age groups. Actors say a set of predetermined sentences with various expressions and intensities. The dataset contains a considerable number of speakers, (each sentence is spoken by many speakers, with different facial expressions) suitable for testing speaker independent speech emotion recognition models. The present work focuses on the speech modality of CREMA-D, which is a multimodal emotional database consisting of audio, video and visual information of faces. As a consequence, only audio recordings are used for all experiments, which enables the proposed framework to study emotional information in speech without considering any visual information.

Every speech recording is saved as a 16 KHz uncompressed WAV file, which has enough temporal resolution to capture detailed acoustic properties, but still allows for efficient computation. The dataset features emotional labels that have been annotated perceptually, which guarantees reliable emotional labelling of the dataset that are consistent with the intended emotional expression. The six emotion categories in the original CREMA-D corpus are matched with each speech recording. These categories are:

Table 1 the samples distribution

Emotion	Count
Angry	1271 samples
Happy	1271 samples
Sad	1271 samples
Fear	1271 samples
Disgust	1271 samples
Neutral	1087 samples
Total	7442 samples

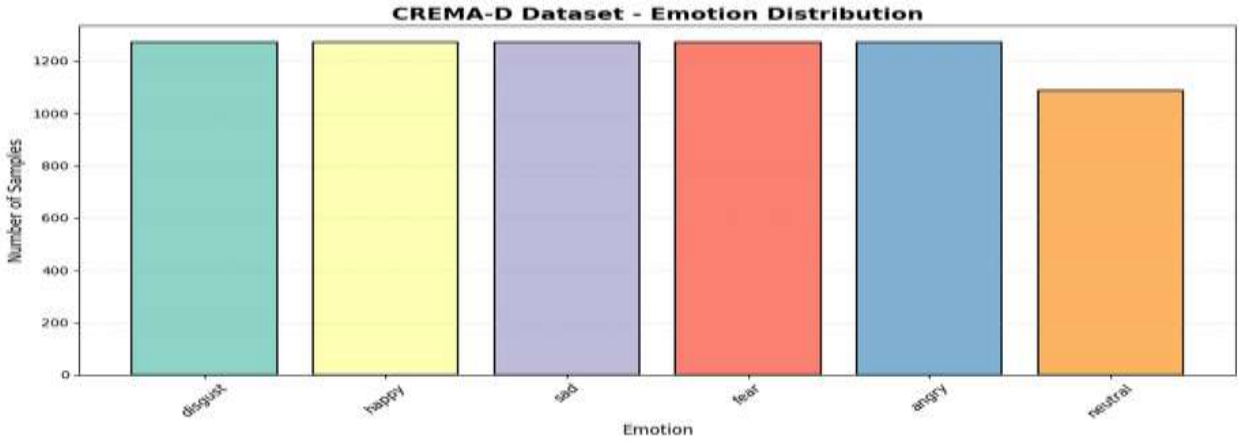

Figure 2: CREMA-D description

Instead of limiting the analysis to a subset of emotion categories, the suggested framework keeps all six original emotion classes from the CREMA-D dataset. This design offers a more thorough assessment of the suggested SER model using all of the emotion categories taken into account in CREMA-D. Speaker identity is maintained throughout data preparation in order to avoid leakage of information due to the speaker. Each recording is also matched with its speaker identifier, sentence identifier, emotion label, gender and intensity; this allows the development of a tight speaker independent evaluation protocol. Each file name is automatically parsed during metadata generation to obtain these attributes and a structured metadata table is created with the recording path, speaker ID, gender, sentence ID, emotion label, emotion intensity, numerical class label, and file name. Pre-processing, feature extraction, model training, and model assessment all rely on the metadata.

3.2 Pre-processing

Prior to training the deep neural network, preprocessing is the primary stage in deep learning (17). In the proposed Speech Emotion Recognition (SER) system, preprocessing plays a crucial role in organizing and preparing the input speech data in a consistent manner for subsequent processing steps. The goal of this stage is to produce high-quality metadata and to ensure that the speech recordings are maintained in a proper manner and the feature representations are standardized for training the models. The whole CREMA-D directory is scanned recursively to find all the available speech recordings. Only audio files stored in the Waveform Audio File Format (.wav) are saved; other files and formats are discarded if they are not audio files. This ensures that the processed data will always be the same. Then each existing name, that is, the name that is considered as valid, is automatically parsed based on the naming convention that has been used for the CREMA-D corpus. Important information of speakers, sentences, emotion label, and intensity of emotion are embedded in the filename. These attributes are automatically extracted and saved in a structured metadata table, avoiding manual annotation errors and ensuring the same structure of the datasets. All six emotion categories are maintained in the proposed framework: Angry (ANG); Disgust (DIS); Fear (FEA); Happy (HAP); Neutral (NEU); Sad (SAD). Therefore, each recording is categorized under one of these six emotion classes without missing out on any category. The identity of the speaker also is kept intact in the pre-processing. The

speaker index scheme used in CREMA-D is that odd speakers have odd speaker ID numbers and female speakers have even speaker ID numbers. This information is used in the metadata and subsequently to create the speaker independent training, validation and testing sets. A metadata entry is created for each recording with the file path, the name of the file, the speaker ID, a gender label, a sentence ID, an emotion label, an intensity level for the emotion, and a numerical label for the emotion class. This metadata is used to build on in the next feature extraction and experimental phases. The features are extracted and then normalized using the z-score normalization to enhance the stability of training and ensure that all acoustic features are comparable in scale. Normalization parameters (mean value and standard deviation) are calculated for the training set only and then uses for the validation and test set. This approach ensures that data does not have to be transferred across the network and still ensures uniformity of features for all data partitions. The speech recordings are not preprocessed for trimming. All the original utterances are retained to preserve the full emotional content of each recording and all the processing is done on the original speech signals.

3.3 Feature Extraction

The majority of researchers in the audio area and all researchers in the SER field are aware that extracting features that are precisely connected to the problem is the most significant and difficult issue (18). Once preprocessed, each speech recording is converted into a sequence of frame-level acoustic features representing the characteristics of the speech signal that could be used for determining emotion. The audio is broken down into overlapping frames with a frame length of 25 ms and a hop length of 10 ms. Each frame is windowed using a Hamming window before the calculation of the features, to reduce spectral leakage and enhance frequency analysis. Each frame is processed with a short set of 42 handcrafted acoustic features. This feature set consists of 13 Mel-Frequency Cepstral Coefficients (MFCCs), 13 Delta MFCCs, 13 Delta-Delta MFCCs, Pitch, Zero-Crossing Rate (ZCR) and Root Mean Square (RMS) Energy. The features are chosen because they carry complementary information regarding the spectral, temporal and prosodic aspects of emotional speech. Each utterance is represented by a feature matrix of variable length, $T \times 42$, with T being the number of speech frames depending on the length of the utterance, and 42 the number of features extract. The system checks whether the feature vectors have the right size before storing them, and it also checks for any NaN or infinite values. The emotion label, speaker ID, gender, sentence ID, intensity level and sequence length are then written along with each feature matrix, which is used for further model training and evaluation. The feature matrices generated are then fed into the proposed CNN1D-BiLSTM-Attention network, which affords the model the ability to learn discriminative patterns of emotion from both local acoustic information and from long-term temporal dependencies. Figure 3 illustrates the features used in the model for certain samples.

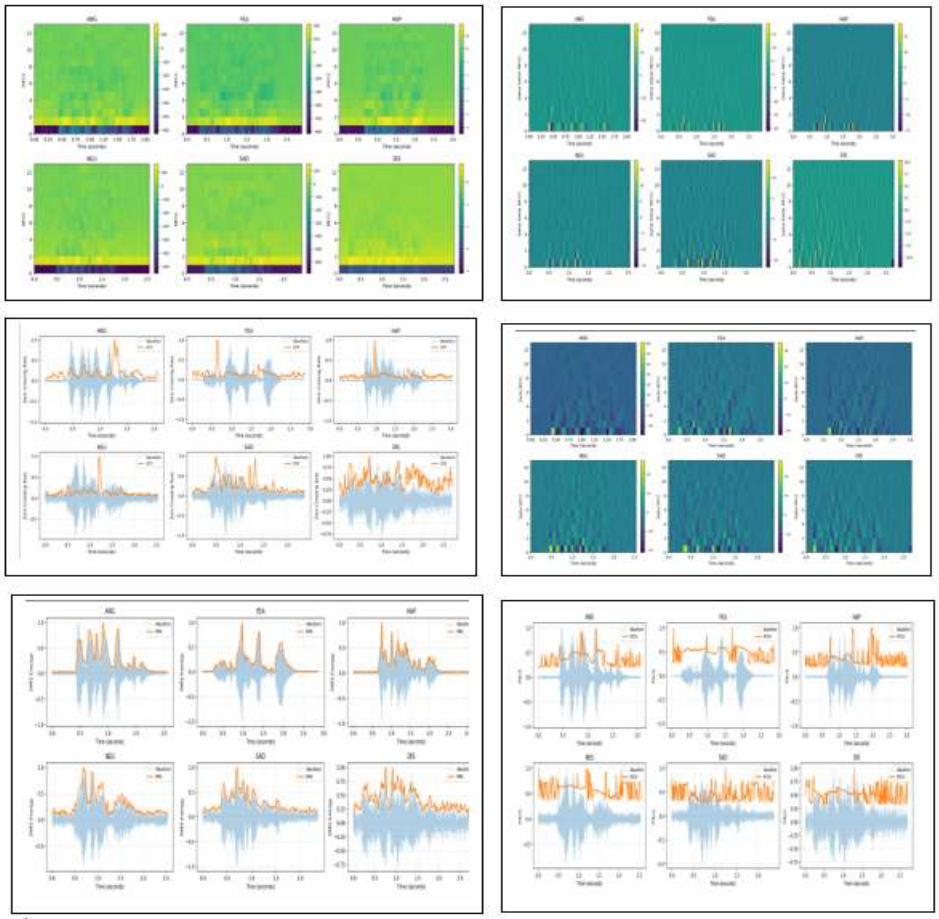

Figure 3: 1st column MFCC, Delta2 MFCC, and ZCR features, 2nd column Delta MFCC, Pitch, and RMS features

3.4 Proposed CNN1D-BiLSTM-Attention Architecture

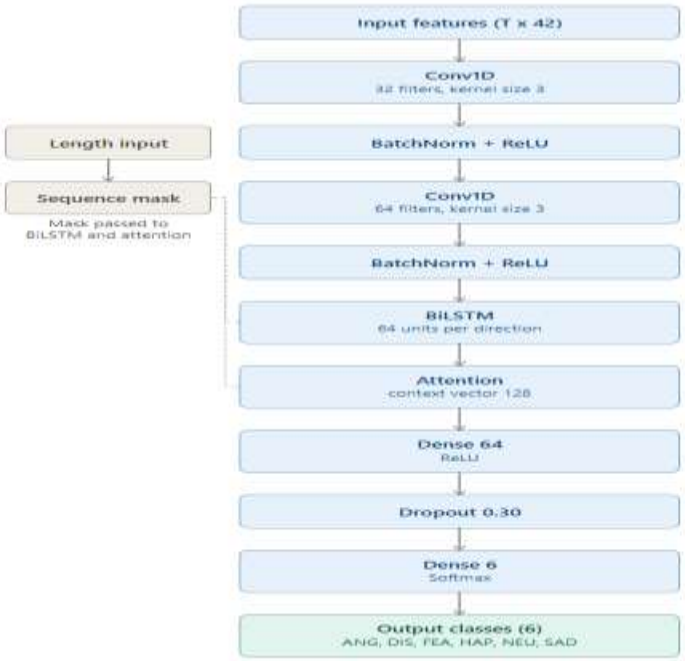

Figure 4: Overall Architecture of the Proposed CNN–BiLSTM–Attention Model.

The principles behind artificial intelligence are learning algorithms, enabling computers to learn from data and gradually become better at their job. These algorithms analyze vast datasets, identify patterns and relationships, and forecast future trends based on the data (19). In order to perform speech emotion recognition, a CNN-BiLSTM-Attention network is designed to capture both the local and temporal representations by learning the complementary representations from the extracted acoustic features. The model takes in variable-length sequences of 42-dimensional acoustic feature vectors. The 42-dimensional feature vector includes 13 MFCC, 13 delta-MFCC, 13 delta-delta MFCC, zero-crossing rate (ZCR), pitch, and root mean square (RMS) energy. Thus, an input utterance could be represented as:

$X=\{x1, x2, \ldots\ldots\ldots, x_T\}$, $x_T \in \mathbb{R}^{42}$.

Where X: Full input feature sequence of an utterance, T: Number of frames (time steps) in the utterance, $X_T \in \mathbb{R}^{42}$: 42 dimensional acoustic feature vector at frame t. The architecture comprises of two 1D convolutional layers, two batch normalization layers, two ReLU activation layers, a sequence masking mechanism, a bidirectional long short-term memory (BiLSTM) layer, an attention mechanism, a fully connected layer, dropout regularization and a final six class emotion classification layer. The summary of the number of parameters, the dimension of the output and input of each layer, and the architecture of the layer is shown in Table 2.

Table 2: Layer-Wise Architecture and Parameter Count of the Proposed Model

Layer	Input Shape	Output Shape	Parameters
Input	T×42	T×42	0
Conv1D-32	T×42	T×32	4,064
Batch Normalization 1	T×32	T×32	128
ReLU 1	T×32	T×32	0
Conv1D-64	T×32	T×64	6,208
Batch Normalization 2	T×64	T×64	256
ReLU 2	T×64	T×64	0
Sequence Masking	T×64	T×64	0
BiLSTM	T×64	T×128	66,048
Attention	T×128	128	129
Dense	128	64	8,256
Dropout	64	64	0
Output Dense	64	6	390
Total	—	—	**85,479**
Model size			**1.1 MB**

Convolutional Feature Extraction

The first convolutional layer contains 32 filters with a kernel size of three. It receives the T×42 feature sequence and produces a T×32 representation. The convolution operation could be expressed as:

$$\mathbf{Z_{t,j} = \sum_{k=0}^{k-1} \sum_{c=1}^{cin} Wk, c, j \; Xt + k, c + bj} \quad \mathbf{(1)}$$

Where $z_{t,j}$: pre-activation output of the j-th filter at frame t

K: kernel size, Cin: number of input channels, Wk,c,j: convolutional weight connecting input channel c at kernel offset k to output filter j, xt+k,c: input value at frame t+k, channel c, and, bj: bias of the j-th filter

$$\mathbf{P_{conv} = (K * C_{in} * C_{out}) + C_{out}} \quad \mathbf{(2)}$$

P_{Conv}: number of trainable parameters in the convolutional layer

C_{out}: number of output channels (filters).

For the first convolutional layer, P=(3×42×32)+32=4,064. Consequently, (T,42)→(T,32).

The convolutional output is subsequently normalized using batch normalization. For an input xx x, batch normalization is expressed as:

$\hat{x} = \frac{x-\mu B}{\sqrt{\sigma^2 B+\epsilon}}$ followed by $y=\gamma x+\beta$ (3)

Where x: input value to be normalized, $\hat{x}$: normalized value, μB: mean of the current mini-batch, $\sigma^2 B$: variance of the current mini-batch, ϵ: small constant added for numerical stability, γ,β: learnable scale and shift parameters, and, y: output of the batch normalization transform.

The first batch normalization layer contains 128 total parameters, including trainable and non-trainable statistics. The normalized features are then passed through the ReLU activation function: $f(x)=\max(0,x)$ (4)

f(x): ReLU activation output

The second convolutional layer increases the number of feature maps from 32 to 64. Its parameter count is P=(3×32×64)+64=6,208. Thus, the second convolutional block transforms the representation as (T,32)→(T,64). A second batch normalization and ReLU activation are then applied, preserving the temporal dimension and producing a final convolutional representation of T×64.

Variable-Length Sequence Handling

Because the utterances do not have identical durations, the model employs sequence masking to distinguish valid speech frames from padded positions. For an utterance containing Ti valid frames, the mask could be defined as $m_t = \begin{cases} 1, t \le Ti \\ 0, t > Ti \end{cases}$ (5)

Where mt: mask value at frame t (1 = valid frame, 0 = padded frame), Ti: number of valid (non-padded) frames in the i-th utterance.

The masking operation introduces no trainable parameters and ensures that padded frames do not contribute to the subsequent recurrent and attention computations. Therefore, the masked representation remains (T,64).

Bidirectional LSTM

The masked convolutional features are passed to a bidirectional LSTM to model temporal dependencies in both forward and backward directions. Each LSTM direction contains 64 hidden units, resulting in a combined output dimension of 128.

For each time step, the LSTM computes the forget, input, candidate, and output states as

$f_t = \sigma (W_f x_t + U_f h_{t-1} + b_f)$ (6)

$i_t = \sigma (W_i x_t + U_i h_{t-1} + b_i)$ (7)

$c_t = \sigma \, Tanh(W_c x_t + U_c h_{t-1} + b_c)$ (8)

$o_t = \sigma (W_o x_t + U_o h_{t-1} + b_0)$ (9)

where f_t, i_t, o_t: forget, input, and output gate activations at time step t,

ct: candidate cell state at time step t.

x_t: input vector at time step t.

$h_t - 1$: hidden state from the previous time step.

Wf,Wi,Wc,Wo: input-to-hidden weight matrices of the forget, input, candidate, and output gates.

U_f, U_i, U_c, U_o: hidden-to-hidden (recurrent) weight matrices of the forget, input, candidate, and output gates.

b_f, b_i, b_c, b_o : bias terms of the forget, input, candidate, and output gates.

σ(·): sigmoid activation function.

tanh(·): hyperbolic tangent activation function.

The cell state and hidden state are subsequently calculated as

$c_t = f_t \odot c_t - 1 + i_t \odot \tilde{c}_t$ (10)

and $h_t = o_t \odot \tanh (c_t)$ (11)

c_t: cell state at time step t.

$c_t - 1$: cell state from the previous time step.

h_t: hidden state (BiLSTM output) at time step t.

⊙: element-wise (Hadamard) product.

For an LSTM with input dimension D and hidden size H, the number of parameters is

$P_{LSTM} = 4[(D \times H) + (H \times H) + H]$,

With D=64 and H=64, $P_{LSTM} = 4[(64 \times 64) + (64 \times 64) + 64] = 33,024$.

Since the architecture is bidirectional, the total number of BiLSTM parameters is

$P_{BiLSTM} = 2 \times 33,024 = 66,048$. The resulting temporal representation is therefore (T,64)→(T,128).

Attention Mechanism

Although the BiLSTM generates a representation for every frame, not all frames contribute equally to emotion recognition. Therefore, an attention mechanism is applied to assign greater importance to emotionally informative frames. Let the BiLSTM output be

$H = \{h_1, h_2, \ldots, h_T\}$, where $h_t \in \mathbb{R}^{128}$

H: the set of BiLSTM hidden states across all time steps

$h_t \in \mathbb{R}^{128}$: BiLSTM hidden state at time step t

An attention score can be calculated as:

$e_t = Tanh (W_a h_t + b_a)$ (12)

where e_t: attention score at time step t, W_a: attention layer weight vector, b_a: attention layer bias term.

The attention weights are then obtained using the softmax function:

$$\alpha t=\frac{\exp(e_t)}{\sum_{j=1}^{T}\exp(e_j)} \quad (13)$$

where αt: normalized attention weight at time step t.

The final utterance-level representation is calculated as the weighted sum

$$C=\sum_{t=1}^{T}\alpha t\, ht \quad (14)$$

Where c: utterance-level context vector produced by the attention mechanism.

Consequently, the variable-length sequence (T,128) is converted into a fixed-dimensional representation 128. The implemented attention layer contains 129 parameters, corresponding to a 128-dimensional attention weight vector and one bias term.

Fully Connected and Dropout Layers

The 128-dimensional attention representation is passed to a fully connected layer containing 64 neurons. The operation is defined as

$$Z= Wh +b \quad (15)$$

Where z: pre-activation output of the dense layer, W: weight matrix, h: input representation to the layer, and b: bias vector.

The number of parameters is $P_{Dense}=(128\times64)+64=8{,}256$.

Where P_{Dense}: number of trainable parameters in the dense layer. Therefore, 128→64.

A dropout layer with a rate of 0.30 is subsequently employed to reduce overfitting. During training, a random subset of activations is dropped. The dropout operation can be represented as

$$\tilde{h} = \frac{m\odot h}{1-P} \quad (16)$$

Where $\tilde{h}$: output of the dropout layer, m: binary random mask applied during dropout

p=0.30: dropout rate. Dropout does not introduce trainable parameters.

Six-Class Emotion Classification

Finally, the 64-dimensional representation is passed to a fully connected output layer containing six neurons corresponding to the six emotion classes of CREMA-D: anger, disgust, fear, happiness, neutral, and sadness. The number of parameters in the output layer is

$P_{Output=(}64\times6)+6=390$, where P_{Output}: number of trainable parameters in the output layer.

The output logits are converted into class probabilities using the softmax function:

$$P(y=i|X)=\frac{\exp(zk)}{\sum_{j=1}^{6}\exp(zj)} \quad ,i=1, \ldots\ldots\ldots, 6 \quad (17)$$

Where P(y=i|X): predicted probability that input x belongs to class k, zk: logit for class k,

k: emotion class index, k=1,…,6. The final output is therefore, (64)→(6).

The complete transformation performed by the proposed model could be summarized as (T,42)→(T,32)→(T,64)→(T,128)→(128)→(64)→(6). The model contains a total of 85,479 parameters, including 85,287 trainable parameters and 192 non-trainable parameters when the final classifier is configured for six emotion classes.

3.5 Model Training

Once the frames-level acoustic feature sequences are extracted, the proposed CNN1D-BiLSTM-Attention model is trained with the extracted feature sequences. The input to the network is variable-length feature matrices, which could process speech utterances of different length without imposing a fixed sequence length on the network. In order to test the generalization capability of proposed framework, a speaker-independent partitioning strategy of data is used. All the speakers used for testing are not included in the training set and validation set, which means they are not part of the training or validation sets. This protocol is chosen to give a realistic evaluation of the robustness of this model in practical speech emotion recognition application. The network is trained with loss function: categorical cross-entropy, which is typically used in multi-class classification problems. The model parameters are optimized using the Adam optimizer with an initial learning rate of (3×10^{-4}). The batch size is set to (16) for all training, and the number of training epochs is set to (100). Early Stopping is used to stop overfitting and for better convergence, the validation loss is used. Training is automatically stopped if it fails to improve by in some way for (12) epochs. Additionally, the learning rate is decreased by 50% on the plateau for (4 epochs) using the (ReduceLROnPlateau) strategy with a minimum learning rate of (1×10^{-6}). This approach allows the model to converge efficiently and enhance the generalization accuracy. The model which has minimum validation loss during training is automatically saved and then evaluated at the end of training. The performance of the model is evaluated on the training, validation, and test sets; and the emotion predicted is the class having the highest Softmax probability of all the emotion classes.

4- Experimental Results and Discussion

In this section, the experimental results obtained with the suggested Speech Emotion Recognition (SER) model using the CREMA-D dataset, through a stringent speaker independent evaluation process are presented and discussed. The performance of the proposed model is evaluated using a number of additional classification measures such as accuracy, precision, recall, Unweighted Average Recall (UAR), and F1-score. The analysis focuses on the overall classification performance, as well as the emotion class recognition performance. In addition, the findings obtained are analyzed to identify the emotional categories that are accurately detected and those that remain challenging, and to evaluate the model's generalization to unseen speakers. Moreover, these results are compared with previous SER studies of the CREMA-D data set. The comparisons are carefully analyzed to make sure that the reported performance differences are significant and supported by science when there are variations in the number of emotion classes or experimental circumstances.

4.1 Experimental Setup

The proposed CNN-BiLSTM-Attention model is used to test the CREMA-D speech emotion data set of speech utterances from six emotion classes: Angry (ANG), Disgust (DIS), Fear (FEA), Happy (HAP), Neutral (NEU), and Sad (SAD). An evaluation protocol is used that is independent of the speaker, to provide a realistic assessment of the model. The CREMA-D dataset is split speaker independently into 80% training, 10% validation and 10% test sets at speaker level. Accordingly, 73 speakers are assigned to training, while 9 speakers were selected for validation (1037, 1061, 1049, 1065, 1027, 1078, 1026, 1010, 1014) and 9 speakers for testing (1013, 1051, 1009, 1003, 1035, 1058, 1030, 1080, 1038). Figure 4 presents the breakdown of the six emotion classes in training, validation and test sets. The speaker-independent partitioning approach is responsible for the slight variations in the class distribution among the three subgroups.

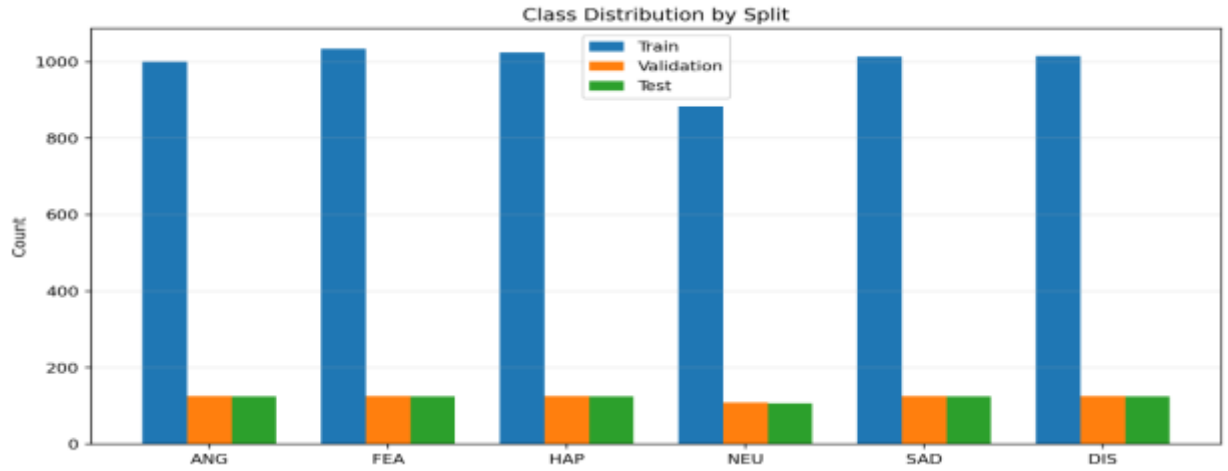

Figure 5: Distribution of the six emotion classes across the speaker-independent training, validation, and test sets.

The 42 features extracted at the frame level are represented as variable length sequences corresponding to each speech recording. The extracted features are 13 MFCCs, 13 MFCC DELTA, 13 DD MFCCs, PITCH, Zero Crossing Rate (ZCR), and Root Mean Square (RMS) Energy. Feature normalization is done using the StandardScaler technique before the model is trained. The normalization parameters (mean and standard deviation) are calculated only from the training set data, and then uses to normalize the validation set and test set data to prevent information leakage. The proposed model is implemented using the deep learning frameworks of TensorFlow and Keras. Model training is performed with Adam optimizer, learning rate 3×10^{-4} and a maximum of 100 epochs. To enhance convergence and minimize overfitting, early stopping with a patience of 12 epochs and ReduceLROnPlateau with a reduction factor of 0.5 are also used. The best model that has minimum validation loss in the training phase is chosen for the final testing. The same training configuration has been used in all the experiments to make the assessment of the proposed framework fair and repeatable.

Table 3: Experimental Setup and Training Parameters

Parameter	Value
Dataset	CREMA-D
Emotion Classes	6

Feature Dimension	42
Frame Length	25 ms
Hop Length	10 ms
Optimizer	Adam
Learning Rate	3×10^{-4}
Batch Size	16
Maximum Epochs	100
Early Stopping	12
LR Scheduler	ReduceLROnPlateau
Dropout	0.30

4.2 Evaluation Metrics

To assess the performance of the proposed model some important metrics are used to evaluate its classification accuracy in a comprehensive way. These metrics are referred to, as accuracy, precision, recall, The UAR, and the F1 score are especially important when assessing the success of SER system. Both the overall performance and emotion-specific classification results are taken into account.

Although accuracy determines the overall percentage of accurate forecasts for each class, it might be misleading when there is a class imbalance. Consequently, complementary criteria like the F1 score and UAR are also used for a more balanced evaluation (20).

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \quad (18)$$

Precision determines the overall percentage of accurate forecasts for each class, although as previously said, it may be misleading when there is a class imbalance. Consequently, for a more impartial evaluation, supplementary metrics like the F1 score and UAR are also included (16).

$$\text{precision} = \frac{\sum_{k=1}^{K} \frac{TPk}{TPk+FPk}}{K} \quad (19)$$

Recall and UAR assess the model's ability to correctly identify each actual instance of each emotion. High recall in SER ensures that emotional expressions are consistently recognized without overlooking relevant examples (21).

$$\text{recall} = \frac{\sum_{k=1}^{K} \frac{TPk}{TPk + FNk}}{K} \quad (20)$$

All emotional events should be identified, which requires recall, hence lowering. the possibility of ignoring important emotional signals. ASER's recall is improved when the high the reliability and effectiveness of the model in realistic variable speech context. By averaging across all emotion classes, the metric “UAR” offers a just one fair measure, independent of their frequency. In particular, it facilitates non-bias towards dominant aspects in ASER. emotions and guarantee equitable evaluation over the whole emotional spectrum (11).

$$\text{UAR} = \frac{1}{k} \sum_{k-1}^{k} \text{Recall}(k) \quad (21)$$

Where K is the number of emotion classes.

A crucial parameter for assessing SER models, particularly in cases of class imbalance, is the F1 score, which strikes a compromise between precision and recall. The model's total capacity to identify and accurately categorize emotional states is reflected in the average F1 score across emotions (F1 Score Total) (11).

$$F1 = 2\frac{\text{precision.recall}}{\text{precision+recall}} = F1 = 2\frac{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} * \frac{TP}{TP+FN}} \quad (22)$$

By using these criteria, we guarantee a thorough assessment of the model's performance, emphasizing both its general classification accuracy and its efficiency in managing characteristics particular to emotions.

4.3. Implementation Details and Material

The proposed model of Speech Emotion Recognition is implemented on a computing system based on an AMD Ryzen 9 12-core processor, 16 GB of RAM DDR5 memory, and an NVIDIA GeForce RTX 4080 graphics card with 16 GB of dedicated memory. The implementation is written in Python 3.12, TensorFlow 2.21.0 and Keras 3.15.0.

4.3 Experimental Results

In this section, the experimental results of the proposed Speech Emotion Recognition (SER) model are shown. The model is tested in an experimental protocol with a speaker independent protocol, training, validation and test sets are used. The following classes of emotions are considered in the experiments: Angry (ANG), Fear (FEA), Happy (HAP), Neutral (NEU), Sad (SAD), Disgust (DIS). The overall accuracy and Unweighted Average Recall (UAR) are used to assess the performance, and the confusion matrices are then examined to examine the classification performance of the model within individual emotion classes.

4.3.1 Model Selection and Training Behavior

The training and validation accuracy for each training epoch is displayed in Figure 5. Both graphs exhibit a general upward trend, suggesting that the emotion classification job is being learned with time. The validation accuracy achieved 75.34% at the chosen 45th epoch, which is utilized for the evaluation that followed.

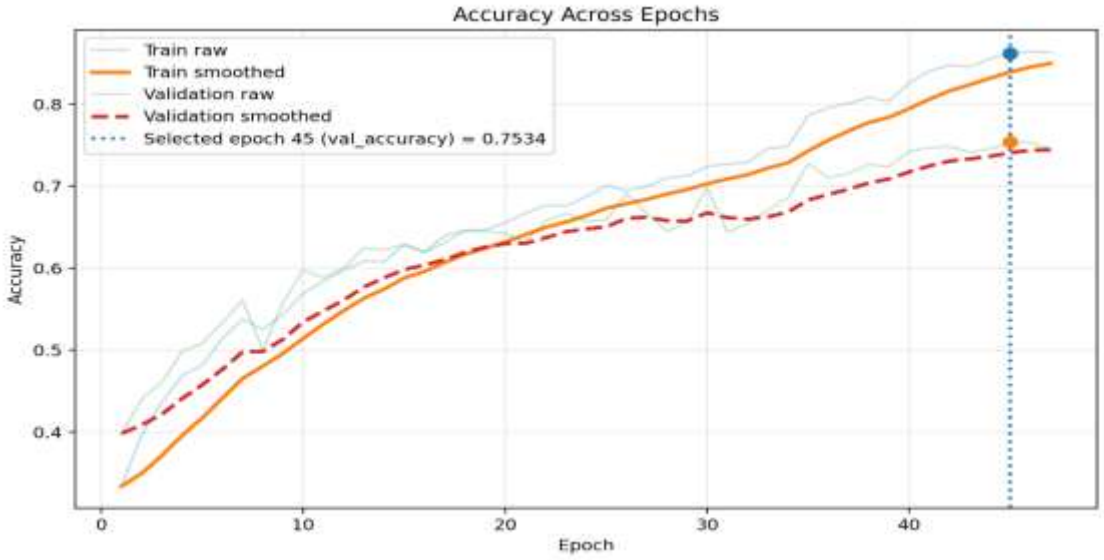

Figure 6: Training and validation accuracy across Epochs

The appropriate training and validation loss is shown in Figure 6. While the validation loss declines in the early training phases and stabilizes in the latter epochs, the training loss steadily declines. A degree of fitting to the training data is indicated by the growing discrepancy between training and validation performance in the later epochs.

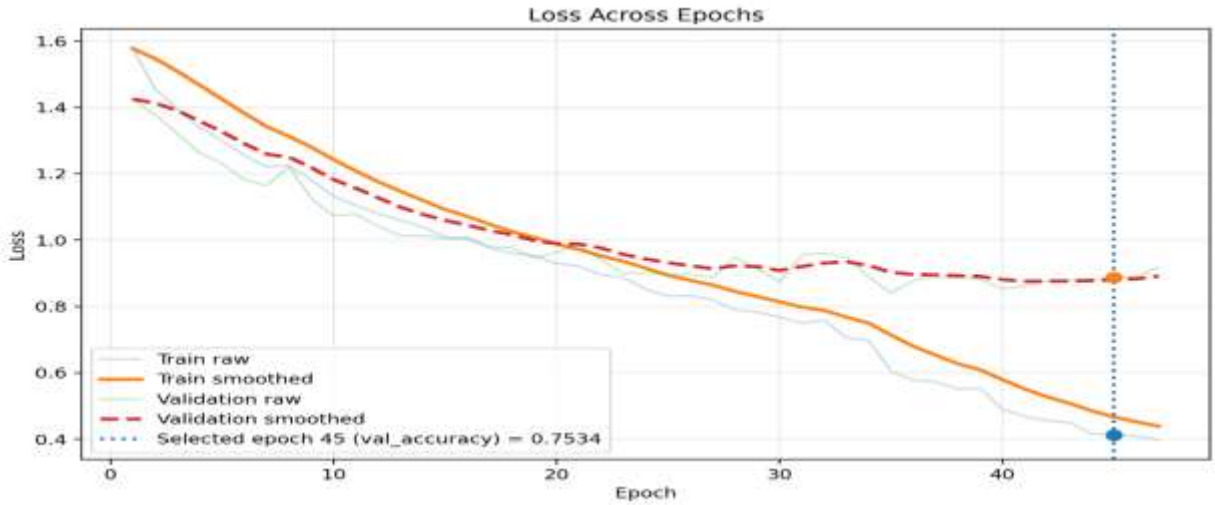

Figure 7: The loss in training and validation over epoch

4.3.2 Overall Classification Performance

The suggested model's overall classification performance on the training, validation, and test sets is shown in Table 4.

Table 4. Overall classification performance of the proposed SER model

Dataset	Accuracy (%)	UAR (%)
Training	87.44	87.69
Validation	75.34	75.53
Test	**71.72**	**71.95**

The suggested model obtains an accuracy of 87.44% and a UAR of 87.69% on the training set, as seen in Table 4. The model's accuracy and UAR on the validation set are 75.34% and 75.53%, respectively. The model's accuracy and UAR for the unseen test set are 71.72% and 71.95%, respectively. One emotion class does not significantly influence overall performance, as evidenced by the comparatively close Accuracy and UAR values across the three datasets. The model maintains comparatively balanced recognition performance throughout the six emotion categories despite differences in the number of samples within the classes, as seen by the fact that the difference between Accuracy and UAR on the test set is only 0.23 percentage points. Since the model must generalize to previously unheard speakers, a decline in performance from the training set to the validation and test sets are anticipated under a speaker-independent evaluation scenario. Thus, compared to the training performance, the test performance offers a more accurate representation of the model's capacity for generalization.

4.3.3 Per-Emotion Performance

Using Precision, Recall, F1-score, and AUC as the primary per-emotion evaluation metrics, Table 4 displays the classification performance of the suggested model for each of the six emotion classes on the test set.

Table 5: Per-Emotion Performance

Emotion	Support	Precision	Recall	F1-score	AUC
ANG	125	83.74%	82.40%	83.06%	96.43%
FEA	125	69.70%	73.60%	71.60%	90.74%
HAP	125	74.79%	71.20%	72.95%	94.79%
NEU	107	76.99%	81.31%	79.09%	97.03%
SAD	125	60.15%	64.00%	62.02%	89.13%
DIS	125	66.07%	59.20%	62.45%	89.02%

The suggested model obtains the greatest F1-score for Angry (83.06%), followed by Neutral (79.09%), according to the per-emotion data. Additionally, Angry has a high recall of 82.40%, meaning that the majority of the Angry samples are correctly recognized by the model. Strong discrimination between Neutral and the other emotion classes is demonstrated by Neutral's highest AUC of 97.03%. F1-scores for Fear and Happy are 71.60% and 72.95%, respectively. Sad and Disgust, on the other hand, has the lowest F1-scores (62.02% and 62.45%, respectively). Despite their very high AUC values of 89.13% and 89.02%, respectively, these results show that Sad and Disgust are the most difficult emotion classes for the suggested model.

4.3.4 Confusion Matrix Analysis

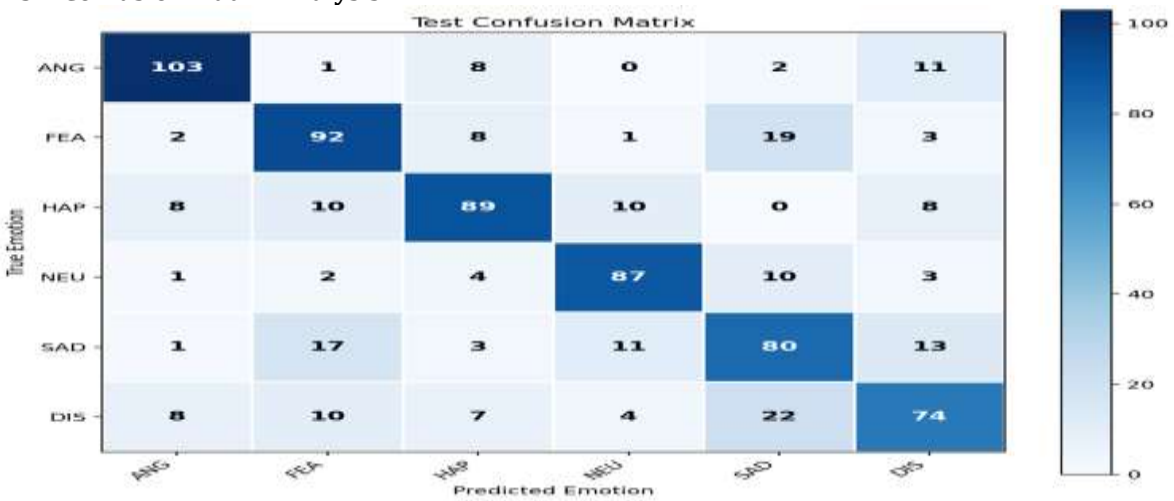

Figure 8: The suggested SER model's test confusion matrix

The confusion matrix in figure 5 offers more information about the suggested model's categorization behavior. The diagonal numbers show that a significant percentage of data for each of the six emotion categories are properly identified by the model. Angry has the most properly identified samples (103), followed by Fear (92), Happy (89), Neutral (87), Sad (80), and Disgust (74). The most noticeable misclassification is seen between sadness and disgust. In particular, 13 sad expressions are mistakenly labeled as disgust, and 22 disgust utterances are mistakenly classed as sad. This is the test set's largest individual misunderstanding. There is

also a significant discrepancy between Fear and Sad, with 19 Fear utterances being categorized as Sad and 17 Sad utterances as Fear. Additional noteworthy categorization errors are found for Happy, where ten utterances are categorized as Neutral and ten more as Fear. In addition, eight angry statements are classified as happy and eleven angry statements are classified as disgusted. The results confirm that the Fear– Sad, Sad – Disgust and, to a smaller extent, Happy – Fear/Neutral combinations are the ones that present the highest classification problem. Such ambiguity implies that the speech signals' acoustic features could not always offer enough unique information to distinguish between various emotional states, especially when assessing speakers who are unseen.

4.4 Comparison with Previous Studies

The proposed SER model is then evaluated in terms of the performance by using its output with the other similar models reported in the previous studies base on CREMA-D dataset. Besides the reported classification performance special attention is given to the data-splitting strategy and the speaker-independency of the evaluation protocol. This is significant because the utterances from the same speaker may be split between the training and test sets so that speaker-related information may be fed into the learning process and lead to overly optimistic estimates of the learning performance. Thus, comparisons are not made just based on the reported performance measures, but also in the context of the evaluation protocol, which differ between studies.

Table 6: Comparison of Deep Learning-Based SER Studies on CREMA-D

Study	Model / Method	Accuracy / Performance	Data Split Strategy	Speaker Independent	Data Leakage Risk	Confusion Matrix	Main Limitations
(5)	CNN + CBAM	WA: 68.30%; UA: 68.22%	80% Train / 10% Validation / 10% Test	No	speaker-overlap risk	Yes	Speaker overlap may be permitted by random split; no Precision, Recall, or F1-score are provided.
(3)	Meta-Learning Optimized CNN	83.28%	Random train–test split	No	speaker-overlap risk	No	Speaker overlap; no class-wise analysis; preprocessing prior to splitting
(4)	1D-CNN + Feature-Selected RF	RF: 69%; CNN: 61%	Not clearly reported	Not reported	Unclear	No	augmentation prior to splitting; ambiguous testing procedure; unreported confusion matrix
(7)	d-vector, x-vector, r-vector + MHA + AM-Softmax	UA: 80.25%; WA: 79.81%	Random 80/20 split	No	speaker-overlap risk	Yes	Overlap of speakers in training and testing sets
(22)	DCRF-BiLSTM	95.10%	Random 80/20 split	No	speaker-	Inconsistent	Speaker overlap,

					overlap risk		augmentation prior to splitting, and discrepancies in the confusion matrix
(12)	Raw-Waveform SER Architecture	69.72%	Random 80/20 split	No	speaker-overlap risk	Questionable	Potential speaker overlap and dubious sample counts during assessment
(9)	IJOA + CP-LSTM + Attention	99.12%	Not specified	Unclear	Unclear	No	Uncertain leakage prevention and an incomplete evaluation protocol
(11)	**CBAM-DenseNet121**	**Accuracy: 71.26%; UAR: 71.01%; F1: 71.25%**	Training / Validation / Test	Not explicitly reported	**Unclear**	**Yes**	Significant performance difference between train and validation; potential overfitting; lack of a clear speaker-independent methodology
(8)	**CNN–BiLSTM–Transformer**	**Acc:93.2 F1:92.5**	5 – folds samples from 7442 to 59,536 samples	Not explicitly reported	speaker-overlap risk	**Yes**	Overlap of speakers in training and testing sets
Proposed Model	**CNN1D-BiLSTM-Attention**	**Accuracy: 71.72%; UAR: 71.95%**	**Speaker-independent**	**Yes**	**Reduced risk**	**Yes**	

As shown in table 5 the suggested model is assessed using a speaker-independent protocol, which lessens the possibility of speaker overlap between the training and test sets and offers a more thorough evaluation of generalization to unseen speakers than a number of earlier studies that used random data splitting.

Discussion

The experimental results show that the proposed CNN-BiLSTM-Attention network could achieve a test accuracy of 71.72% and a UAR of 71.95% in the task of recognizing six speech emotions with a strict speaker-independent evaluation protocol. The similarity of these two values (-0.23 percentage points) suggests that there are not one or two emotion classes that dominate the performance and that the model spreads its recognition capability relative evenly across all six emotion classes. The architecture's design leads directly to this balance: The CNN layers derive local spectral cues, while the BiLSTM layers capture long-range temporal context in forward/backward directions, and the attention mechanism reweights the frames based on their emotional content rather than any particular feature type that may be most sensitive to a particular emotion.

Why the difference between training, validation and test performance. The training accuracy of 87.44% and the validation accuracy of 75.34% and test accuracy 71.72% would not indicate poor learning; rather, it is a natural consequence of a speaker independent partitioning strategy used in this work. The validation and test sets consist of speakers with which the model has not seen anything during training, and therefore could not rely on speaker-specific acoustic features (such as an individual's base pitch, vocal timbre or speech rate) that a model trained on a random split of the utterances could have implicitly memorized. A further indication of training to speaker specific patterns in the training data is the increasing separation between the training and validation curves in the later epochs, which is the exact phenomenon that speaker-independent evaluation is designed to reveal and not mask. This is because the test performance is numerically poorer than the training performance, but it is regarded as the more reliable measure of the real-world performance of the model over new speakers in a deployment environment.

Why Angry and Neutral are most successful, angry obtained the highest F1-score (83.06%), recall (82.40%), and AUC (96.43%), while Neutral achieved the highest AUC overall (97.03%) and the second-highest F1-score (79.09%). These two emotions are at opposite ends of the arousal dimension: Angry speech is generally characterized by high energy, higher pitch, faster unpredictable zero-crossing rate, while Neutral speech is characterized by comparatively flat prosody, low pitch variability, and stable energy contours. Since the 42-dimensional feature set employed in this study is directly related to these prosodic and energy related properties, the acoustic representation of the Angry and Neutral emotions have more distinct separation properties from the other four emotions, making them easier to separate from the other emotions and from the other classes for the CNN-BiLSTM-Attention model.

Why the classes of sadness and disgust are the most difficult ones, in contrast, in both cases, F1 scores are lowest for classes Sad (62.02%) and Disgust (62.45%), while Disgust class has the lowest recall (59.20%) with high AUC values (89.13% for Sad and 89.02% for Disgust). Such a paradoxical result (low F1/recall and high AUC) indicates that the model may be capable of distinguishing these emotions comparatively well from the rest of the emotions, however it performs not so well specifically at the decision boundary between Sad and Disgust. However, acoustically, this is expected: Sadness and Disgust are voiced with similar levels of vocal energy, lower pitch, and fewer articulations per second than the more clearly distinguished high-arousal emotions, Angry for example, in acted emotional speech corpora like CREMA-D. Most of the handcrafted features employed in this work are prosodic and energy based, which are less useful in capturing the more subtle qualitative variations (e.g., breathiness or vocal creak) that usually distinguish Sad from Disgust. This effect is also observed under the speaker independent protocol; in this case, the model is unable to make up for this acoustic overlap by learning speaker-specific cues, which is reflected in the confusion matrix results with high cross-class misclassification between Sad and Disgust.

The importance of not making direct comparisons to earlier CREMA-D studies. Comparisons with previous CREMA-D based studies ought to be made with that protocol in mind, rather than solely on the accuracy of the results of the study. Many of the earlier papers reported their results with random splitting of data at the utterance level: utterances from the same speaker might be in the training set as well as the test set, or it is not clearly specified whether speaker independence is imposed or not. With such splitting strategies, the model is able to learn implicitly speaker-identifying features during training that inflates the performance in a manner unrelated to the performance on unseen speakers. This is the reason the results of the present study could not be directly compared to those studies claiming higher accuracy numbers without a speaker independent protocol; the two numbers are not quantifying the same underlying capability.

Comparison with CBAM-DenseNet121. The CBAM-DenseNet121 model is the most comparable among the reviewed studies, as it also reports similar accuracy (71.26%), UAR (71.01%), and F1-score (71.25%) on the same set of data. The only slight numerical improvement that the proposed model has over the existing model in terms of accuracy and UAR is a difference of 0.46 and 0.94 percentage points, respectively, but it is not clear whether the evaluation protocol used in that study is speaker independent or not, and so this small number ought not to be interpreted as evidence of clear architectural superiority, but rather as a comparison between two models being evaluated under potentially different degrees of evaluation rigour.

Explanation of the properties of the proposed architecture. This global trend of more accurate results for high-arousal/flat-prosody emotions and less accurate results for low-arousal emotions acoustically similar to the high-arousal emotions is a direct consequence of the design choices that are made in the proposed architecture. The CNN component is good at recognizing short, localised spectral bursts, hence its role in the high Angry recognition. The BiLSTM models the entire temporal profile of an utterance, which is especially important for Neutral speech, where there is no strong local profile variation over time but this is in fact an informative pattern. The attention mechanism also underscores those of the speaker's frames that are most emotionally salient, which could be useful if emotion is encoded in part of an utterance, but not in the entirety of it. This same architecture is less discriminative in the case that two emotions, e.g., Sad and Disgust, share the same emotion prosodic and emotional energy pattern across the whole utterance, because there is not a highly emotion specific segment to which the attention layer is applied.

The results overall show that the proposed framework could provide competitive and, more significantly, balanced SER performance in a significantly more challenging speaker independent evaluation setting. Several earlier studies report on higher raw accuracy values, but these are not directly comparable without taking into account the evaluation and data-splitting procedures of the different studies, some of which involve a significant likelihood of speaker leakage. Therefore, the reported accuracy (71.72%) and UAR (71.95%) ought not to be seen only as the performance metrics but rather as a more realistic and interpretable estimate of the actual performance of the proposed CNN-BiLSTM-Attention framework upon actual deployment with unseen speakers.

Limitations

Although the proposed CNN-BiLSTM-Attention framework has achieved competitive and balanced performance, there exist some limitations. The first is the fact that the proposed model is based purely on a 42-dimensional handcrafted acoustic feature set, namely MFCC, delta MFCC, delta-delta MFCC, pitch, ZCR and RMS energy. This compact representation, while important to the lightness of the model, is mostly prosodic and energy-based and does not explicitly encode more subtle, spectral or voice quality (e.g. jitter, shimmer, harmonic-to-noise ratio) aspects. This is also likely to be a cause of the low performance of the model to distinguish emotions that are acoustically similar, such as Sad and Disgust, which had the lowest F1-scores, namely 62.02% and 62.45% respectively, and the highest number of overlaps in the confusion matrix analysis.

Second, the study uses one single database namely CREMA-D, including acted, sentence-level emotional speech with professional actors in a studio environment in a controlled way. The reported results might not be directly applicable, however, to spontaneous emotional speech, typically where there is background noise, overlapping speech, less overt or mixed emotions, and higher linguistic and prosodic variation than is present in acted speech corpora.

Third, in this work the strict speaker-independent evaluation protocol is used, which better reflects the generalization properties of the speech recognition system than splitting it randomly at the level of utterances, but this also leads to a relatively small number of speakers being available for validation and testing, 9 speakers for each. This makes the performance estimates per emotion somewhat less statistically robust and could result in less statistical stability in the reported metrics if they are sensitive to the specific speaker who is in each partition.

Thirdly, the study's comparison to previous CREMA-D based studies is limited due to the different methods used to report the assessment protocol. Not all of the studies uses for comparison that makes it clear whether speaker independence is enforced; hence direct numerical comparisons of accuracy or UAR should not be fully judged or interpreted as a completely fair or valid comparison.

Lastly, the current framework lacks the possibility of using complementary modalities such as facial expressions or textual/linguistic content, and does not use transformer-based, self-supervised speech representations (e.g., HuBERT or Wav2Vec2) that, although having demonstrated greater raw accuracy in some SER benchmarks recently, are more computation-intensive. In this research, the lightweight properties of the proposed model and the possible accuracy improvement that could be provided by such larger models is not directly quantified.

Conclusion

The proposed model in this study is a lightweight hybrid CNN1D-BiLSTM-Attention (CNN-BL) model for SERR, which is tested on the CREMA-D dataset with six emotion classes: Angry, Disgust, Fear, Happy, Neutral, and Sad. The proposed architecture consists of convolutional layers to extract local patterns of the spectral information, followed by a bidirectional LSTM to capture the long-range temporal dependencies in both

directions of an utterance, and an attention mechanism to underscore the most emotion-charged frames per utterance using a small 42-dimensional handcrafted acoustic feature set that includes MFCCs, delta MFCCs, delta-delta MFCCs, pitch, zero-crossing rate and RMS energy.

The proposed model performance is evaluated using a strict speaker-independent evaluation protocol where the speakers used for the test are not involved in the training and validation. The test accuracy of the proposed model is 71.72% and UAR of the model is 71.95%, which shows that the model has performed well for this evaluation set, as the gap between test accuracy and test UAR is just 0.23 percentage points, which indicates a balance performance across emotion classes rather than a dominant class. As has been well reported in the SER literature, there are two most easily recognized emotions: Angry and Neutral; and two most difficult emotions: Sad and Disgust, the latter two having the similar acoustic properties. The comparison with previous works, which adopted the CREMA-D, also shows that the proposed framework gives a good trade-off between recognition performance and computational complexity, without using large-scale transformer-based architectures. It also points to a methodological problem in the existing SER literature, namely that most of the previous CREMA-D research uses random splitting of the data by utterances, which could cause speaker leakage between the training and test sets, and could overestimate the results. This work explicitly defines speaker independence and thus offers a more realistic and reproducible benchmark for assessing model generalization to unseen speakers.

Overall, the results validate the effectiveness of a lightweight CNN-BiLSTM-Attention architecture with a compact handcrafted feature representation to obtain competitive, well-balanced speech emotion recognition performance under realistic, speaker-independent deployment conditions. Future research could incorporate more acoustic descriptors to participate in the voice quality component of the framework, Test the model on spontaneous, and multilingual emotional speech datasets, as well as multimodal fusion with text or visual features, and compare and contrast the proposed light model design with larger transformer based or self-supervised speech representation models within the same strict speaker-independent evaluation procedure.

Conflict of Interest

The authors affirm that the work presented in this study is not impacted by any known competing financial interests or personal relationships.

Funding

This research received no external funding.

Ethics declaration

Because this study uses the publicly available CREMA-D dataset and does not directly recruit or collect data from human participants, ethical approval is not necessary.

Data Availability

The CREMA-D dataset makes the data used in this investigation accessible to the general audience. No fresh data was gathered for the study. The original dataset source is where the data can be found.

References

1. Nasir RJ, Abdulmohsin HA. Noise Reduction Techniques for Enhancing Speech. Iraqi Journal of Science. 2024:5798-818.
2. George SM, Ilyas PM. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. Neurocomputing. 2024;568:127015.
3. Ottoni LTC, Ottoni ALC, Cerqueira JdJF. A deep learning approach for speech emotion recognition optimization using meta-learning. Electronics. 2023;12(23):4859.
4. Rezapour Mashhadi MM, Osei-Bonsu K. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. PloS one. 2023;18(11):e0291500.
5. Dal Ri FA, Ciardi FC, Conci N. Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks. IEEE Access. 2023;11:116638-49.
6. Bhanbhro J, Memon AA, Lal B, Talpur S, Memon M. Speech emotion recognition: Comparative analysis of CNN-LSTM and attention-enhanced CNN-LSTM models. Signals. 2025;6(2):22.
7. Jakubec M, Lieskovska E, Jarina R, Spisiak M, Kasak P. Speech emotion recognition using transfer learning: Integration of advanced speaker embeddings and image recognition models. Applied Sciences. 2024;14(21):9981.

8. Tatapudi SR, Suma GJ. A hybrid CNN–BiLSTM–transformer fusion model for emotion recognition from partial and disordered speech. Discover Computing. 2026;29(1):363.
9. Paramasivam R, Lavanya K, Divakarachari PB, Camacho D. A Robust Framework for Speech Emotion Recognition Using Attention Based Convolutional Peephole LSTM. International Journal of Interactive Multimedia and Artificial Intelligence. 2025;9(4):45-58.
10. Poorna S, Menon V, Gopalan S. Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition. Biomedical Signal Processing and Control. 2025;100:106967.
11. Kahhoul ZS, Terki N, Benaissa I, Aldwoah K, Hassan E, Osman O, et al. Mathematical analysis and performance evaluation of CBAM-densenet121 for speech emotion recognition using the crema-D dataset. Applied Sciences. 2025;15(17):9692.
12. Kilimci ZH, Bayraktar Ü, Küçükmanisa A. Evaluating raw waveforms with deep learning frameworks for speech emotion recognition. Multimedia Tools and Applications. 2025;84(36):45119-49.
13. Alhussein G, Ziogas I, Saleem S, Hadjileontiadis LJ. Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. Artificial Intelligence Review. 2025;58(7):198.
14. Chakhtouna A, Sekkate S, Adib A. Speech emotion recognition: A systematic mega-review of techniques and pipelines. Information Fusion. 2026:104161.
15. Portal F, De Lope J, Graña M. A performance benchmarking review of transformers for speaker-independent speech emotion recognition. International Journal of Neural Systems. 2025;35(10):2530001.
16. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing. 2014;5(4):377-90.
17. Al-Atroshi SJA, Ali AM. Facial Expression Recognition Based on Deep Learning: An Overview. Iraqi Journal of Science. 2023:1401-25.
18. Abdulmohsin HA. A new proposed statistical feature extraction method in speech emotion recognition. Computers & Electrical Engineering. 2021;93:107172.
19. Alzaki A, Al-Tamimi M. An Analytical Study on the Most Important Methods and Data Sets Used to Identify People Through ECG: Review. Journal of Applied Engineering and Technological Science (JAETS). 2024;5:842-59.
20. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern recognition. 2011;44(3):572-87.
21. Fahad MS, Ranjan A, Yadav J, Deepak A. A survey of speech emotion recognition in natural environment. Digital signal processing. 2021;110:102951.
22. Chowdhury SY, Banik B, Hoque MT, Banerjee S. A Novel Hybrid Deep Learning Technique for Speech Emotion Detection using Feature Engineering. arXiv preprint arXiv:250707046. 2025.