applied sciences

MDPI

*Article*

# Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network

**Ala Saleh Alluhaidan [1], Oumaima Saidani [1,*], Rashid Jahangir [2], Muhammad Asif Nauman [3] and Omnia Saidani Neffati [4]**

[1] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[2] Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari 61100, Pakistan; rashidjahangir@cuivehari.edu.pk
[3] Department of Computer Science, University of Engineering and Technology, Lahore 54890, Pakistan; asif.nauman.uet@gmail.com
[4] Computer Science Department, College of Sciences and Arts in Sarat Abida, King Khalid University, Abha 64734, Saudi Arabia; oneffati@kku.edu.sa
* Correspondence: ocsaidani@pnu.edu.sa

**Abstract:** Speech emotion recognition (SER) is the process of predicting human emotions from audio signals using artificial intelligence (AI) techniques. SER technologies have a wide range of applications in areas such as psychology, medicine, education, and entertainment. Extracting relevant features from audio signals is a crucial task in the SER process to correctly identify emotions. Several studies on SER have employed short-time features such as Mel frequency cepstral coefficients (MFCCs), due to their efficiency in capturing the periodic nature of audio signals. However, these features are limited in their ability to correctly identify emotion representations. To solve this issue, this research combined MFCCs and time-domain features (MFCCT) to enhance the performance of SER systems. The proposed hybrid features were given to a convolutional neural network (CNN) to build the SER model. The hybrid MFCCT features together with CNN outperformed both MFCCs and time-domain (t-domain) features on the Emo-DB, SAVEE, and RAVDESS datasets by achieving an accuracy of 97%, 93%, and 92% respectively . Additionally, CNN achieved better performance compared to the machine learning (ML) classifiers that were recently used in SER. The proposed features have the potential to be widely utilized to several types of SER datasets for identifying emotions.

**Keywords:** speech emotion recognition; feature fusion; MFCCs; time-domain; convolutional neural networks



**Citation:** Alluhaidan, A.S.; Saidani, O.; Jahangir, R.; Nauman, M.A.; Neffati, O.S. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. *Appl. Sci.* **2023**, *13*, 4750. https://doi.org/10.3390/app13084750

Academic Editors: Ya Li, Kai Yu and Yan Song

Received: 10 March 2023
Revised: 3 April 2023
Accepted: 9 April 2023
Published: 10 April 2023



## 1. Introduction

Speech is the natural and widespread method of human communication and carries both paralinguistic and linguistic information. The linguistic information includes the context and language of the speech, while paralinguistic information includes the gender, emotions, age, and other unique attributes of the human. Several studies have shown that audio signals can be a simple mean to establish a connection between machines and humans [1]. Nonetheless, this involves the machine becoming familiar with the human voice so that the machine can predict the emotion, similar to humans. This has led to a growing interest in the area of SER, which involves identifying the emotions of speakers from their voices.

SER is a crucial area of research that has various applications in the field of call centers [2,3], human–computer interaction (HCI) [4], automatic translation systems, driving a vehicle [5,6], and in healthcare, where patient emotional states are identified through voice, and necessary facilities are provided [7,8]. However, as individuals have different

speaking styles, and cultural backgrounds, the selection of relevant acoustic features becomes complex and challenging. Currently, voice features used for SER are categorized into spectral, continuous (formants, energy, pitch), Teager Energy Operator, and qualitative (voice quality) features [9]. Nevertheless, these expert-driven acoustic features typically rely on the experience of area experts, representing low-level features that are not helpful for recognizing emotions in complex situations. In summary, the major drawbacks of expert-extracted features are:

- The ability to recognize emotional declines in complex situations, such as inter-speaker differences, variations in expressions, and environmental factors [10];
- The expert-driven features require significant time, money, and human expertise in order to train ML classifiers to enhance the efficiency [11];
- Currently, there is no established algorithm available to extract the ideal features to identify emotions [12].

In order to address the challenges mentioned above, it is necessary to implement effective techniques that can extract emotionally relevant and significant features for SER. To this end, various studies have proposed techniques that involve automatic feature extraction from voice signals. For example, a study [13] utilized a single-layer CNN to derive automatic feature, while another study [14] implemented a CNN with two convolutional layers (CLs) followed by a Long Short-term Memory (LSTM) layer for a SER system. However, shallow architectures such as single-layer and double-layer CNNs may not be able to learn salient features. In Ref. [15], a deep CNN was utilized to derive the discriminative frequency features by employing a rectangular filter along with a customized pooling technique for the SER. The suggested deep CNN was trained on the derived features from audio data.

In this study, a novel approach for SER is proposed, which combines the MFCCs and time-domain features derived from each audio signal in dataset. The proposed approach consists of four main components: (1) data collection (2) feature extraction, (3) model training, and (4) prediction, as shown in Figure 1 . In the feature extraction stage, the traditional features including MFCCs, and time-domain are extracted. In the feature fusion stage, the extracted features are concatenated. In the final stage, a CNN model that comprises three 1D CLs, following an activation, dropout, and max-pooling layers, as well as a fully connected (FC) layer, is used for SER. To estimate the performance of methodology, three publicly datasets: Emo-DB, Surrey Audio-Visual Expressed Emotion (SAVEE), and The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) were employed. It is worth noting that the expression of emotion varies among different speakers due to factors such as gender, cultural background, and accent. The proposed method achieved better recognition performance for SER. In summary, the main contributions of this research include an algorithm for extracting emotionally relevant and robust features by combining frequency and time-domain (MFCCT) for SER and implemented a lightweight CNN that obtained improved recognition results over the baseline SER methods.

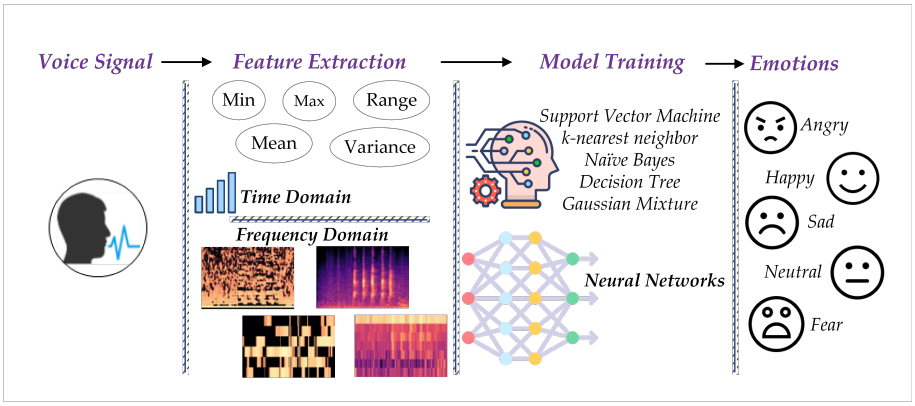


**Figure 1.** An illustration of the SER process using conventional approaches.

The rest of the paper is organized as follows. Section 2 provides a brief overview of related work on SER, including feature extractions and classification approaches. Section 3 describes the proposed approach in detail, including feature extraction, feature fusion, classification, and evaluation parameters. Section 4 provides the experimental results and discussion, and Section 5 concludes this paper.

## 2. Related Work

The categorization of SER systems in the literature involves two main stages. The initial stage is to derive a suitable and unique feature from the audio signals, while the second stage is to select the classifier that can accurately identify emotions. A brief overview of both stages for SER is provided in the following subsections.

### 2.1. Feature Extraction

Two approaches are widely used in the literature to derive features from audio signals. The first approach involves dividing the signal into voice frames of a specific duration and extract low-level features from each single frame. The features used for SER are generally categorized into four categories: linguistic, contextual, acoustic, and hybrid. Acoustic features are the most popular and effective features used in SER. These comprise voice quality features (jitter, shimmer, first three formants, and harmonic to noise ratio), prosodic features (pitch, loudness, and duration), as well as spectral features (MFCCs) [9]. Ref. [16] extracted 6 distinct acoustic features, namely energy, amplitude energy, log energy, formants, pitch, 10-order Linear Predication Cepstral Coefficients (LPCCs), and 12-order MFCCs from each frame. Afterwards, the Hidden Markov Model (HMM) was used to identify emotions, which obtained an accuracy of 76.1% on the Beihang University Mandarin dataset.

Recently, the use of Deep Learning (DL) architectures for SER has gained interest due to the rapid development in computational technology and challenges faced by traditional ML algorithms in handling large datasets. For this, Akçay and Oğuz [17] provided a comprehensive review on SER and highlighted three effective approaches to identify emotions. The review further reported the importance of employing optimized classification algorithms to enhance the robustness of SER system. In another study, Jahangir et al. [9] reviewed DL techniques with characteristics, pros, and cons. This study also classified DL techniques into discriminative (Recurrent Neural Network, CNN), generative (Deep Belief Network, restricted Boltzmann machine, and deep autoencoders), and hybrid techniques. Additionally, the authors investigated the optimization of DL methods for SER. Ref. [18] proposed a method to derive multimodal feature representations using multiple CNNs. The reported method derived suitable features, including two-dimensional log Mel-spectrograms, and three-dimensional temporal dynamics from voice. The multiple CNNs (1D, 2D, and 3D) were trained separately, and finally a fusion based on scores fusion was performed to identify emotions. The study achieved a weighted accuracy of 35.7% on AFEW5.0, and 44.1% on BAUM-1 datasets. In another study [19] a DL technique was proposed for emotion prediction from audio files. The technique involved preprocessing raw signals to derive features such as energy, pitch, and MFCCs followed by the selection of relevant features using a feature selection method. A CNN model was trained for classification. The proposed technique achieved an accuracy of 93.8% on the Emo-DB dataset, outperforming the baseline k-NN classifier's accuracy of 92.3%. Ref. [20] presented a novel SER framework that employed ID CNN and combined five features (Chromagram, MFCCs, Mel-Spectrogram, Tonnetz, and Contrast) as the input. The model was evaluated on IEMOCAP, RAVDESS, and EMO-DB datasets and achieved good accuracy. To improve the accuracy and reduce computational cost and processing time, Ref. [21] proposed a novel SER method. The study obtained a useful sequence from signals using the K-means clustering algorithm and generated spectrograms using the Short-time Fourier transform (STFT) method. A ResNet CNN model was then utilized to extract relevant features from spectrograms, which were given to the BiLSTM model to predict the emotions. The pro-

posed technique achieved 77.0% accuracy for the RAVDESS dataset, 72.3% for IEMOCAP, and 85.6% for EMO-DB.

Badshah, et al. [22] reported a CNN-based model to automatically derive features from audio files and identify emotions. The authors generated spectrograms with a 50% overlap from signals and trained two CNNs with different kernel sizes and pooling techniques on resized spectrograms. The results indicated that rectangular kernel and max-pool operations are the most efficient techniques for SER. Ref. [23] presented a technique to combine handcrafted and automatic features. The method was evaluated on well-known datasets, including IEMOCAP, EMO-DB, and RADVES. Initially, acoustic features and spectrograms were derived from audio files. Then, data enhancement was used to create additional training data. Next, deep features were derived from spectrogram images using pre-trained SqueezeNet, VGG16, DenseNet201, ResNet18, ResNet50, and ResNet101 CNNs. Finally, acoustic, and deep features were combined to get hybrid features, and linear SVM was used to identify emotions. The authors achieved 85.4% accuracy for IEMOCAP, 90.1% for EMO-DB, and 79.4% for RADVES dataset.

### *2.2. Classification Methods*

Various ML classifiers have been utilized by researchers to predict emotions from audio files. These classifiers include Multilayer Perceptron (MLP) [24], Random Forest (RF) [25], Support Vector Machine (SVM), hidden Markov model (HMM) [26], Gaussian mixture model (GMM), and k-NN. These classifiers are commonly employed for speech-related problems such as emotion recognition [11,27,28], and speaker identification [29–31]. This study employed RF, J48, SVM, NB, and k-NN to recognize emotions.

## 3. Materials and Methods

In this section, the methodology (Figure 2) utilized for emotion recognition is elaborated in detail. Initially, emotional audio files were collected for experimentation purposes. Subsequently, various valuable features were obtained from these collected audio files to create a vector called master feature vector (MFV). This MFV was then given to a CNN to develop the SER model. To assess the performance of SER models, two metrics—AUROC (Area Under the Receiver Operating Characteristics) and overall accuracy were employed. Lastly, the performance of proposed SER model was evaluated by comparing it with existing SER baseline techniques, using a percentage split approach, where 80% of the data was utilized to train the CNN model and the remaining 20% data was utilized to test the model [32,33]. Existing research [34] have concluded that the optimum results are achieved if 20 to 30% of the data is utilized for model testing and the remaining 70 to 80% of the data is used to train the model. Further details regarding these methods are presented in the following sections.

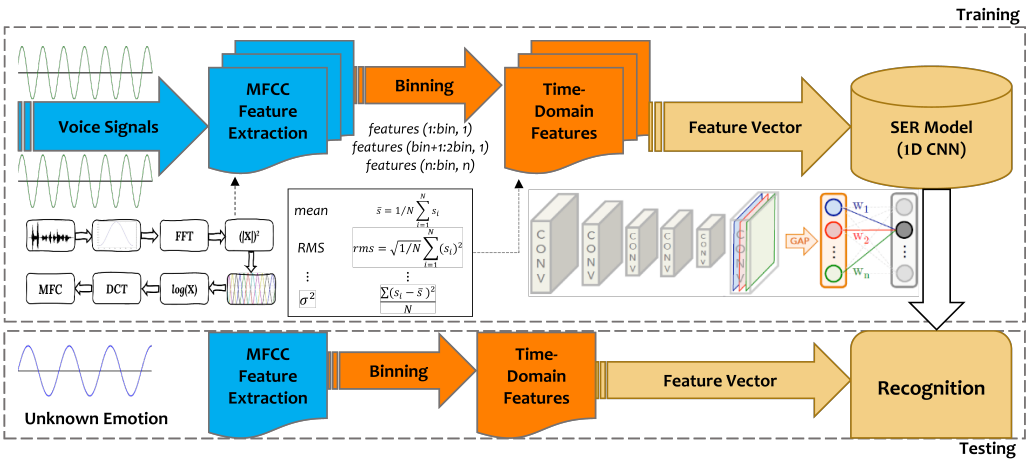


**Figure 2.** Proposed research methodology for SER.

### *3.1. Datasets*

EMO-DB: The proposed methodology begins by collecting publicly available datasets. For this purpose, EMO-DB [35] was chosen as one of the datasets due to its widespread usage in the area of SER. EMO-DB is a German database that contains 535 audios of varying duration, recorded by 5 male and 5 female professional speakers. These audios are categorized into seven emotions: anger, boredom, disgust, fear, happiness, neutrality, and sadness. To ensure consistent speech quality, all audio files were recorded at a 16 kHz sampling rate, saved in wav format, with mono-channel and 16 bits per sample. The waveforms of the seven emotions are illustrated in Figure 3.


**Figure 3.** The waveforms representing each emotion in the EMO-DB dataset.

RAVDESS: RAVDESS [36] is the second dataset used in the experiments due to its greater accessibility. This dataset includes 1440 audio files, recorded by 12 male and 12 female actors speaking English scripts with 8 different emotions: anger, calmness, disgust, fear, neutrality, happiness, sadness, and surprise. All recorded audio files have a 48 kHz sample rate and a 16-bit resolution.

SAVEE: This study also utilized the SAVEE [37] dataset, which was recorded by 4 male researchers and students at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey. Each speaker was asked to speak 120 phonetically balanced English sentences in 7 emotional categories: anger, disgust, fear, happy, neutrality, sad, and surprise. This resulted in a total of 480 utterances.

### *3.2. Data Pre-Processing*

The pre-processing of voice is an extremely important stage in systems that cannot tolerate background noise or silence. Such systems include SER and speech recognition, which require efficient extraction of features from audio files, where the majority of the spoken part comprises emotion-related characteristics. To achieve this objective, this study utilized silence removal and pre-emphasis techniques.

By increasing the power of high frequencies in speech signals and leaving low frequencies unchanged, the pre-emphasis technique can enhance the signal-to-noise ratio. This is achieved through the use of a high-pass filter called finite impulse response (FIR), which enhances the high-frequency energy. Essentially, pre-emphasis enhances the high-frequency components of speech signals using Equation (1).

$$H(z) = 1 - \alpha z^{-1}, \alpha = [1, -0.97] \quad (1)$$

The use of FIR in pre-emphasis results in changes to the distribution of energy across the frequencies, as well as alterations to the overall energy level, which can have a significant

influence on the features associated to energy. In contrast, signal normalization utilizes Equation (2) to ensure that voice signals are comparable, regardless of any variations in magnitude.

$$S_{Ni} = \frac{S_i - \mu}{\sigma} \tag{2}$$

The $S_i$ denotes the ith part of signal, while the mean and standard deviation of the signal are represented by $\mu$ and $\sigma$, respectively. $S_{Ni}$ refers to the normalized ith part of the signal.

### 3.3. Feature Extraction

The recognition performance of the model is largely determined by the quality of the feature set. Therefore, unsuitable features may lead to poor recognition results. In the context of machine learning (ML), extracting a relevant feature set is a crucial task for achieving reasonable recognition performance. Ref. [38] reported that feature extraction is an important phase in ML, as the failure or success of the SER model largely depends on the variability of features utilized in the recognition task. Recognition becomes accurate if the derived features are highly correlated with the emotion class, while it becomes difficult and inaccurate if the derived features do not strongly correlate with emotion. In SER, the quality of the feature set greatly affects recognition performance. Therefore, useful features must be extracted from collected audio files, which are suitable to learn recognition rules. Feature extraction is an essential step in this process and requires significant effort and creativity. It involves extracting suitable features from emotional audio files and transforming them into an MFV. The MFV is then utilized by a ML or DL technique to develop a recognition system. It is more challenging to derive features than classification due to its domain-specific nature. In this paper, a new feature extraction technique was implemented to derive effective and suitable features, called MFCCT features, from audio files for constructing an accurate SER model. Details of the derived MFCCT features are presented in the following subsection.

#### The Proposed MFCCT Features

The hybrid features were derived from audio files into two steps: (1) extracting MFCC features, and (2) fusion of time-domain and MFCC features.

#### (a) Extracting the MFCCs features

MFCCs features are derived to capture the vocal tract characteristics of an emotion. The process of MFCCs feature extraction involves framing, windowing, the discrete Fourier transform (DFT), taking the log of magnitude, frequency warping on Mel scale, and applying discrete cosine transform (DCT). To prevent information loss, each utterance was divided into frames of 25 ms with a 10 ms overlap between successive frames. The total frames for single audio can be computed using Equation (3), while the number of samples ($N$) can be calculated using Equation (4).

$$Total_{frames} = \frac{Number of Samples}{(Frame_{step}) \times (SampleRate)} \tag{3}$$

$$N = \left(Frame_{length}\right) \times SampleRate \tag{4}$$

Once the framing steps were completed, each individual frame underwent hamming windowing to ensure that the edges of the frame were smoothed by Equation (5).

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N \tag{5}$$

where $N$ represents the total samples per frame.

Next, the magnitude spectrum of all frames was computed through DFT. These magnitude spectrums were passed to the multiple Mel-filter banks, where Mel represents the perceived frequency of ears. The Mel can be computed as:

$$Mel(f) = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{6}$$

where $f$ denotes the actual frequency and the $Mel(f)$ denotes the corresponding perceived frequency.

To mimic the frequency perception of ears, frequency axis was warped through Equation (6). The Mel-frequency warping with a triangular filter bank is commonly used to achieve this. The Mel-spectrum was then obtained by multiplying all magnitude spectra of triangular filters, $X(k)$, as given below.

$$s(m) = \sum_{k=0}^{N-1}\left[|X(k)|^2 \times H_m(k)\right]; 0 \le m \le M-1 \tag{7}$$

where $M$ represents the triangular filters. $H_m(k)$ denotes the weight assigned to $k$th bin of spectrum, which contributes to the $m$th output band. Mathematically it is written as:

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases} \tag{8}$$

where m differs from 0 to $M-1$.

Ultimately, the MFCCs features were derived by using the DCT to the log of the Mel spectrum of each frame. This step was performed using Equation (9).

$$c(n) = \sum_{m=0}^{M-1} \log_{10}\left(s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right)\right); n = 0, 1, 2, \ldots, c-1 \tag{9}$$

where $C$ represents the MFCCs.

(b) Extracting MFCCT Features

The MFCCT features were obtained through a process outlined in Algorithm 1. This process comprised of three different steps. Firstly, binning method was used on the derived MFCC features, with each bin comprising 1500 rows of each single column. This bin size (1500) was selected as it attained improved accuracy. Secondly, 12 distinct time-domain (*t-domain*) features, as shown in Table 1, were derived from all bins of the MFCCs feature.

In Algorithm 1, the extracted MFCCs feature are represented by the variable '*matrix*', which is in a matrix form. The '*size*' variable shows the bin size (1500). The Algorithm 1 returns the final MFV for SER. The variables '*rows*' and '*cols*' represent the total rows and columns of '*matrix*'. The '*bins*' variable contains the total bins, while '*n*' shows the MFCCT features (12). Therefore, for each emotion, the '*bins*' are multiplied by '*n*' to get the number of rows, and the columns of the matrix represents the number of audio files for a single emotion.

**Table 1.** List of time-domain features.

| Label | Time-Domain Features |
|---|---|
| MIN | Minimum value of each bin |
| MAX | Maximum value of each bin |
| Mn | Mean value of each bin |
| Md | Median of each bin |
| Mo | Mode of each bin |
| STD | Standard deviation of each bin |
| VAR | Variance of each bin |
| COV | Covariance of each bin |
| RMS | Root mean square of each bin |
| Q1 | 25th percentile of each bin |
| Q2 | 50th percentile of each bin |
| Q3 | 75th percentile of each bin |

**Algorithm 1:** Constructing the Master Feature Vector for CNN.

$$matrix \leftarrow \begin{bmatrix} v_{11} & \cdots & v_{1c} \\ \vdots & \ddots & \vdots \\ v_{r1} & \cdots & v_{rc} \end{bmatrix}_{rows \times cols}$$

$bins \leftarrow (rows/size)$
$n \leftarrow 12$
**for** $i := 1$ *to* $cols$ **do**
  $initial \leftarrow 1$
  $\Phi \leftarrow 1$
  **for** $j := 1$ *to* $bins$ **do**
    $\hat{M} \leftarrow matrix[initial : j \times size, i]$
    integer array $[1 \ldots 12] \leftarrow$
    extract 12 features from $\hat{M}$
    **for** $k := 12$ *to* $1$ **do**
      $MFV[\Phi + n - k, i] \leftarrow C[k]$
    **end**
    $\Phi \leftarrow \Phi + n$
    $initial \leftarrow initial + size$
  **end**
**end**
**return** $MFV$

### 3.4. Convolutional Neural Network Model

The proposed methodology for SER utilized a CNN model to process the derived MFCCT features. The CNN model was employed due to its ability to generate a feature map of the time series data, which can achieve enhanced performance on MFCCT features. The derived features were used as input into the CL to produce local features from the input. The CNN model was comprised of three 1D CLs, each followed by a dropout, activation, and max-pool layers. The first (Conv1) layer received an array of features with a 1-pixel stride, 64 filters, and a 5-size kernel, which is a critical parameter to fine-tune the CNN model. The output of every neuron was activated using the activation function called Rectifier Linear Unit (ReLU), followed by a dropout rate of 0.2. The ReLU accelerates convergence and solves the issue of vanishing gradient, while the dropout layer reduces the overfitting issue. The output of all neurons after using ReLU was given to a 1D max-pool layer with a pooling size of 4, implementing batch normalization. The next CL was comprised of 128 filters with 5-size kernel size and 1-pixel stride, followed by an activation, 0.2 dropout rate, and max-pool layer of same size. The final CL was comprised of 256 filters with the same size of kernel and stride, followed by an activation, dropout, and flattening layer to convert the CLs output into a 1D feature vector, utilized as input to the FC layer.

The number of neurons in the FC layer were selected based on the number of emotion classes in the dataset, integrating the global feature obtained from the preceding layers, and generating a vector for SoftMax activation to predict emotions. The model used Adam optimizer with a learning rate of 0.0001, 100 epochs, and a batch size of 16. The proposed configuration of CNN with parameters is given in Table 2.

**Table 2.** Layer structure of the lightweight CNN model.

| CNN Configuration |
| --- |
| Conv_1 (5, @64), activation function = "ReLU", Dropout Rate (0.2), MaxPooling_1 (4), Conv_2 (5, @128), activation function= "ReLU", Dropout Rate (0.2), MaxPooling_2 (4), Conv (5, @256), activation function= "ReLU", Dropout Rate (0.2), Flatten, Dense (7 neurons), Softmax (7 emotions) |
| **Parameters:** Adam Optimizer, 100 epochs , 0.0001 learning rate |

### 3.5. Evaluation Metrics

The classification performance in all experiments was evaluated using weighted accuracy and AUCROC. The weighted accuracy is defined by the ratio of correctly recognized audio files to the sum of all audio files for emotion recognition. Equation (10) provides the mathematical expression of weighted accuracy [39].

$$Accuray = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP+TN}{TP+TN+FP+FN}\right)_i \quad (10)$$

The AUROC is a commonly used metric in ML applications that requires imbalanced datasets. It offers a comprehensive evaluation of classifier performance across all classes and summarizes the ROC curve's performance by calculating the area under it. A high AUC value (close to 1) indicates good classifier performance, whereas a low value (less than 0.5) indicates poor performance [40].

## 4. Experimental Results and Discussion

The performance of the constructed SER model was evaluated through an extensive set of experiments and compared with baseline SER models. Five different experiments were conducted to measure the performance of SER models using the proposed MFCCT features. In the first experiment, the MFCCT features were derived from audio files and fed to six different algorithms to construct the SER models. Six analyses were carried to evaluate the performance of classifiers coupled with the MFCCT features. In the second experiment, the performance of MFCCs and the *t-domain* feature was compared with the MFCCT feature. In the third and fourth experiments, a different binning size and *t-domain* were used to derive the MFCCT feature and get the best learning curve for DL architecture. In the last experiment, the performance of the hybrid feature coupled with the CNN was evaluated using three SER datasets to examine the effectiveness of the constructed SER models. The results of these experiments are reported in the subsequent subsections.

### 4.1. Results of Experiment I

In this experiment, the derived MFCCT features were utilized as inputs for five distinct ML classifiers, including k-NN, RF, J48, NB, and SVM, as well as a CNN. The weighted accuracies of these classifiers for three datasets are demonstrated in Figure 4. The CNN model outperformed the classifiers, obtaining a 97% weighted accuracy for the Emo-DB dataset. Additionally, the CNN model for SER achieved 92.6% and 91.4% weighted accuracy for the SAVEE and RAVDESS datasets, respectively. Among the other five ML classifiers, an irregular pattern was observed in weighted accuracy. The SVM and RF classifiers for SER achieved the highest weighted accuracy (80.7% and 86.9%) on the Emo-DB dataset SER model, compared to other three ML classifiers. Moreover, the SER models for the SAVEE

and RAVDESS, RF, and k-NN classifier achieved the highest weighted accuracy (74% and 54.1%, respectively) compared to the other four ML classifiers. In all the analysis, the NB classifier achieved the lowest weighted accuracy.

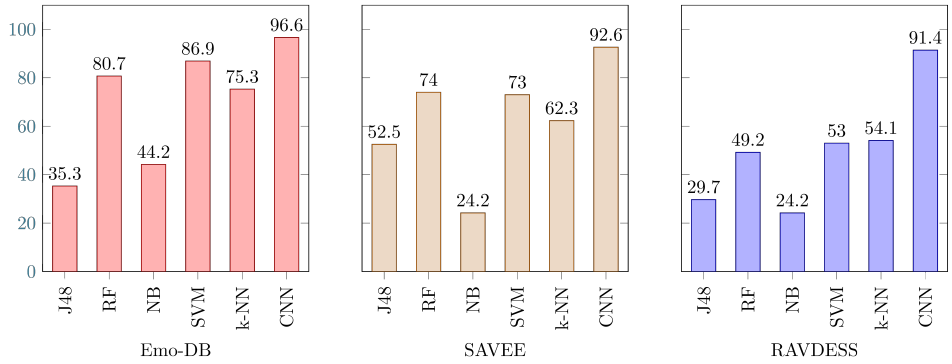


**Figure 4.** Weighted accuracies of the classifiers for the three datasets.

In summary, the CNN model for SER demonstrated superior performance over the other ML classifiers in achieving improved recognition accuracy using all datasets. Furthermore, the ROC diagrams for the three datasets achieved through the CNN model are presented in Figure 5. Figure 5a illustrates the ROC diagram for Emo-DB. The performance of boredom and disgust emotions is slightly better than the anger emotion. This could be because several analysis techniques, such as formant and pitch, are less precise for high-pitch emotions as compared to the low-pitch emotions. Figure 5b illustrates the ROC diagram for all 7 emotions, indicating acceptable recognition accuracy for all emotions. Figure 5c shows the ROC diagram for all emotions in RAVDESS dataset, demonstrating reasonable recognition accuracy for all emotions.

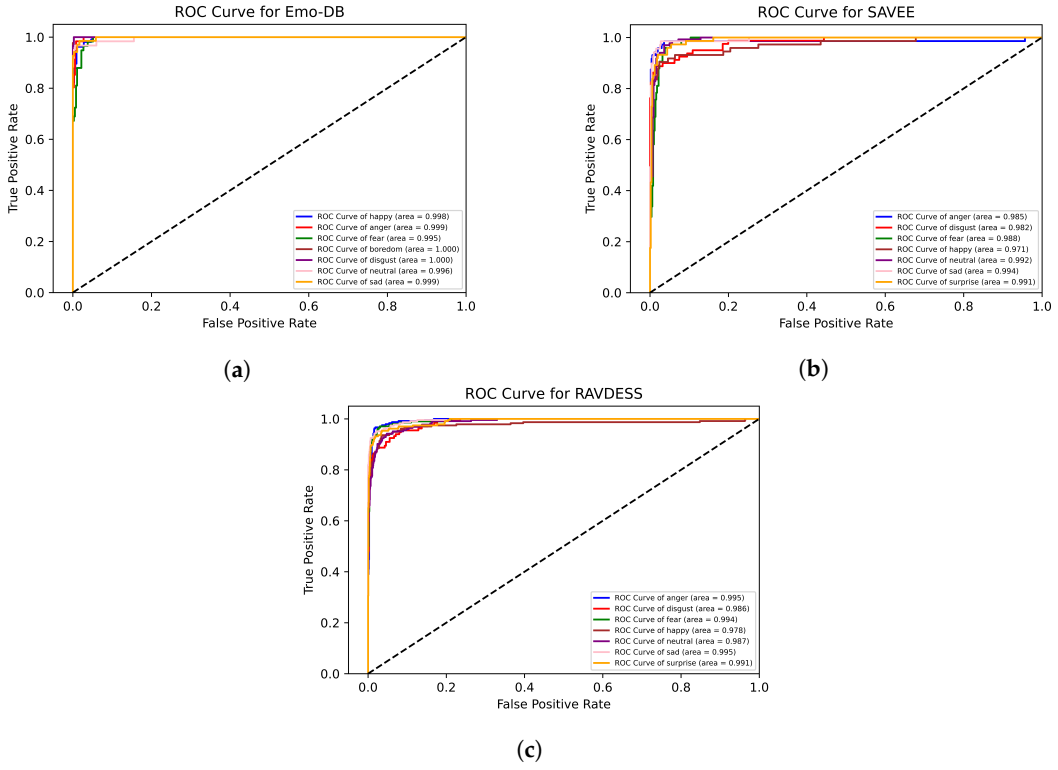


**Figure 5.** ROC Curves of the (**a**) EMO-DB, (**b**) SAVEE, and (**c**) RAVDEES Models.

### *4.2. Results of Experiment II*

In this experiment, the performance of MFCCs and t-domain feature was compared to MFCCT feature using the CNN model. To evaluate the performances of the MFCCT feature, 9 SER models were constructed (3 different feature sets × 3 different datasets). Figure 6 presents the results of all 9 SER models, where the MFCCT features achieved improved weighted accuracy. In addition, the t-domain features showed the lowest weighted accuracy. The MFCCT features achieved around 50% more weighted accuracy compared to the MFCCs and t-domain features.

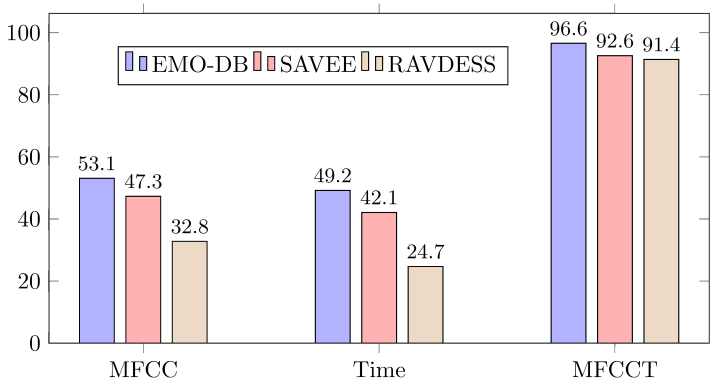


**Figure 6.** Performance comparison of MFCC, t-domain, and MFCCT.

### *4.3. Results of Experiment III*

In this experiment, the weighted accuracy of MFCCT feature was compared at various binning sizes. The SER models were evaluated by applying distinct binning sizes ranging from 500 to 3000. The weighted accuracy of each binning size is presented in Table 3. The results revealed that the binning size of 1500 obtained the highest weighted accuracy when the proposed MFCCT feature was utilized. Beyond this bin size, the weighted accuracy decreased gradually, with the lowest accuracy occurring at a binning size of 3000.

**Table 3.** Comparison of weighted accuracy (×100) at different binning sizes.

| | **500** | **1000** | **1500** | **2000** | **2500** | **3000** |
|---|---|---|---|---|---|---|
| EMO-DB | 0.76 | 0.87 | **0.97** | 0.81 | 0.81 | 0.82 |
| SAVEE | 0.77 | 0.84 | **0.93** | 0.78 | 0.74 | 0.74 |
| RAVDESS | 0.72 | 0.81 | **0.92** | 0.73 | 0.64 | 0.76 |

### *4.4. Results of Experiment IV*

This experiment evaluated the effectiveness of combining the different *t-domain* features listed in Table 1 to derive MFCCT features. Initially, the first two *t-domain* features from Table 1 were utilized to derive the MFCCT features. These *t-domain* features were then increased to 4, 6, 8, 10, and 12. Thereafter, the 18 different SER models (Table 4) were constructed to evaluate recognition accuracy across datasets. The table shows an incremental pattern in recognition accuracy, with the highest weighted accuracy achieved when 12 *t-domain* features were utilized to derive MFCCT features. Conversely, the lowest recognition accuracy was achieved when MFCCT features were derived using only 2 *t-domain* features.

**Table 4.** t-domain features used to obtain MFCCT features.

| No of *t-Domain* Features | EMO-DB | SAVEE | RAVDESS |
|---|---|---|---|
| 2 | 81.6 | 53.8 | 48.2 |
| 4 | 84.6 | 73.6 | 69.3 |
| 6 | 87.4 | 78.5 | 73.9 |
| 8 | 88.3 | 80.1 | 76.8 |
| 10 | 90.6 | 83.7 | 79.8 |
| **12** | **96.6%** | **92.6%** | **91.4%** |

### *4.5. Comparison with Baseline*

To show the effectiveness and robustness of proposed features for SER, the performance was compared with baseline methods using EMO-DB, SAVEE, and RAVDESS datasets. A detailed overview of comparisons is provided in Table 5. As shown in the table, the MFCCT feature coupled with the CNN model achieved better performance than the baseline SER methods, which reveals the effectiveness of our approach. Nevertheless, in some cases the recognition rate of our approach for a particular emotion is slightly less than the existing SER methods. For example, the SER model for Emo-DB dataset in Ref. [1] recognized the boredom with an accuracy of 95%, while the proposed SER model achieved 91% accuracy for boredom emotion in same dataset. However, the proposed approach outperformed existing methods by achieving a weighted accuracy of 97% compared to 93% of the baseline. The proposed method for SER recognized all emotions with higher accuracy, less computation time, and is suitable for real-time. Therefore, it can be established that the proposed approach is generic, more accurate, and reliable than the existing methods.

**Table 5.** Comparative analysis of the SER method and the baseline methods utilizing the EMO-DB, SAVEE, and RAVDESS datasets.

| Study | Datasets | Accuracy (×100%) of Individual Emotion | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **A** | **B** | **C** | **D** | **F** | **H** | **N** | **S** | **U** | **Avg** |
| [20] | EMO-DB | 1.00 | 0.61 | × | 0.67 | 0.67 | 1.00 | 1.00 | 0.87 | × | 0.86 |
| | RAVDESS | 0.92 | × | 0.57 | 0.72 | 0.76 | 0.68 | 0.75 | 0.52 | 0.80 | 0.71 |
| [41] | EMO-DB | 0.92 | 0.88 | × | 0.99 | 0.92 | 0.92 | 0.90 | 0.93 | × | 0.90 |
| | RAVDESS | 0.80 | × | 0.90 | 0.71 | 0.74 | 0.65 | 0.68 | 0.66 | 0.67 | 0.73 |
| | SAVEE | 0.90 | × | × | 0.48 | 0.50 | 0.47 | 0.82 | 0.58 | 0.53 | 0.67 |
| [19] | EMO-DB | 0.88 | 0.95 | × | 0.84 | 0.95 | 0.84 | 0.95 | 0.95 | × | 0.93 |
| [1] | EMO-DB | 0.88 | 0.95 | × | 0.84 | 0.95 | 0.84 | 0.95 | 0.95 | × | 0.93 |
| [21] | EMO-DB | 0.91 | 0.90 | × | 0.87 | 0.92 | 0.66 | 0.85 | 0.88 | × | 0.85 |
| | RAVDESS | 0.95 | × | 0.95 | 0.86 | 0.91 | 0.43 | 0.50 | 0.61 | 0.95 | 0.77 |
| Proposed model | EMO-DB | 0.96 | 0.91 | × | 0.96 | 0.99 | 1.00 | 0.97 | 0.99 | × | **0.97** |
| | SAVEE | 0.94 | × | × | 0.90 | 0.91 | 0.95 | 0.95 | 0.94 | 0.93 | **0.93** |
| | RAVDESS | 0.95 | × | 0.93 | 0.92 | 0.94 | 0.95 | 0.89 | 0.91 | 0.93 | **0.92** |

A = Anger, B = Boredom, C = Calm, D = Disgust, F = Fear, H = Happy, N = Neutral, S = Sad, U = Surprise.

### *4.6. Summary*

In this study, the MFCCT features were proposed using a 1D CNN model for an effective SER, utilizing the EMO-DB, SAVEE, and RAVDESS datasets. The results revealed that the MFCCT feature coupled with 1D CNN can efficiently identify emotions from speech. Furthermore, the results revealed that the MFCCT features, utilized as input to 1D CNN to retrieve the high-level patterns, are more discriminative at obtaining enhanced performance of the SER system. Several research studies have proposed various methods for SER employing DL models and handcrafted features, but these methods are still not effective in terms of size, computational time, and recognition accuracy [42,43]. This research adopted

a novel method to address these issues through hybrid features (MFCCs and t-main) and 1D CNN to extract high-level patterns from audio data. The proposed method for each dataset recognized the emotions from audio signals with high accuracy. Secondly, this research resolves the problem of computation time by employing a lightweight and simple 1D CNN that used three CLs, two max-pool layers, three dropouts, and an FC layer with softmax classifier to identify emotions from audio files. To show the effectiveness, this research evaluated the efficiency of the proposed method with existing methods and the results showed that the proposed method outperformed the baseline methods by enhancing the number of accurately recognized instances. The potential reason for the enhanced recognition rate is that this study combined both the f-domain and t-domain representations of the audio signal to enhance the robustness, diversity, reliabilities, and to increase the generalization of the proposed SER method.

## 5. Conclusions

SER is a complex task that includes two main challenges: feature extraction and classification. This research proposed the novel fusion of MFCCs and t-domain features coupled with a 1D CNN for emotion classification. The CNN model comprised three CLs, two max-pooling layers, one flattening layer, and one FC dense layer. The use of a small number of layers decreased the computation time and size. To evaluate the performance, the proposed method employed three datasets: EMO-DB, SAVEE, and RAVDESS. The accuracy of the CNN model using three datasets outperformed the baseline methods. The proposed technique achieved an accuracy of 96.6% for the EMO-DB datasets, 92.6% for the SAVEE dataset, and 91.4% for the RAVDESS dataset. Moreover, the proposed method enhanced the accuracy by 10% for EMO-DB, 26% for SAVEE, and 21% for RAVDESS datasets. This demonstrates the significance and robustness of the proposed method for SER. A comparison analysis of SER methods based on DL approaches employing other datasets is planned for future work. Furthermore, the implementation RNNs while using optimum acoustic features can enhance the accuracy of the SER, since it offers high-level acoustic features accurately.

**Author Contributions:** Methodology, A.S.A. and R.J.; Writing—original draft, O.S.N.; Writing—review & editing, M.A.N.; Supervision, O.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R234), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Chen, L.; Su, W.; Feng, Y.; Wu, M.; She, J.; Hirota, K. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction. *Inf. Sci.* **2020**, *509*, 150–163. [CrossRef]
2. Hansen, J.H.; Cairns, D.A. Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard effect environments. *Speech Commun.* **1995**, *16*, 391–422. [CrossRef]
3. Koduru, A.; Valiveti, H.B.; Budati, A.K. Feature extraction algorithms to improve the speech emotion recognition rate. *Int. J. Speech Technol.* **2020**, *23*, 45–55. [CrossRef]
4. Zheng, W.; Zheng, W.; Zong, Y. Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition. *Virtual Real. Intell. Hardw.* **2021**, *3*, 65–75. [CrossRef]
5. Schuller, B.; Rigoll, G.; Lang, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 17–21 May 2004; pp. 577–580.

6. Spencer, C.; Koç, İ.A.; Suga, C.; Lee, A.; Dhareshwar, A.M.; Franzén, E.; Iozzo, M.; Morrison, G.; McKeown, G. *A Comparison of Unimodal and Multimodal Measurements of Driver Stress in Real-World Driving Conditions*; ACM: New York, NY, USA, 2020.
7. France, D.J.; Shiavi, R.G.; Silverman, S.; Silverman, M.; Wilkes, M. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 829–837. [CrossRef]
8. Uddin, M.Z.; Nilsson, E.G. Emotion recognition using speech and neural structured learning to facilitate edge intelligence. *Eng. Appl. Artif. Intell.* **2020**, *94*, 103775. [CrossRef]
9. Jahangir, R.; Teh, Y.W.; Hanif, F.; Mujtaba, G. Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimed. Tools Appl.* **2021**, *80*, 23745–23812. [CrossRef]
10. Fahad, M.S.; Ranjan, A.; Yadav, J.; Deepak, A. A survey of speech emotion recognition in natural environment. *Digit. Signal Process.* **2021**, *110*, 102951. [CrossRef]
11. Jahangir, R.; Teh, Y.W.; Mujtaba, G.; Alroobaea, R.; Shaikh, Z.H.; Ali, I. Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion. *Mach. Vis. Appl.* **2022**, *33*, 41. [CrossRef]
12. Ayadi, M.E.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. [CrossRef]
13. Abdel-Hamid, O.; Mohamed, A.-R.; Jiang, H.; Deng, L.; Penn, G.; Yu, D. Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1533–1545. [CrossRef]
14. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
15. Anvarjon, T.; Kwon, S. Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors* **2020**, *20*, 5212. [CrossRef] [PubMed]
16. Rybka, J.; Janicki, A. Comparison of speaker dependent and speaker independent emotion recognition. *Int. J. Appl. Math. Comput. Sci.* **2013**, *23*, 797–808. [CrossRef]
17. Akçay, M.B.; Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. [CrossRef]
18. Zhang, S.; Tao, X.; Chuang, Y.; Zhao, X. Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Commun.* **2021**, *127*, 73–81. [CrossRef]
19. Pawar, M.D.; Kokate, R.D. Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. *Multimed. Tools Appl.* **2021**, *80*, 15563–15587. [CrossRef]
20. Issa, D.; Demirci, M.F.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control.* **2020**, *59*, 101894. [CrossRef]
21. Sajjad, M.; Kwon, S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access* **2020**, *8*, 79861–79875.
22. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. *Multimed. Tools Appl.* **2019**, *78*, 5571–5589. [CrossRef]
23. Er, M.B. A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. *IEEE Access* **2020**, *8*, 221640–221653. [CrossRef]
24. Nicholson, J.; Takahashi, K.; Nakatsu, R. Emotion recognition in speech using neural networks. *Neural Comput. Appl.* **2000**, *9*, 290–296. [CrossRef]
25. Noroozi, F.; Sapiński, T.; Kamińska, D.; Anbarjafari, G. Vocal-based emotion recognition using random forests and decision tree. *Int. J. Speech Technol.* **2017**, *20*, 239–246. [CrossRef]
26. Nwe, T.L.; Foo, S.W.; Silva, L.C.D. Speech emotion recognition using hidden Markov models. *Speech Commun.* **2003**, *41*, 603–623. [CrossRef]
27. Aljuhani, R.H.; Alshutayri, A.; Alahdal, S. Arabic Speech Emotion Recognition From Saudi Dialect Corpus. *IEEE Access* **2021**, *9*, 127081–127085. [CrossRef]
28. Al-onazi, B.B.; Nauman, M.A.; Jahangir, R.; Malik, M.M.; Alkhammash, E.H.; Elshewey, A.M. Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion. *Appl. Sci.* **2022**, *12*, 9188. [CrossRef]
29. Jahangir, R.; Teh, Y.W.; Memon, N.A.; Mujtaba, G.; Zareei, M.; Ishtiaq, U.; Akhtar, M.Z.; Ali, I. Text-independent speaker identification through feature fusion and deep neural network. *IEEE Access* **2020**, *8*, 32187–32202. [CrossRef]
30. Jahangir, R.; Teh, Y.W.; Nweke, H.F.; Mujtaba, G.; Al-Garadi, M.A.; Ali, I. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Syst. Appl.* **2021**, *171*, 114591. [CrossRef]
31. Khan, A.A.; Jahangir, R.; Alroobaea, R.; Alyahyan, S.Y.; Almulhi, A.H.; Alsafyani, M. An efficient text-independent speaker identification using feature fusion and transformer model. *Comput. Mater. Contin.* 2023, 75, 4085–4100.
32. Garcia-Ceja, E.; Riegler, M.; Kvernberg, A.K.; Torresen, J. User-adaptive models for activity and emotion recognition using deep transfer learning and data augmentation. *User Model. User-Adapt. Interact.* **2020**, *30*, 365–393. [CrossRef]
33. Nie, W.; Ren, M.; Nie, J.; Zhao, S. C-GCN: Correlation based Graph Convolutional Network for Audio-video Emotion Recognition. *IEEE Trans. Multimed.* **2020**, *23*, 3793–3804. [CrossRef]
34. Gholamy, A.; Kreinovich, V.; Kosheleva, O. *Why 70/30 or 80/20 Relation between Training and Testing Sets: A Pedagogical Explanation*; University of Texas at El Paso USA: El Paso, TX, USA, 2018.

35. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of German emotional speech. In Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005.
36. Livingstone, S.R.; Russo, F.A. 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* **2018**, *13*, e0196391. [CrossRef] [PubMed]
37. Jackson, P.; Haq, S. *Surrey Audio-Visual Expressed Emotion (Savee) Database*; University of Surrey: Guildford, UK, 2014.
38. Domingos, P. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
39. Tahon, M.; Devillers, L. Towards a small set of robust acoustic features for emotion recognition: Challenges. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *24*, 16–28. [CrossRef]
40. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
41. Farooq, M.; Hussain, F.; Baloch, N.K.; Raja, F.R.; Yu, H.; Zikria, Y.B. Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network. *Sensors* **2020**, *20*, 6008. [CrossRef] [PubMed]
42. Zhao, Z.; Li, Q.; Zhang, Z.; Cummins, N.; Wang, H.; Tao, J.; Schuller, B.W. Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-Based discrete speech emotion recognition. *Neural Netw.* **2021**, *141*, 52–60. [CrossRef]
43. Kwon, S. Att-Net: Enhanced emotion recognition system using lightweight self-attention module. *Appl. Soft Comput.* **2021**, *102*, 107101.