Multi-Modal Emotion Aware System Based on Fusion of Speech and Brain Information

Rania M. Ghoniem [1,2,*]**, Abeer D. Algarni** [2] **and Khaled Shaalan** [3]

[1] Department of Computer, Mansoura University, Mansoura 35516, Egypt
[2] Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 84428, Saudi Arabia
[3] Faculty of Engineering & IT, The British University in Dubai, Dubai 345015, United Arab Emirates
[*] Correspondence: RMGhoniem@pnu.edu.sa

Abstract: In multi-modal emotion aware frameworks, it is essential to estimate the emotional features then fuse them to different degrees. This basically follows either a feature-level or decision-level strategy. In all likelihood, while features from several modalities may enhance the classification performance, they might exhibit high dimensionality and make the learning process complex for the most used machine learning algorithms. To overcome issues of feature extraction and multi-modal fusion, hybrid fuzzy-evolutionary computation methodologies are employed to demonstrate ultra-strong capability of learning features and dimensionality reduction. This paper proposes a novel multi-modal emotion aware system by fusing speech with EEG modalities. Firstly, a mixing feature set of speaker-dependent and independent characteristics is estimated from speech signal. Further, EEG is utilized as inner channel complementing speech for more authoritative recognition, by extracting multiple features belonging to time, frequency, and time–frequency. For classifying unimodal data of either speech or EEG, a hybrid fuzzy *c*-means-genetic algorithm-neural network model is proposed, where its fitness function finds the optimal fuzzy cluster number reducing the classification error. To fuse speech with EEG information, a separate classifier is used for each modality, then output is computed by integrating their posterior probabilities. Results show the superiority of the proposed model, where the overall performance in terms of accuracy average rates is 98.06%, and 97.28%, and 98.53% for EEG, speech, and multi-modal recognition, respectively. The proposed model is also applied to two public databases for speech and EEG, namely: SAVEE and MAHNOB, which achieve accuracies of 98.21% and 98.26%, respectively.

Keywords: multi-modal emotion aware systems; speech processing; EEG signal processing; hybrid classification models

1. Introduction

In human–computer interaction (HCI), comprehending and discriminating emotions turned into a principal issue to construct intelligent systems that could perform purposed actions. Emotions can be discriminated utilizing distinct forms of sole modalities, like facial expression, short phrases, speech, video, EEG signals, and long/short texts. These modalities vary over the computer applications, e.g., the well-known modality in computer games is video.

Recent studies unveil several merits of employing physiological signals for recognizing emotions [1]. For instance, the electroencephalogram (EEG) signals have been shown to be a robust sole modality [2,3]. The control of these bio-signals is managed by our central nervous system; thus, it cannot be affected intentionally, while actors can pretend with emotion on their faces deliberately. Further, physiological signals are emitted constantly and since sensors are directly attached to subject's body,

they are not out of reach. Moreover, physiological information could also be utilized as supplementary to emotional information gained from facial expressions or speech to optimize the recognition rates.

In this regard, several multi-modal emotion recognition approaches have achieved significant contributions [4–6]. Though there are distinctive input modalities for recognizing emotions, the most widely recognized are bimodal inputs that combine both speech and video. The two modalities have been selected frequently by the researchers because of being captured in a noninvasive way and are expressive than the other modalities. Nevertheless, face expressions might be falsified by not returning accurate information regarding the person's internal state.

From between the modalities that tend to be in the literature, speech is an intuitive measure for computers to comprehend human emotion, whereas EEG [7] is an internal measurement from the brain that makes an intriguing alternative for recognizing multi-modal emotion. Up until now, there are no proposals that have endeavored to consider speech and EEG simultaneously for recognizing spontaneous emotion. This has motivated us to propose a new multi-modal emotion aware system that fuses speech and EEG modalities for discriminating emotional state of the subject.

In the multi-modal fusion, input information from many modalities, like audio, EEG, video, and electrocardiogram (ECG) can be fused together coherently [8]. More than one modality cannot be combined in a context-free way; a context-dependent model has to be utilized. Information fusion is classified into three progressive stages: (I) Early fusion, (II) intermediate fusion, and (III) late fusion. As for early fusion, integration of information is implemented at signal-/feature-level, while in late fusion, information at a semantic level is to be fused. The foundation in combining multi-modal information is the modalities number [2,3], information derivation synchronization, and fusion procedure, besides finding the proper fusion level of information. However, it is not constantly fundamental that diverse modalities give complimentary information through the fusion stage; therefore, it is essential to comprehend each modality's contributions with respect to accomplishment of distinct tasks.

One of the important problems faced in the multi-modal emotion recognition is the fusion of features belonging to different modalities. By reviewing the literature on the multi-modal emotion analysis, it has been observed that the majority of works have concentrated on the concatenation of feature vectors obtained from different modalities. However, this does not take into consideration the conflicting information that may be carried by the unimodal modalities. On the other hand, little work has addressed this issue through new feature-fusion methods to enhance the multi-modal fusion mechanism [9–12].

In [9], the authors proposed a feature fusion strategy that proceeds from the unimodal to bimodal data vectors and later, the bimodal to trimodal data vectors. In [10], a context-aware audio–video system is proposed. Accordingly, a contextual audio–video switching approach is presented to switch between visual-only, audio-only, and audio–visual clues. The proposed approach integrates the convolutional neural network (CNN) and the long-short-term memory (LSTM) network. Moreover, a multi-modal hybrid deep neural network architecture is presented in [11] for audio–visual mask estimation. In [12], an image-text emotion analysis model is proposed. To utilize the internal correlation between the image and textual features, an intermediate fusion-based multimodal model was proposed. Eventually, a late fusion scheme was applied to combine the models of sentiment prediction.

One of the drawbacks of feature-level emotion fusion is that it might exhibit high dimensionality and poor performance because of redundancy and inefficiency. For overcoming the high dimensionality issue, hybrid intelligent models can introduce several choices for unorthodox handling of complex problems, which carry uncertainty, vagueness, and high dimensionality of data. They can exploit a priori knowledge and the raw data to introduce innovative solutions. In this regard, hybridization is a crucial phase in many domains of human activity [13,14].

This paper suggests a new hybrid emotion recognition model based on the fuzzy clustering, genetic search-based optimization. The proposed model selects and trains the neural network (NN) with the optimal fuzzy clusters representing each modality, without any prior knowledge about the fuzzy clusters number. Unlike the previous works on multimodal emotion fusion, the fusion

of two modalities is implemented on the decision-level instead of the feature-level fusion that may carry conflicting information. For comparison purposes, the proposed hybrid model is likewise compared to another developed hybrid *c*-means-genetic algorithm-neural network model, namely, $FCM - GA - NN_{fixed}$ model that relies on defining a fixed fuzzy clusters number.

2. Related Work

This study suggests a novel multi-modal emotion aware system by fusing together speech and EEG modalities. Consequently, the literature review section is progressive. So, the emotion recognition topic of interest is divided into three parts speech emotion recognition, EEG-based emotion recognition, and multi-modal emotion recognition.

2.1. Literature Review on Speech Emotion Recognition

The related work on speech emotions classification reveals three crucial aspects. Firstly, preparing the emotional database is critical to validate the system performance. Secondly, selecting features properly for speech characterization. Eventually, design of accurate classification model. For emotional speech databases, acoustic analysis has been utilized to recognize emotions using three kinds of databases: Natural spontaneous emotions, acted emotions, and elicited emotions.

Most emotional speech databases depend on inviting professional actors for expressing pre-determined sentences related to the purposed emotion. Though, in some databases like the Danish Emotional Speech (DES) database [15], semi-professional actors are invited for avoiding exaggeration during expressing emotions. As for spontaneous speech, databases may be collected from interaction with robots, or call center data.

The performance of the speech emotion recognition system depends on the features extracted from the speech signal. A challenging issue in recognizing speech emotions is extracting speech features that efficiently describe the speech emotional content and, simultaneously, do not rely on the speakers or the lexical contents. The majority of published works on speech emotion classification have concentrated on analyzing speech spectral information and prosodic features. Some new parameters are utilized for recognizing speech emotion, like the Fourier parameters [16]. Although numerous acoustic parameters have been found to carry emotional content, little success was proved to define a feature set that performs invariably over diverse conditions [17].

Therefore, the majority of works use mixing feature set, which comprises several types of features involving more emotional information [18]. This is implemented by dividing the signal into k frames/segments containing n samples per frame/segment, which causes a high dimensionality problem. As a result, the computational cost and over-fitting likelihood of the speech classifier are increased. Thus, feature selection approaches are imperative to minimize feature redundancy and speed up the learning process of speech emotion recognition.

Table 1 [19–22] summarizes some works on speech emotion recognition. By reviewing these works, it is obvious that the most used classifiers are artificial neural networks, Gaussian mixture model, and multiple different one-level standard methodologies. Satisfactory outcomes are obtained using these standard classifiers. Nevertheless, the improvements in their performances are usually restricted. Thus, the fuzzy genetic search-based optimizations and fusion of speech classifiers could constitute a new step toward robust emotion classification [23,24]. Therefore, we hypothesize that the proposed hybrid soft computing model can overcome the existing limitations of speech emotion recognition.

Table 1. Recent Research on speech emotion recognition.

Reference	Publication Year	Corpus	Speech Analysis	Feature Selection	Classifier	Classification Accuracy
[19]	2019	Berlin EmoDB	The local feature learning blocks (LFLB) and the LSTM are used to learn the local and global features from the raw signals and log-mel spectrograms.	-	1D & 2D CNN LSTM networks.	95.33% and 95.89% for speaker-dependent and speaker-independent results, respectively.
		IEMOCAP				89.16% and 52.14% for speaker-dependent and speaker-independent results, respectively.
[20]	2018	Chinese speech database from Chinese academy of sciences (CASIA)	Speaker-dependent features, and speaker-independent features.	A correlation analysis and Fisher-based method.	Extreme learning machine (ELM)	89.6%
[21]	2015	BES	Two prosodic features and four paralinguistic features of the pitch and spectral energy balance.	None	Support vector machines (SVM)	94.9%
		LDC Emotional Prosody Speech and Transcripts				88.32%
		Polish Emotional Speech Database				90%
[22]	2015	BES	Spectral and prosodic features.	None	Ranking SVMs	82.1%
		LDC				52.4%
		FAU Aibo				39.4%

2.2. Literature Review on EEG-Based Emotion Recognition

The initial stage of EEG-based emotion classification is to take accurately labeled EEG signals induced by picture, video clips, or music. Subsequent to presentation of visual stimuli to a subject, multi-channel signals of EEG are recorded, and afterwards, the signal labeling is done based on subject ratings. The previous work on EEG emotional features extraction [25–29] has revealed that there exist several valuable features of time, frequency, and time–frequency that have been evidenced to be efficient in differentiating emotions. Furthermore, there is no standard feature set that has been agreed as the most appropriate for EEG emotion classification. This causes a high dimensionality issue in EEGs, because not all features would contain significant information concerning emotions. The redundant and irrelevant features maximize the feature space, making the detection of patterns more difficult, and maximizing the over-fitting risks.

However, the majority of classifiers that have been proposed in the literature for EEG emotion classification are based on the conventional classification algorithms such as ANN, SVM, and k-nearest neighbor. Some of these approaches are demonstrated in Table 2. Few works have presented hybrid methods of evolutionary computer algorithms and classification methods [25], where the objective of these works is to deal with the high dimensionality issue of EEG emotion recognition. Therefore, in this work, we propose a novel fuzzy *c*-means-genetic algorithm-neural network (FCM-GA-NN) model for EEG emotion recognition. The meta-heuristic can provide the optimal initial centroids for the FCM algorithm, which will be trained to the NN as optimized solutions for emotion recognition. In this regard, the proposed algorithm has been experimented on our collected database and another publicly available database, namely MAHNOB. The comparative results are demonstrated in the experimental results section.

Table 2. Recent research on electroencephalogram (EEG)-based emotion recognition.

Reference	Corpus	Publication Year	Feature Extraction	Feature Selection	Classifier	Emotion Classes	Classification Accuracy
[26]	DEAP database	2018	Multivariate synchrosqueezing transform.	Non-negative matrix factorization and independent component analysis.	Artificial Neural Network (ANN)	Arousal/valence states (high arousal-high valence, high arousal-low valence, low arousal-low valence, low arousal -high valence).	82.03% and 82.11% for valence and arousal state recognition.
[27]	DEAP database	2018	Liquid State Machines (LSM)	-	Decision Trees	Valence, arousal as well as liking classes.	84.63%, 88.54%, and 87.03% for valence, arousal, and liking classes, respectively.
[28]	DEAP database	2016	Empirical mode decomposition and sample entropy.	None	SVM	Arousal/valence states (high arousal-high valence, high arousal-low valence, low arousal-low valence, low arousal-high valence).	94.98% for binary-class tasks, and 93.20% for the multi-class task.
[29]	Collected database	2016	Features of time (i.e., Latency to Amplitude Ration, Peak to Peak Signal Value, etc.), frequency (i.e., Power Spectral Density, and Band Power), and wavelet domain.	None	Three different classifiers (ANN, k-nearest neighbor, and SVM)	Happiness, sadness, love, and, anger	78.11% for ANN classifier.

2.3. Literature Review on Multi-Modal Emotion Recognition

Recently, a little work has investigated multiple modalities to recognize emotions [9–12]. Many studies have fused facial expression together with physiological signals [30]. Table 3 [30–33] presents some of the surveyed studies on the multi-modal emotion fusion, including the corpus, modalities of fusion, feature extractors, fusion approach, classifier, and classification accuracy. As indicated in the literature and to the best of our knowledge, there are no works reported on fusing speech with EEG for recognizing emotions. Furthermore, the resulting accuracies are still below expectations.

It is also observed that the large number of multi-modal emotion fusion works is based on the feature-level fusion that concatenates the features obtained from the signals of multiple modalities before feeding them to the classifier. This may exceed the risks of conflicting information that may be carried by the sole modalities. On the other hand, a little work has been implemented on the decision-level fusion, which attempts to manipulate each sole modality separately, and integrates later the results from their classifiers for making the final recognition.

In the two cases of multi-modal emotion fusion, the high dimensionality or the redundant features resulting from the processing of each modality may make the learning process complex for the most used machine learning algorithms like the NN classifier, which will be optimized in this work. The training step of NN classifier is a crucial procedure. A high computational time is needed if NN is trained by high-dimensional data. It presupposes network architecture of a huge input layer, which significantly maximizes the weights number, usually causing an infeasible training.

This issue could be solved by minimizing the input space dimensionality to a manageable size, then a network is trained on fewer dimensions. Clustering is a strictly necessary solution that has been exploited to make the dimensionality reduction of an NN training data through organizing group of objects or patterns into clusters. Accordingly, objects within the cluster itself reveal common attributes and others within different clusters reveal dissimilarity. *k*-means [34] and FCM [35] are the most employed clustering algorithms to train NN. As for FCM, a soft partitioning is executed as a pattern and is esteemed as a member of all clusters, but distinct membership degrees are set for distinct clusters. Despite that, the two algorithms are centroid-based, so they previously assume a fixed cluster number, and are fully sensitive for centroid initialization [36].

On the contrary, several problems do not have knowledge on the clusters number a priori. Numerous studies suggested solutions for these problems through running the algorithm repeatedly along with diverse fixed centroid value *k* and with diverse initializations. Nevertheless, this might not be doable with the big datasets. Moreover, running the algorithm through using a limited centroid number might be inefficient because one solution relies on a limited initialization set. This is referred to as "clusters number dependency" issue [37].

For solving this issue, evolutionary approaches reveal alternative optimization approaches utilizing stochastic principles for evolving clustering solutions. They also are based upon probabilistic rules for returning the near-optimal solution among the global search space. A few evolutionary methods have been suggested for optimizing clusters number in data partitioning issues [38,39]. In [38], an algorithm of artificial bee colony was executed to mimic an intelligent foraging conduct of honey bee swarms. In [40], a *k*-means was optimized through a GA, which esteems the impact of isolated points. Several studies also suggested a number of approaches for NN optimization by GA [41–44].

To overcome the high dimensionality issue, this paper suggests a new multi-class NN model, which is optimized by hybridization of the FCM and GA. For each sole modality, the GA selects the optimal centroids for the FCM algorithm. Then, the NN is automatically trained with the optimized solutions from each modality, which reduce the classification error without any prior knowledge about the fuzzy clusters number. For comparison purposes, we developed another hybrid model of FCM, GA, and NN, which is trained using a fixed number of fuzzy clusters. The fusion is then implemented on the decision level.

Table 3. Research on multi-modal emotion recognition utilizing multiple modalities.

Reference	Publication Year	Corpus	Fused Modalities	Feature Extraction	Fusion Approach	Classifier	Classification Accuracy
[30]	2019	RML, Enterface05, and BAUM-1s	Speech and image features	2D convolutional neural network for audio signals, and 3D convolutional neural network for image features.	Deep belief network	SVM	85.69% for Enterface05 dataset in case of multi-modal classification based upon six discrete emotions.
							91.3% and 91.8%, respectively, for binary arousal-valence model, in case of multi-modal classification using Enterface05 dataset.
[31]	2019	Private database, namely, Big Data and a publicly available database of Enterface05.	Speech and video	For speech feature extraction, Mel-spectrogram is obtained. For video feature extraction, a number of representative frames are selected from a video segment.	Extreme Learning Machines (ELM)	The CNN is separately fed with speech and video features, respectively.	An accuracy of 99.9% and 86.4% for the favor of ELM fusion using the Big Data and Enterface05, respectively.
[32]	2017	MAHNOB-HCI database	Facial expressions and EEG	As for facial expressions, the appearance features are estimated from each frame block then the expression percentage feature is computed. The EEG features are calculated using the Welch's Averaged Periodogram.	Feature-level fusion by concatenating all features within a single vector, in addition to decision-level fusion by processing each modality in a separate way, and amalgamating results from their special classifiers in the recognition stage.	For decision-level fusion, LWF and EWF are used. For feature-level fusion, the paper used several statistical fusion methods like Canonical Correlation Analysis (CCA) and Multiple Feature Concatenation (MFC).	For decision-level fusion, LWF achieved recognition rates of 66.28% and 63.22%, for valence and arousal classes, respectively. For feature-level fusion, MFC achieved recognition rates of 57.47% and 58.62%, for valence and arousal classes, respectively.

Table 3. *Cont.*

Reference	Publication Year	Corpus	Fused Modalities	Feature Extraction	Fusion Approach	Classifier	Classification Accuracy
[33]	2015	Private database, which were collected from university students.	EEG images along with speech signals	For EEG images, the features were extracted using the threshold, the Sobel edge detection, and some statistical measures (e.g., mean, variance, standard deviation, etc.). While the intensity, the RMS Energy, and the pitch were used for speech signal feature extraction.	None (the study investigated correlation of EEG images as well as Speech signals)	None	The significance accuracy of correlation coefficient was about 95% for the favor of said emotional status.

3. Proposed Methodology

The proposed system architecture is as depicted in Figure 1, which comprised five steps: (i) Multi-modal data acquisition, (ii) pre-processing, (iii) feature extraction, (iv) classification using the proposed hybrid fuzzy *c*-means-genetic algorithm-neural network model (CM-GA-NN) model, and (v) fusion on the decision-level. Speech and EEG signals are acquired from subjects simultaneously. Recorded signals are then pre-processed to eliminate noise of external interferences. Next to pre-processing, some speaker-dependent and -independent features are estimated from speech signals. For EEG signals, features are estimated from three domains: Time, frequency, and time–frequency. For classifying unimodal data (speech and EEG), the proposed hybrid FCM-GA-NN model is used. To fuse speech and EEG information, two algorithms are experimented, where a separate FCM-GA-NN classifier is used for each modality, then the output is computed by integrating posterior probabilities of sole modalities.
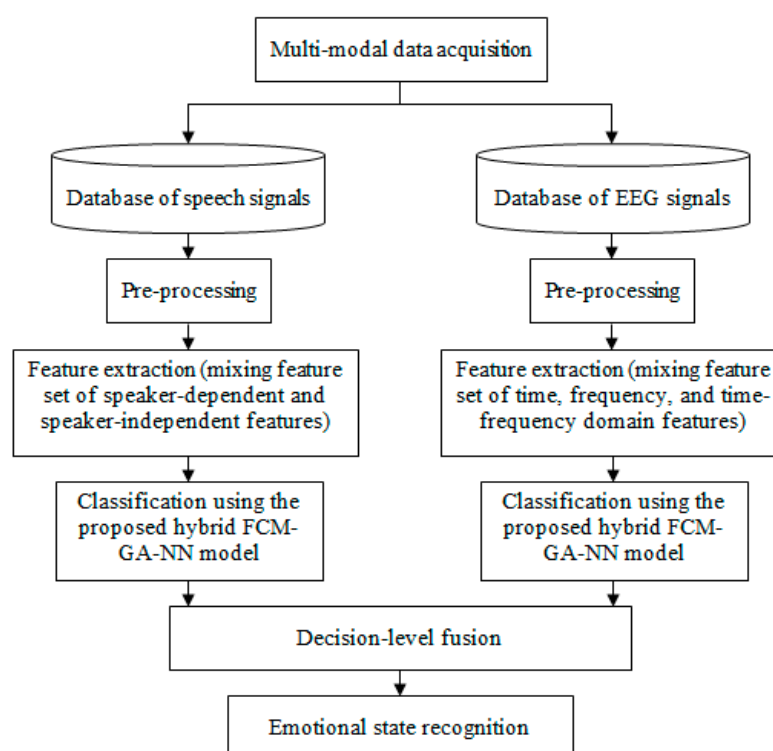

Figure 1. Architecture of the multi-modal emotion aware system.

3.1. Multimodal Data Acquisition

There are a number of multi-modal emotional databases that comprise either speech or facial expressions with different types of physiological signals. However, to the best of our knowledge, there are no available databases that integrate both speech and EEG information. The details about multi-modal data collection could be found in [45]. In the literature, there are different types of stimuli that have been utilized by the researchers to induce emotions. In [7], different genres of music were chosen as stimuli for inducing EEG-based emotions. In [46], a hypermedia system, namely MetaTutor, was utilized as stimuli for students learning about a complicated science topic.

To collect the multi-modal data in the current study, the subjects were 36 girls from the College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University (PNU) for girls, KSA, who voluntarily participated. Their ages range was between 18 and 20 years. The subjects are all Arabic native speakers. The study with its methodology was implemented under institutional ethical review of the University. The steps of the research ethics approach that we adopted are applied as follows:

- Before the experiment, students were informed of the experiment purpose, and each completed a consent form subsequent to an introduction about the steps of simulation.
- Identities of subjects were kept anonymous and confidential, where the personal information will not be ever associated or disclosed with any answer.
- Information acquired from subjects was employed only for the aim of the current research.
- The accuracy and suitability of health and medical information used in current research were confirmed by experts.

The materials involved 35 music tracks, which were employed as external stimuli for inducing emotions. The English music utilized in this research involved 5 genres: Namely, electronic, metal, rock, rap, and hip-hop. These genres are considered to cover a range that produces discernable emotions. Each participant listened to 7 songs from each genre used (one by one); each of which was followed by 15 s of silence in order to allow for emotion labeling. In this context, the total number of tracks that every participant listened to was 35 (5 genres × 7 songs, each of which represents a type of emotion classes), as demonstrated in Table 4.

Only 60 s of every song was employed to avoid inducing anxiety and boredom in a participant's brain, as depicted in Figure 2. The subjects annotated their emotional status by the end of each music track picking one of the seven available emotional keywords: Fear, surprise, happy, disgust, neutral, anxiety, and sadness [7,45]. To avoid the participant's subjectivity and exaggeration while expressing emotions, the experiment was also noticed by an expert of psychology who judged through listening to the participant's emotional speech and did the final labeling of emotions to exclude the spoofed labeled emotion. According to the keyword used by the subject and the evaluation of the expert, the approved speech/EEG recording was mapped to its representative class. Otherwise, the unapproved one was excluded.
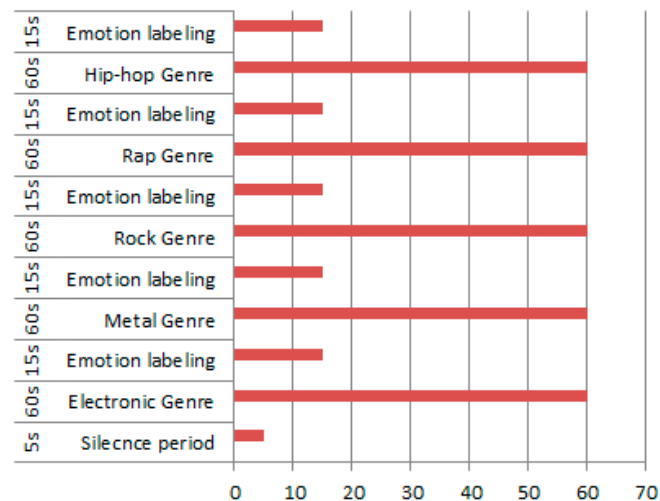

Figure 2. Structure of audio tracks that were played for participants: 5 s of silence, then 60 s from every track, and in between 15 s for performing emotion labelling.

While the participants were listening to the music tracks through headphones, the speech and EEG signals were acquired simultaneously. The speech signals were acquired spontaneously using a microphone, namely: "SHURE dynamic cardioid microphone C660N". The distance from microphone to speaker was kept at 3 m. The speech signals were sampled at 16 KHz. For Arabic speech signal acquisition, each subject had to describe the feeling induced while listening to the audio i.e., "أنا مكتئبة جداً/I am feeling deep depression".

Likewise, the EEG recordings were taken using the Emotiv-EPOC System. The device involves 14 electrodes together with 2 reference channels, which present accurate spatial resolution. The inner

sampling rate of the device was 2048 Hz prior to filtering, and its output sampling equaled 128-samples/second. The electrode placement was chosen as 10/20. This placement is commonly used in studies for EEG-based emotions recognition using visual and audio stimuli. This system relies on relationships of diverse electrode positions existing on the scalp as well as the main cerebral cortex side [38]. Figure 3 depicts the 16 electrodes, "AF3, F7, F3, FC5, T7, CMS, P7, O1, O2, P8, DRL, T8, FC6, F4, F8, and AF4", which were embedded to the record of the EEG signals. In the locations of P3/P4, the channels of average reference (CMS/DRL) were situated. The MATLAB and Puzzle box synapse were utilized for recording signals of Emotiv EPOC headset.
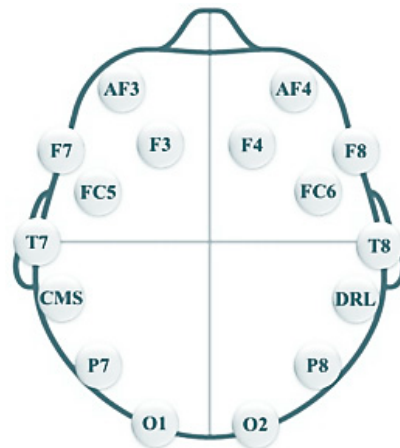

Figure 3. Emotiv-EPOC headset with 16-channel placement; 14-channels were intended to detect signals of human brain, along with 2 reference channels situated adjacent to the ears. Channel locations were based upon the 10/20 system of electrodes placement.

The data collection was implemented during the analysis [7], by using only the parts of speech and EEG data that the subjects were listening to music through, leaving out the parts of annotation. In this regard, 1260 speech samples, as well as 1260 EEG signals, were acquired from students. Thus, the total number of samples within our multi-modal emotional database was 2520. The duration of each speech/EEG signal was 60 s. Characteristics of the corpus and its statistics are demonstrated in Table 4, where emotions are sorted by arousal and valence dimensions [46], along with the representation of their classes in the corpus.

Table 4. Corpus statistics, where emotions are sorted by arousal and valence, along with the representation of their classes in the corpus.

Emotion Dimension		Valence						Others		
		Positive valence	Num. of music tracks inducing emotion	Samples number in database	Negative valence	Num. of music tracks inducing emotion	Samples number in database	Class	Num. of music tracks inducing emotion	Samples number in database
Arousal	High arousal	Happy	5	5 music tracks × 36 participants = 180 samples	Fear	5	5 music tracks × 36 participants = 180 samples	Neutral	5	5 music tracks × 36 participants = 180 samples
					Anxiety	5	180	Surprise	5	180
					Disgust	5	180			
	Low arousal				Sadness	5	180			
Total		180 samples representing positive valence-high arousal emotion classes of each single modality × 2 modalities (speech and EEG) = 360			540 samples representing negative valence-high arousal emotion classes of each single modality × 2 modalities = 1080 180 samples representing negative valence-low arousal emotions of each single modality × 2 modalities = 360			360 samples for each single modality × 2 modalities = 720		

3.2. Speech Signal Pre-Processing

To identify diverse speech multi-styles and emotion states, silent features had to be isolated. As the recorded signals are of distinct sampling frequency rates, we down-sampled all signals to 8 KHz. Then, signals were sub-divided into non-overlapping segments (frames) of 32 ms (256 samples). This frame length gave the best performance in this paper. According to [47], the unvoiced parts are eliminated from signals based upon existing energy within the frames and the frames with lower energy are eliminated before the feature extraction takes place. The speech segment (of 256 samples) was classified as voiced, unvoiced, or silence by computing its discrete wavelet transform (DWT) at scales $m = 2^1$ through $m = 2^5$ and computing the signal energy over each scale. Then, a decision scheme based on scale-energy was employed for detecting voiced speech segments.

As the unvoiced speech is of a high frequency nature, its DWT energy was high at scale $m = 2^1$. Thus, this method determines the scale at which the DWT of signal reaches the highest energy. The segment is considered unvoiced if its DWT reaches its highest energy at scale $m = 2^1$. Otherwise, it may be a silence or voiced segment relying on its energy over the higher scales. Accordingly, the DWT energy over the scale $m = 2^3$ of each segment, which was not classified as unvoiced, is compared to a predefined threshold. Thus, the segments that exceed this threshold are classified as voiced. Otherwise, they are classified as silence.

According to [24], the median of segment energies over the scale $m = 2^3$ gives a good criterion for discriminating the silence speech signals from the voiced ones. Thus, the threshold used in this regard was the median of segment energy distribution at the silence and voiced speech segments computed over scale $m = 2^3$. The other voiced frames were then concatenated and the glottal waveforms were obtained using inverse filtering as well as linear predictive analysis approach. The resulting speech and glottal waveforms were filtered utilizing a 1st-order pre-emphasis filter [24] expressed by Equation (1). The parameters of such filter were set as in [48].

$$H(S) = 1 - a * S^{-1} 0.9 \leq a \leq 1.0 \tag{1}$$

where $S \rightarrow$ the speech signal and $a \rightarrow$ takes a value of 0.9375.

3.3. EEG Signal Pre-Processing

EEG signals comprise multiple extrinsic and intrinsic artifacts that obscure the waves of brain. Extrinsic artifacts (i.e., wiring noise of the EEG sensor) have different frequencies that may interfere with the brain waves. Eliminating these artifacts requires filtering frequencies that are out of EEG signals scope. Accordingly, we employed a band-pass filter with 64 Hz and 0.5 Hz of the maximum cutoff frequency and the minimum cutoff, respectively [49]. Likewise, the notch filter that filters out the signal narrow frequencies band [50], was utilized for noise isolation from the ambient electrodes' wire resulting from the signal of power line interference (i.e., 60 Hz).

To isolate intrinsic artifacts, the independent component analysis algorithm is employed to determine artifactual elements (i.e., blinking, eye movement) found in EEG recordings. This algorithm eliminates intrinsic artifacts from signal without loss through determination of the artifactual EEG elements and subtraction of the elements, which are associated with intrinsic artifacts for obtaining a cleaner EEG signal. It has also been frequently utilized in EEG clinical studies for detecting and eliminating intrinsic artifacts [51].

3.4. Speech Feature Extraction

Recently, most of the speech emotion recognition systems shown in the literature are trained only using a specific emotional database of particular persons. This may cause drawbacks like the lake of generality. As a result, the emotion recognition accuracy of any person who does not belong to this database will be low. Thus, eliminating the speech individual difference is the primary step to enhance the speech emotion recognition accuracies and the system universality. In this stage,

feature extraction is implemented after dividing the speech signals into a sequence of non-overlapped sub-frames/segments, each of which has a length of 256 samples. Then, the speaker-dependent and speaker-independent features [20,21,52] are computed from each frame as follows. The selected features have been shown in the literature to carry an important emotional content [20,21,52].

(a) Speaker-dependent features: Usually involve valuable personal emotional information [52], as demonstrated in Table 5. They comprise fundamental frequency, fundamental frequency four-bit value [20], Mel-frequency cepstral coefficients (MFCC), etc.

(b) Speaker-independent features: These features are used to suppress the speaker's personal characteristics. In [20], the speaker-independent features from emotional speech involve the fundamental frequency average change ratio, etc.

3.5. EEG Feature Extraction

In this paper, we use a mixing set of EEG features that have been proven in the literature to be highly effective in emotion recognition [26,29]. These features belong to time, frequency, and time–frequency. The EEG signal is blocked into windows/frames. From the literature, the efficient window size ranges from 3 to 12 s, when classifying the individuals' mental state utilizing EEG signals [53]. Two methods of windowing i.e., fixed windowing and sliding windowing, were tested to choose the superior method in reference to classification accuracy. In this paper, we estimate 23 measures of time, frequency, and time–frequency presented in Table 6, from 14 EEG channels.

Table 5. The estimated speaker-dependent and speaker-independent features.

Speech Features	Characteristics				
	Prosodic Features			**Speech Quality Features**	**Spectrum Features**
	Fundamental frequency-related	Energy-related	Time length correlation-related		
Speaker-dependent	Fundamental frequency	Short-time maximum amplitude	Short-time average crossing zero ratio	Breath sound	12-order MFCC
	Fundamental frequency maximum	Short-time average energy	Speech speed	Throat sound	SEDC for 12 frequency bands (equally-spaced)
	Fundamental frequency four-bit value	Short-time average amplitude		Maximum and average values of first, two as well as three formant frequencies	Linear Predictor Coefficients (LPC)
Speaker-independent	Fundamental frequency average change rate	The Short-Time-Energy average rate	Partial time ratio	Average variation of 1st, 2nd, and 3rd formant frequency rate	1st-order difference MFCC
	Standard deviation of fundamental frequency	The amplitude of short-time energy		Standard deviation of 1st, 2nd, and 3rd formant frequency rate	2nd-order difference MFCC
	Change rate of four-bit point frequency			Every sub-point value of 1st, 2nd, and 3rd formant frequencies change ratio	

Table 6. Time, frequency, and time–frequency domains features, extracted from EEG signals in this work.

Domain of Analysis	EEG Feature	Explanation	Equation
Time domain	Cumulative maximum	Highest amplitude of channel M until sample R	$Max_{RM} = max(EEG_{1:R,M})$
	Cumulative minimum	Lowest amplitude of channel M until sample R	$Min_{RM} = min(EEG_{1:R,M})$
	Mean	Amplitude average absolute value over the various EEG channels	$Mean_M = \dfrac{\sum_{R=1}^{K} EEG_{RM}}{K}$
	Median	Signal median over the various channels	$Median_M = sort(EEG)\dfrac{K+1}{2}, M$
	Standard deviation	EEG signals deviations over the various channels in every window	$SD_M = \sqrt{\dfrac{1}{K-1}\sum\limits_{R=1}^{K} EEG_{RM}{}^2}$
	Variance	EEG signal amplitude variance over the various channels	$V_M = \dfrac{1}{K-1}\sum\limits_{R=1}^{K} EEG_{RM}{}^2$
	Kurtosis	Reveals the EEG signal peak sharpness	$Kur_M = \dfrac{\frac{1}{K}\sum_R (EEG_{RM} - Mean_M)^4}{\left(\frac{1}{K}\sum_R (EEG_{RM} - Mean_M)^2\right)^2}$
	Smallest window components	Lowest amplitude over the various channels	$S_M = \min\limits_{R} EEG_{RM}$
	Moving median using a window size of n	Signal median with channel M and a widow of n -samples size	$Mov_{R,M} = Median\left(EEG_{R:R+n-1,M}\right)$
	Maximum-to-minimum difference	Difference among highest and lowest EEG signal amplitude over the various channels	$Max - Min_M = \max\limits_{R} EEG_{RM} - \min\limits_{R} EEG_{RM}$
	Peak	Highest amplitude of EEG signal over the various channels in time domain	$P_M = max EEG_{RM}$

Table 6. *Cont.*

Domain of Analysis	EEG Feature	Explanation	Equation
Frequency domain	Peak to Peak	Time among EEG signal's peaks over the various windows	$PTP_M = PL_M - \underset{R,M \neq LP_M RM}{arg\ max}$
	Peak location	Location of highest EEG amplitude over channels	$PL_M = \underset{RRM}{arg\ max}$
	Root-mean-square level	EEG signal's Norm 2 divided by the square root of samples number over the various EEG channels	$Q_M = \sqrt{\dfrac{\sum_{R=1}^{K} EEG_{RM}{}^2}{K}}$
	Root-sum-of-squares level	EEG signal's Norm over the distinct channels in every window	$RL_M = \sqrt{\sum_{R=1}^{K} |EEG_{RM}|^2}$
	Peak-magnitude-to-root-mean-square ratio	Highest amplitude of EEG signal divided by the Q_M	$PM_M = \dfrac{\left\| EEG_{:,M} \right\|\infty}{\sqrt{\dfrac{\sum_{R=1}^{K}|EEG_{RM}|^2}{K}}}$
	Total zero crossing number	Points number where the EEG amplitude sign changes	$ZC_M = |\{R|EEG_{RM} = 0\}|$
	Alpha mean power	EEG signal power *Pow* in channel *M* in an interval of $[[8H, 15H]]$	$\alpha_M = Pow(EEG_{:,M}, F \in [8Hz, 15Hz])$
	Beta mean power	EEG signal power in Beta interval	$\beta_M = Pow(EEG_{:,M}, F \in [16Hz, 31Hz])$
	Delta mean power	EEG signal power in Delta interval	$\delta_M = Pow(EEG_{:,M}, F \in [0Hz, 4Hz])$
	Theta mean power	EEG signal power in Theta interval	$\theta_M = Pow(EEG_{:,M}, F \in [4Hz, 7Hz])$
	Median frequency	Signal power half of channel *M* which is distributed over the frequencies lower than MF_M.	$Pow(EEG_{:,M}, F \in [0Hz, MF_M]) = Pow(EEG_{:,M}, F \in [MF_M, 64Hz])$
Time-frequency domain	Spectrogram	The spectrogram (short-time Fourier transform), is computed through multiplying the time signal by a sliding time-window, referred as (M). The time-dimension is added by window location and one outputs time-varying frequency analysis. *o* refers to time location and *K* is the discrete frequencies number.	$SP_R = \sum_{R=0}^{M-1} EEG(M)W(M-o)exp\left(-\dfrac{j2\pi MR}{K}\right),$ $where\ 0 \leq M \leq (K-1)$

3.6. The Classifier

3.6.1. The Basic Classifier

NN is a parallel processor distributed massively [41]. It involves single input layer, besides one output layer, as well as one hidden layer. Such a hidden layer is like a connection in between input/output layers through multiple weights, nodes, biases, in addition to activation functions. The operational procedure of NN is denoted by the subsequent equations:

$$W_j = \sum_{t=1}^{k} \omega_{jt} X_t \tag{2}$$

$$Y_j = \phi\left(W_j + A_j\right) \tag{3}$$

where j indicates a neuron in the network, k is input parameters number, $X_t(t = 1, 2, \ldots, k)$ represents the t-th input parameter, in addition to $\omega_{jt}(t = 1, 2, \ldots, k)$, denotes respective synaptic weight. Initially, every synaptic weight ω_{jt} multiplies the input sample X_t, which conforms to it. Values of weight are put in a summator $\sum$, getting an output, indicated as W_j. Afterwards, W_j is applied to the activation function ϕ for producing an output signal Y_j, and A_j denotes the bias.

This study deals with a multi-class classification problem where N classes of emotions are to be recognized. According to the literature [54], the multi-class neural network (NN) classification can be implemented using: (1) Either a single neural network model with K outputs or (2) a multi-NN model with hierarchical structure. In this regard, there are two methods that can be employed to model the pattern classes, one-against-all (OAA) approach, and one-against-one (OAO) approach [54–56].

In this paper, we present a multi-class back propagation neural network for emotion classification. We adopt the OAA approach scheme, which operates on a system of $M = N$ binary neural networks, NN_j, $j = 1, \ldots, N$, where each single network, NN_j, holds one output node P_j that has an output function F_j being modulated to output $F_j(\bar{x}) = 1 or 0$ to determine whether or not the input sample $\bar{x}$ belongs to class j. Accordingly, each neural network NN_j is trained using the same dataset but with different class labels.

For training the *jth* neural network NN_j, the training set Ω_T is subdivided into two sets, $\Omega_T = \Omega_T^j \cup \overline{\Omega}_T^j$, where Ω_T^j includes all the class j samples, which obtain the label 1, and $\overline{\Omega}_T^j$ includes all the samples belonging to all of the rest classes, which obtain the label 0. The decision function considers the activation function output at every neural network NN_j, and outputs the label of the class that confronts to neural network NN_j, which obtains the highest output value by the activation function of the P_j output node:

$$F(\bar{x}, y_1, \ldots, y_M) = \underset{j=1,\ldots,M}{arg\ max}\left(y_j\right). \tag{4}$$

In this study, we constructed an OAA system of 7 binary NNs with 20 hidden nodes for each of which. Selecting the appropriate parameters for the multi-class back-propagation neural network was based upon minimizing the root-mean-square-error (RMSE) between target and the predicted output. The parameters giving the best accuracy are demonstrated in Section 4.

3.6.2. Proposed Classifier Design

The proposed hybrid FCM-GA-NN model is depicted in Figure 4. The chromosome fitness (feature selection outcome) is evaluated based on the optimal fuzzy clusters that reduce the NN classification error with no prior knowledge about the number of fuzzy clusters. Thereafter, chromosomes are encoded utilizing the resulting fuzzy memberships. Since the cluster number u is not determined, the population may have chromosomes with uniform or different lengths. Therefore, a modified single-point crossover algorithm is developed. Accordingly, the mutation algorithm changes every

dimension of a solution by a probability P_T. Then, the fuzzy memberships are computed, and the clustering indices of the object are taken based on the maximum fuzzy membership values respecting different fuzzy clusters. In this regard, some empty classes may be found when crossing two parents of different lengths, or if a class number is of big size. In case of an empty j-*th* class, the center u_j will be eliminated from $\{u_1, u_2, \ldots\ldots, u_J\}$. Update of the new memberships and centroids with the remaining centroids is accordingly implemented. In the last generation, the algorithm generates a non-dominated set of solutions whose number changes in reference to population size. Such solutions are regarded as equal in relation to the fitness values obtained by fitness function. The final clustering outcome is determined based upon a clustering ensemble method. Algorithm A1 reveals the pseudo-code of the whole proposed model Appendix A.
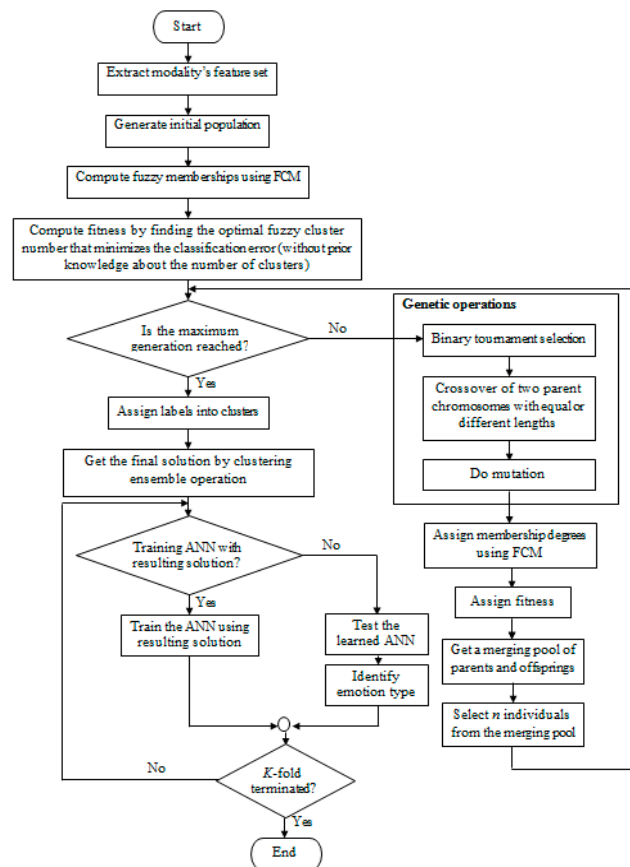
Figure 4. Flowchart of the proposed hybrid FCM-GA-NN model for recognizing human emotions.

The steps of the proposed algorithm are as follow:

(a) Fuzzy-clustering of each modality's emotional dataset

The purpose of a clustering algorithm goes to separation of resembling objects, i.e., feature vectors into shared cluster and unalike objects into distinct clusters; thus, the within-cluster objects similarity measure is large and is contrarily small for the between-cluster objects. In this paper, each modality's feature set is separated into training and testing sets. Consequently, the FCM algorithm is executed to cluster the relevant emotion vectors in each modality's training set individually into several categories, i.e., emotional speech classes as well as emotional EEG classes. The developed clustering procedure is demonstrated in Equations (5)–(9) of Algorithm A2,

$$\sum_{i=1}^{u} \mu_{i,k}^m = 1, k = 1, 2, \ldots, K \tag{5}$$

$$0 < \sum_{k=1}^{K} \mu_{i,k}^m < K, i = 1, 2, \ldots, u \tag{6}$$

$$R_i^m = \frac{\sum_{k=1}^{K} (\mu_{i,k}^m)^a \cdot x_k}{\sum_{k=1}^{K} (\mu_{i,k}^m)^a}, i = 1, \ldots, u, 1 < a \tag{7}$$

$$\mu_{i,k}^{m+1} = \left[\sum_{h=1}^{u} \left(\|x_k - R_i^m\|^2 / \|x_k - R_h^m\|^2 \right)^{\frac{1}{a-1}} \right]^{-1} \tag{8}$$

$$K_a^{m+1} = \sum_{k=1}^{K} \sum_{i=1}^{u} \left[(\mu_{i,k}^{m+1})^a \cdot \|x_k - R_i\|^2 \right] \tag{9}$$

where:

$K \rightarrow$ the number of emotional classes in each modality's dataset,

$x_k \rightarrow$ the emotion feature vector of the sample k,

$\mu_{ik}^m \rightarrow$ the emotion feature vector's membership grade belonging to cluster i at time m, $0 < \sum_{k=1}^{K} \mu_{ik}^m < K$,

$\varepsilon \rightarrow$ decides time complexity and precision for clustering. (in this paper, it was set as 10^{-10}).

(b) Chromosome representation

After applying fuzzy clustering, a $c \times b$ matrix, representing fuzzy cluster centers, can be considered as a chromosome in GAs terminology. Every chromosome's gene is indicated by an element of U of Equation (10). To express the decision variables, a matrix U is transformed into a vector $= [B_{11}, \ldots, B_{K1}, .., B_{1n}, \ldots, B_{Kn}]$, which is encoded as a single chromosome whose length is changing according to n (see Figure 5).

$$U = \begin{bmatrix} B_{11} & B_{12} & B_{13} & \ldots & B_{1n} \\ B_{21} & B_{22} & B_{23} & \ldots & B_{2n} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ B_{K1} & B_{K2} & B_{K3} & \ldots & B_{Kn} \end{bmatrix} \tag{10}$$
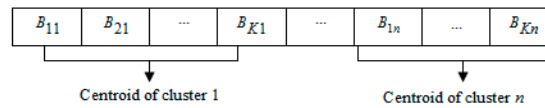

Figure 5. Chromosome representation.

(c) Fitness evaluation

Applying the FCM algorithm attempts to determine optimal number of clusters, that is, for cluster validity. The WB index [57] is utilized to find cluster number u that diminishes the intra-cluster variance $acv(\mu, R)$ and increases the inter-cluster variance. On this basis, the feature selection result (the fitness) of the chromosome is assessed by finding the optimal fuzzy clusters that minimizes the classification error without prior knowledge on the clusters number.

To handle that multi-criteria decision-making issue, we used a standalone weighted fitness function combining two single processes into one objective. According to Equation (11), two predefined weights W_u and W_c are associated with optimal number of clusters u_p and classification error , respectively.

$$Fit_{FCM-GA-NN} = W_u[u_p] + W_c[ce] \tag{11}$$

The optimal cluster number u_p is obtained by:

$$u_p = \arg \max_u WB = \frac{ecv(\mu, R)}{acv(\mu, R)} \tag{12}$$

where:

$$acv(\mu, R) = \sum_{i=1}^{u} \sum_{k=1}^{K} (U_i)^{-1} \|x_k - R_i\|^2 \tag{13}$$

$$ecv(\mu, R) = \frac{1}{U_2^u} \sum_{\lambda=1}^{u-1} \sum_{a=\lambda+1}^{u} (U_i)^{-1} * \left(\sum_{k \in v_\lambda \cup v_a} \frac{(\mu_{\lambda k} \cdot \mu_{ak})}{|v_\lambda| + |v_a|} \right)^{-1} \|R_\lambda - R_a\|^2 \tag{14}$$

$$U_i = \sum_{k=1}^{K} \mu_{ik}^2 / |v_i|, i = 1, 2, \ldots, u \tag{15}$$

$$v_i = \{k | I_{ik} = 1\}, \ I_{ik} = \begin{cases} 1, & f \ \mu_{ik} = \max_{1 \le h \le u} \mu_{hk} \\ 0, & otherwise \end{cases} \tag{16}$$

and U_2^u indicates a combination computation.

The classification error is computed using:

$$ce = \frac{NE}{NT} \tag{17}$$

where $NE \leftarrow$ the number of incorrectly classified samples. $NT \leftarrow$ the total number of training instances.

(d) Binary tournament selection

Binary tournament selection technique [58] is utilized in this work to choose parents for creating new generation. Accordingly, two individuals are picked haphazardly for playing a tournament. During that, the winner is picked by $\prec_n$, referred to as a crowded comparison operator. Such operator relies on two attributes denoted as non-domination rank (A_{rank}) and crowding distance (A_{dist}). If A and B are assumed to be two individuals, $\prec_n$ can be defined as follows. In applying this approach, chromosomes of 40% top-ranking are picked for producing the child chromosomes through crossover and mutation procedures.

$$A \prec_n B if (A_{rank} < B_{rank}) or (A_{rank} = B_{rank}) and (A_{dis} < B_{gist}) \tag{18}$$

(e) Crossover procedure

In this framework, the length of every chromosome within the population is denoted by $K \times n$. The two parent chromosomes might be with equal or unequal lengths relying on the K values; therefore, each cluster centroid may be indivisible while crossing two parents. Thus, a modified single-point crossover algorithm is developed in this paper.

Assuming that $R = \{r_1, r_2, r_3, r_4\}$ and $M = \{m_1, m_2, m_3, m_4\}$ are two parent solutions having four cluster centers, where each r_i and m_i represents a feature vector, D^1 and D^2 are two children created. To perform uniform crossover, each two centers from parents are crossed with a probability of 0.5. As depicted in Figure 6, the crossover is not implemented to gene 3, and values of r_3 and m_3 are copied to gene 3 of children D^1 and D^2, respectively. Values of genes 1, 2, and 4 have been updated to y_1, y_2, y_4 and z_1, z_2, z_4 on D^1 and D^2, respectively. If the two parents are with different lengths (see Figure 6), the centroids considered for crossover in the longer parent are randomly picked. Thus, r_1, r_2, r_3, r_5 are chosen and crossed with m_1, m_2, m_3, m_4, respectively. Then, the omitted centers values (r_4, r_6) are copied to an offspring. Algorithm A3 demonstrates the crossover procedure between parents. The new centers' values for two offspring are calculated using Equations (19)–(20) of Algorithm A3.

$$D_j^1 = \frac{[(1 + beta) * Q_1 + (1 - beta) * Q_2]}{2} \tag{19}$$

$$D_j^2 = \frac{[(1 - beta) * Q_1 + (1 + beta) * Q_2]}{2} \tag{20}$$

(f) Mutation procedure

In this procedure, a small probability of mutating P_T is assigned to each gene, determined by randomly generating a number (the gene is mutated if the generated number is less than P_T, otherwise not). According to this framework, a change of a gene within a chromosome will trigger a sequence of

gene changes. Thus, the fuzzy memberships of a chromosome point will be chosen to mutate together using a probability P_T. The mutation procedure is demonstrated in Algorithm A4.
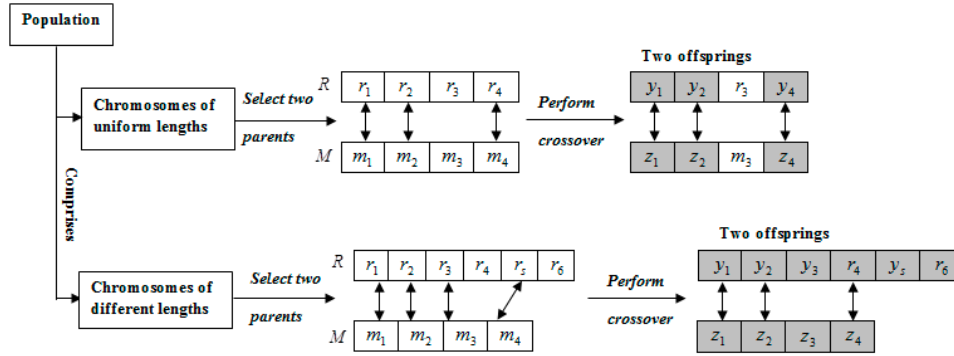

Figure 6. Crossover procedure of two chromosomes in our framework.

(g) Obtaining final solution

The cluster validity indices are utilized by taking their ranking numbers, so we can choose some solutions to integrate. The ranking of every cluster validity index is illustrated as follows. Assuming the cluster validity indices number is V, a corresponding ranking vector of every solution is $R(s) = (r_1(s), r_2(s), \ldots, r_V(s))$, $r_j(s) \in \{1, 2, \ldots, n\}$, $1 \le j \le V$. Besides that, the sum of cluster validity indices for every solution s is computed. Accordingly, the sum ranking number for every solution is computed as $r_{sum}(s) \in \{1, 2, \ldots, n\}$, which is integrated to $R(s)$ for generating a new vector referred as $R'(s) = (r_1(s), r_2(s), \ldots, r_{V+1}(s))$.

Therefore, a total ranking number of s is $R_t(s) = \sum_{j=1}^{V+1} r_j(s)$. The smallest value of R_t may be $V + 1$ if all cluster validity indices for a single solution are the best. Algorithm A5 shows the steps to get final clustering solution. The whole process is illustrated in Figure 7. Many clustering ensemble algorithms have been compared in the literature [59]. As the meta-clustering algorithm showed the highest performance in terms of accuracy average rates (AVR), it has been applied in Step 13 of Algorithm A5.
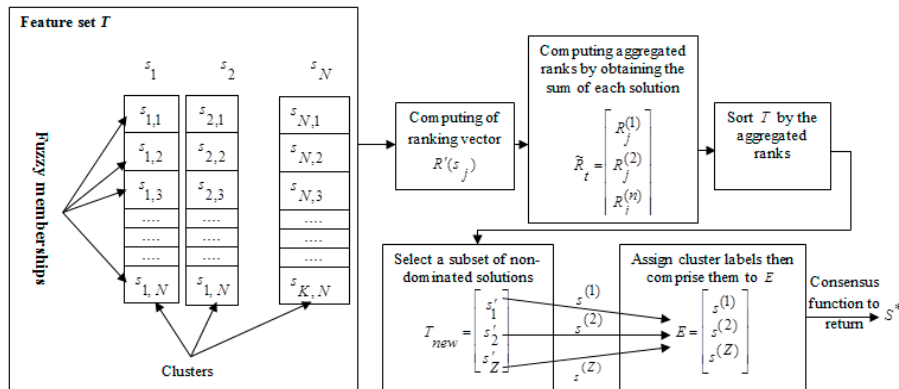

Figure 7. Clustering ensemble based final solution.

(h) Decision-level fusion

A decision-level fusion first processes every modality separately, and then integrates the outcomes from their classifiers to reach the final decision. Motivated by [60], for every trial, assume that M^x_{Speech}, M^x_{EEG} and $M^x_0 \in [0, 1]$ denoting the probability of classifier for class $x \in [1, 2, 3]$ for speech, EEG, and fusion, each in order. Subsequently, the class probabilities are obtained using

$$M^x_0 = \beta M^x_{Speech} + (1 - \beta) M^x_{EEG} \tag{21}$$

where β is a modality weight. Subsequently, we tested two approaches to calculate β.

Equal weights fusion method (EWF) that makes fusion on a decision-level, where each class final probabilities are estimated using the class probability taken from every sole modality. That is,

$$M_0^x = M_{Speech}^x + M_{EEG}^x.\tag{22}$$

Learned weights fusion method (LWF) is a different method for the decision-level fusion, where an optimal decision weight from each modality is numerically approximated. This is implemented by changing from 0 to 1 and selecting the value that outputs the superior accuracy for the training set. Then, an estimated weight is implemented to the sample, by Equation (22).

4. Experimental Results

This section discusses how the system is evaluated and compared to the state-of-the-art systems. For testing the proposed FCM-GA-NN model performance, we not only conduct our experiments on the collected multi-modal emotion database, but also on two different public unimodal datasets speech and EEG, which are used in the literature for the purposes of comparisons.

4.1. Performance Evaluation Metrics

In this paper, the performance is assessed through four criteria: AVR, root mean squared error (RMSE), percent deviation (PD), and correlation coefficient (ρ). Accuracy is referred to as a truly classified samples ratio, T over the total samples number considered for classification, N. The RMSE is employed for measuring the difference between both predicted and actual values. The correlation coefficient reveals the degree of approximation between estimated and actual values. Calculating the aforementioned indicators is as follows:

$$AVR = \frac{T}{N}\tag{23}$$

$$RMSE = \sqrt{\frac{1}{Num} \sum_{i=1}^{Num} (\mu_x - \mu_y)^2}\tag{24}$$

$$PD = \frac{1}{Num} \sum_{i=1}^{Num} \frac{|\mu_x - \mu_y|}{\mu_x} \times 100\tag{25}$$

$$\rho = \frac{\sum_{i=1}^{Num}(\mu_y - \overline{\mu_Y})(\mu_x - \overline{\mu_X})}{\sqrt{\sum_{i=1}^{Num}(\mu_y - \overline{\mu_Y})^2 \sum_{i=1}^{Num}(\mu_x - \overline{\mu_X})^2}}.\tag{26}$$

$Num \leftarrow$ the data points number.
$\mu_y \leftarrow$ the output predicted by the model.
$\mu_x \leftarrow$ the sample actual value.
$\overline{\mu_Y} \leftarrow$ the mean of μ_y .
$\overline{\mu_X} \leftarrow$ the mean of μ_x .

4.2. Cross Validation (CV)

For each single modality, the accuracy of the hybrid classification algorithm is evaluated using the nested cross validation [61,62]. In this method, two nested CV loops are used to test the classifier, as shown in Figure 8. For the outer loop, the 1260 samples of each modality are split into seven folds. This means that in each fold we will have 210 testing samples and 1050 training samples. Every 210 samples are used as validation set (outer testing set), and the other 1050 samples are combined as outer training set. This procedure is repeated for each fold. Then, each outer training set is split into seven folds. Therefore, each fold will have 175 testing samples and 875 training samples.

A single set of 175 samples is accordingly used for validation (inner testing set), and the 875 samples are used as inner training set. This is implemented repeatedly for each fold. In this manner, the

inner CV is used for selecting the best parameters that reduces the *RMSE* of our algorithm. For instance, the activation functions for the hidden and output layers, the learning rate, the maximum generations, the mutation ratio, the reproduction ratio, etc. The outer CV is used for final model testing.
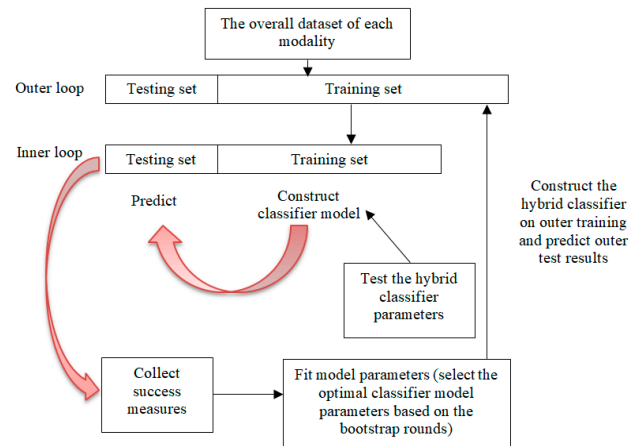

Figure 8. Nested cross validation.

4.3. First Experiment: Comparison of the Proposed FCM-GA-NN Model to the Developed $FCM-GA-NN_{fixed}$

To reveal the efficiency of the proposed FCM-GA-NN model, it is crucial to compare it with another developed model that uses FCM and GA as hybrid to train NN [63]. The second model is centroid-based and requires previously a fixed cluster number. Contrarily, our model finds optimal fuzzy clusters with no prior knowledge on their number. Accordingly, the proposed model's errors are compared with the second hybrid model. Figure 9 depicts the bar graph illustrations representing the AVR, *RMSE* s, *PDs*, and correlation coefficients ρs among both measured and predicted values from the two models.
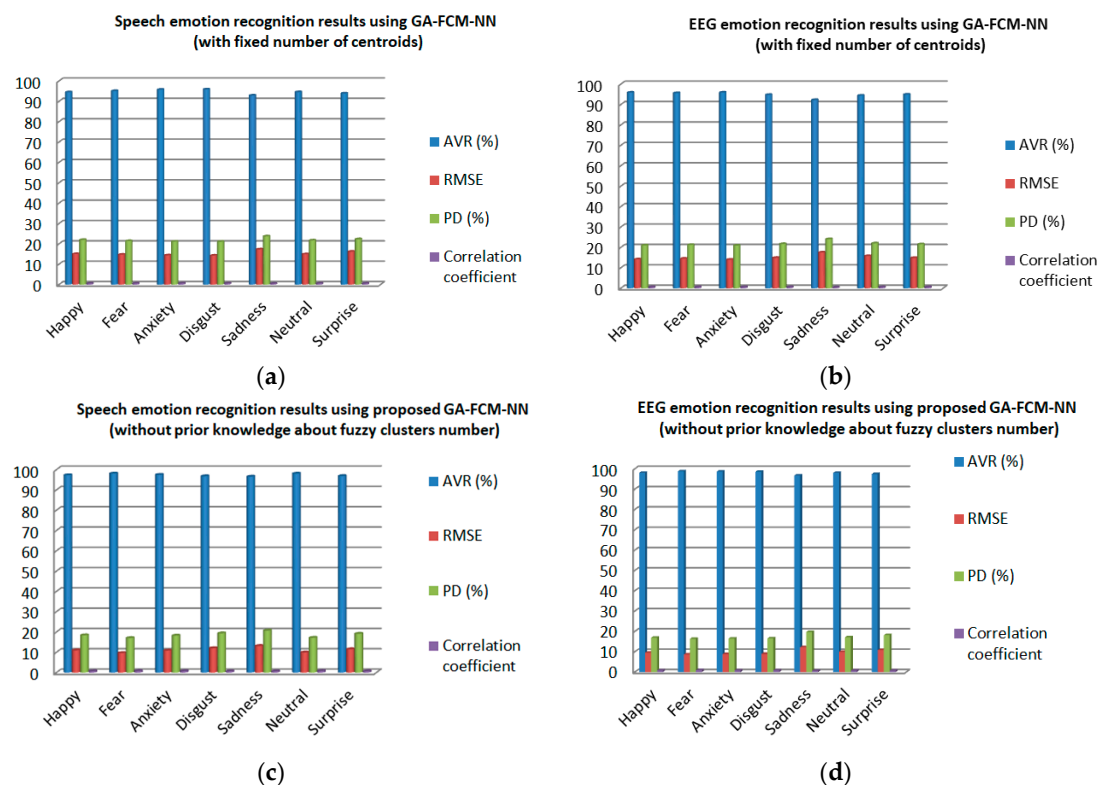

Figure 9. The overall evaluation results for the proposed FCM-GA-NN model, and $FCM-GA-NN_{fixed}$ model, on different emotion classes of each single modality (speech and EEG).

Taking fear class as an instance, the AVR rates of the proposed FCM-GA-NN model are 98.85% and 98.08% for EEG and speech modalities, respectively. By comparison, the AVRs using $FCM - GA - NN_{fixed}$ model are smaller than those of the proposed model, 95.66%, and 95.11% for EEG and speech modalities, respectively. The percent deviations (*PDs*) are 21.42% and 21.24% for speech and EEG modalities, respectively, which are also higher than the values from the proposed model. Furthermore, the proposed model exhibits ρ s of 0.9511 and 0.841, for EEG and speech modalities, respectively, which are higher than those obtained using the other model. The parameters that give the best accuracy for the two models are illustrated in Table 7.

Table 7. The parameters that give the best accuracy in this work.

Parameter	Value
Number of layers for each NN	3 (input, hidden, output)
Number of hidden nodes in each NN	20
Activation functions for the hidden and output layers	tansig-purelin
Learning rule	Back-propagation
Learning rate	0. 1
Momentum constant	0. 7
Units of population	150
Maximum generations	50
Mutation rate	0.2
Crossover rate	0.5

4.4. Second Experiment: Speech Emotion Recognition

The second experiment depicted in Figure 10 was intended to investigate how robust the proposed hybrid fuzzy-evolutionary NN model is for speech emotion recognition. Accordingly, two databases for speakers of two different languages were used to test the model. The first dataset is collected during this study. The second dataset is a public one (SAVEE) [24]. For each dataset, the emotional features are extracted from the speech testing samples. Further, the proposed classifier is carried out.
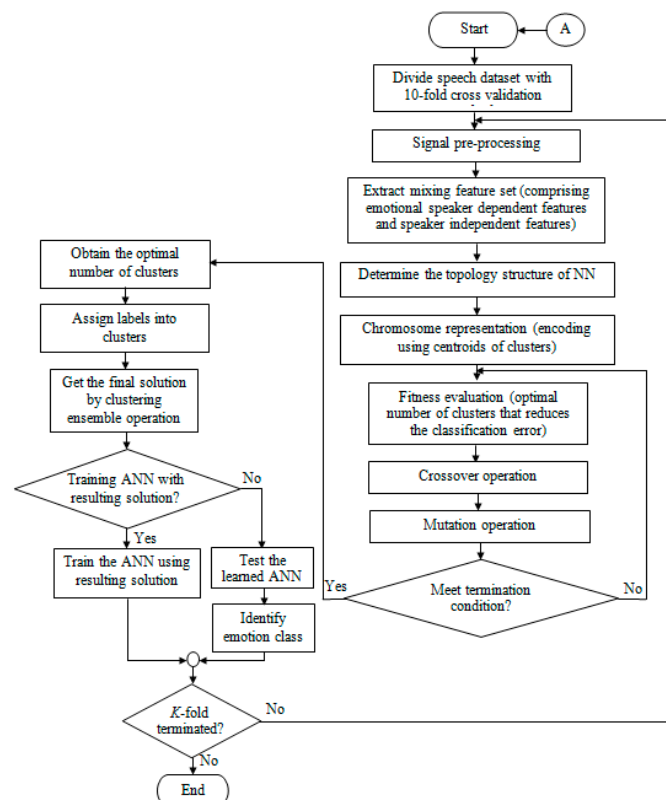

Figure 10. Cross validation of speech emotion recognition using the proposed FCM-GA-NN model.

The comparative results for the proposed model over the two speech datasets are illustrated as follows:

(a) From Figure 9c, it can be vividly recognized that using the collected dataset, the proposed model achieved higher AVRs of 98.08%, 98.03%, 97.44%, 97.24%, 96.87%, 96.76%, and 96.55% for fear, neutral, anxiety, happy, surprise, disgust, and sadness classes, respectively. Likewise, it can be observed that during training the proposed model exhibits lower *RMSE* of 9.6233, 9.9299, 10.8219, 10.9501, 11.5206, 12.0312, and 13.1127, respectively, for the aforementioned classes. The model also gives lower *PDs* of 16.99%, 17.15%, 18.19%, 18.35%, 19.13%, 19.32%, and 20.67%, respectively, for the same classes.

(b) From Figure 11a, the proposed model achieved higher AVRs using the SAVEE dataset, 98.98%, 98.96%, 98.93%, 98.11%, 97.83%, 97.42%, and 97.21% for surprise, anxiety, sadness, happy, fear, neutral, and disgust classes, respectively. The minimum and maximum *PDs* occur at surprise class (15.31%) and disgust class (18.49%), respectively. The highest correlation coefficient ρ equals 0.9801, which is found at surprise class, while the lowest value is 0.733 and noted at disgust class.

By comparison, it is observed that the estimated measures give superior results over the two datasets. These results reflect the robustness of the mixing speech feature set in combination with the proposed hybrid model. On the other hand, there are low noticeable differences with respect to AVRs, *RMSE* s,*PDs* , and ρ s, obtained on emotion classes of the two datasets, for the favor of the SAVEE dataset. These differences may be due to the difference in the stimuli mode used to induce emotions. Whereas our speech dataset was acquired using music stimuli, the SAVEE dataset was collected using video stimuli.

Eventually, we compared our results to the other published results obtained using the public database of SAVEE, as illustrated in Table 8 [24,64,65]. It is obvious that our model yields superior results which outperform the state-of-the-art on speech emotion recognition, in terms of classification accuracy, where the total AVR achieved was 98.21%. This result supports the previously published works that have reported the powerful of the hybrid classification approaches in optimizing the speech recognition results in general [66,67].
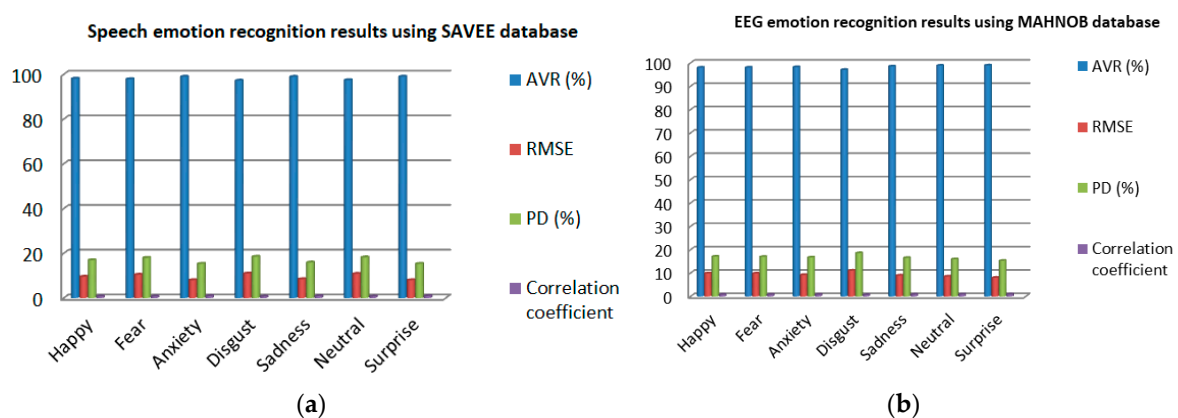

Figure 11. The overall evaluation results for the proposed FCM-GA-NN model over two publicly available databases: (**a**) SAVEE database for speech emotion recognition, and (**b**) MAHNOB for EEG emotion recognition.

Table 8. Comparison of the proposed model accuracy to the state-of-the-art accuracies obtained using SAVEE database.

Corpus	Reference	Year of Publication	Feature Extraction	Feature Selection	Classifier	Classification Accuracy
SAVEE	[24]	2017	The study used 50 higher order features (28 Bispectral feature + 22 Bicoherence), which were combined with Inter-Speech 2010 features for improving the recognition rate.	Feature selection included to phases: Multi-cluster feature selection, and proposed hybrid method of Biogeography-based Optimization as well as Particle Swarm Optimization.	SVM and ELM.	The speaker-independent accuracies were 62.38% and 50.60%, for SVM and ELM, respectively, whereas the speaker-dependent accuracies were 70.83%, and 69.51%, respectively.
	[64]	2018	Feature set of 21 statistics	Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), PCA + LDA.	Genetic algorithm- Brain Emotional Learning model.	The speaker independent accuracy was 44.18%, when using PCA as feature selector
	[65]	2019	The openSMILE toolbox was used to extract 1582 features from each speech sample	Proposed feature selection method relies on the changes in emotions according to acoustic features.	SVM, k-nearest neighbor (k-NN), and NN.	77.92%, 73.62%, and 57.06% for SVM, k-NN, and NN, respectively.
	Proposed speech emotion recognition model		Mixing feature set of speaker-dependent and speaker-independent characteristics.	Hybrid of FCM and GA.	Proposed hybrid Optimized multi-class NN	98.21%

4.5. Third Experiment: EEG Emotion Recognition

In terms of performance comparison purposes, we used two different datasets to check the robustness of the proposed model regarding EEG emotion recognition. The first dataset was collected during current research, which contains EEG signals of only 14 channels. The second dataset is a public one (MAHNOB) [45], which comprises EEG signals of 32 channels. MAHNOB dataset comprises user response recordings to a multimedia content. Video fragments from online sources, which last among 34.9 s and 117 s, using different content, were chosen for inducing nine emotions in the participants.

Next to each video clip, the subjects were asked to express their emotional status by a keyword like neutral, amusement, surprise, happiness, anger, disgust, sadness, fear, and anxiety. For each dataset, we estimated a set of extracted features from three domains, frequency, time, and time–frequency. Thereafter, the proposed model is applied to classify each dataset into seven-class emotions of EEG signals as depicted in Figure 12. The overall comparative results for the proposed model over the two datasets are overviewed as:

Using the collected dataset, the proposed model achieved higher AVRs of 98.85%, 98.69%, and 98.57%, respectively, for fear, anxiety, and disgust classes. It also gives lower *RMSE* s of 8.7233, 8.9111, and 8.9555, respectively for these classes, as shown in Figure 9d. The lowest and highest *PDs* found at fear class (16.33%) and sadness class (19.72%), respectively. The greatest ρ is 0.9511, also found at fear class.
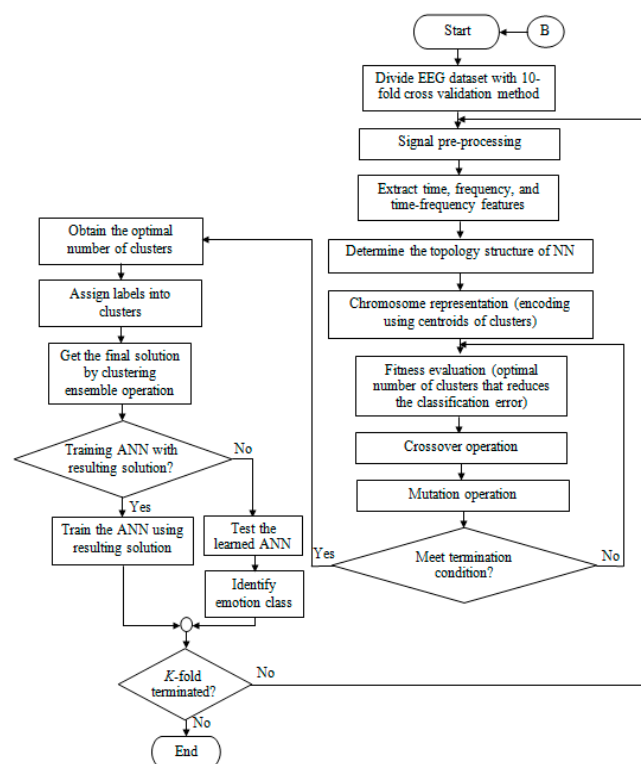

Figure 12. Cross validation of EEG emotion recognition using proposed model.

With respect to the MAHNOB dataset, the proposed model gives higher AVRs of 98.96%, 98.91%, and 98.55%, respectively, for surprise, neutral, and sadness classes. On contrast, it gives lower *RMSE* s of 8.0112, 8.5402, and 8.9555, respectively, for the aforesaid classes, as shown in Figure 11b.

In terms of performance comparisons over the two datasets, the overall AVR of the proposed classifier is 98.26%, which shows low noticeable differences on the emotion classes for the favor of MAHNOB database. This can be interpreted by the differences in stimuli mode employed to induce emotions, where MAHNOB database uses visual stimuli. From these results, we deduce that the selected EEG feature set in combination to the proposed FCM-GA-NN model is robust for classifying

emotions from EEG signals. We also conducted a comparison of the impact of fixed/sliding windowing methods on AVRs over the two datasets. As demonstrated in Table 9, the comparison reveals that fixed windowing returns better AVRs over the two compared datasets. The window size of 6 S was found to give the best AVR. We also compared our results to the other state-of-the-art results obtained using MAHNOB database, as shown in Table 10 [68–70]. The results indicate that the proposed model outperforms the state-of-the-art concerning EEG emotion recognition.

Table 9. Comparison of using fixed windowing and sliding windowing on AVRs over the two compared databases.

Dataset	AVR (%)	
	Fixed window	Sliding window
Collected database	98.06	96.93
MAHNOB	98.26	97.41

Table 10. Comparison of the proposed model accuracy to the state-of-the-art accuracies obtained using MAHNOB database.

Corpus	Reference	Year of Publication	Feature Extraction	Feature Selection	Classifier	Classification Accuracy
MAHNOB	[68]	2017	Features of time, frequency, and time–frequency.	Genetic Algorithm, Ant Colony Optimization, Particle Swarm Optimization, and Differential Evolution.	Probabilistic Neural Network	96.97 ± 1.893%
	[69]	2018	Empirical Mode Decomposition (EMD) and the Wavelet Transform.	Heuristic Black Hole Algorithm	Multi-class SVM	92.56%
	[70]	2018	EMD and the Bat algorithm	Autonomous Bat Algorithm	Multi-class SVM	95%
	Proposed model			Hybrid of FCM and GA.	Optimized multi-class NN	98.06%

4.6. Fourth Experiment: Multi-Modal Emotion Recognition

The experiment is depicted in Figure 13. Figure 14 depicts comparisons of the multi-modal emotion recognition results obtained using the two decision level algorithms i.e., EWF and LWF, when changing the value of the NN weight. From the figure, LWF outperforms the EWF over all cases of changing the learned NN weight. This recommends that predicting or approximating a weighting among the modalities performs better than classifying the samples employing individual modality outputs just as features.
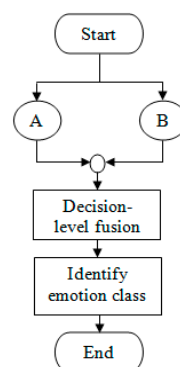

Figure 13. Flowchart of multi-modal emotion recognition using speech and EEG information.
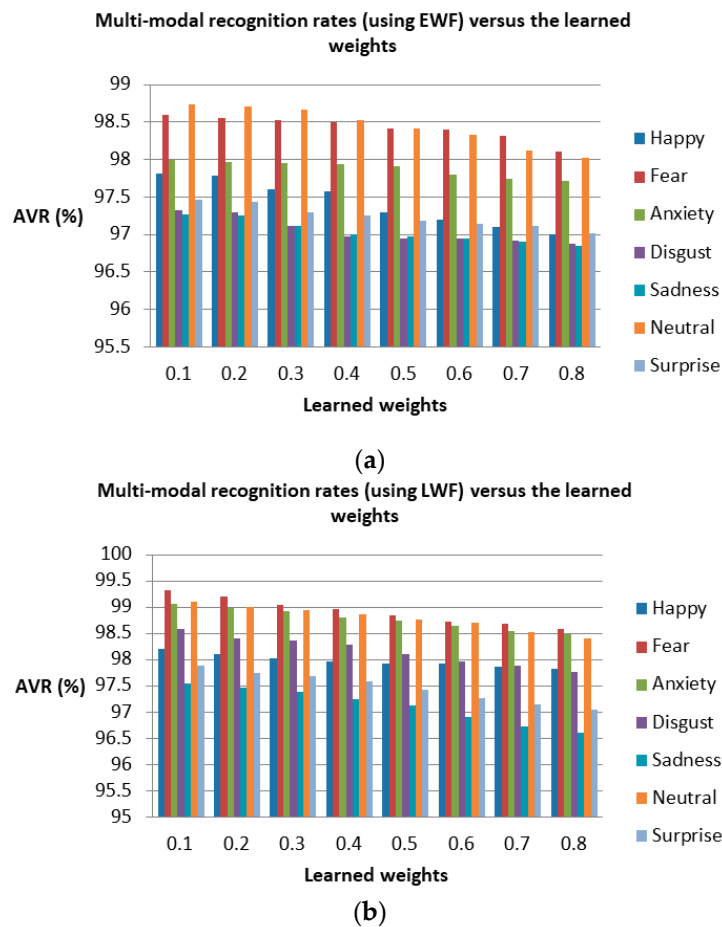
(**a**)

(**b**)

Figure 14. Multi-modal emotion recognition results using (**a**) EWF and (**b**) LWF versus the learned weights.

Although the results taken by speech modality are only lower than those obtained with EEG only, this recommends that it is imperative to fuse speech and EEG signals, because there are emotional states that are recognized in a superior way from EEG than from speech. Eventually, the overall performance of multi-modal emotion aware system is 98.53%, which is higher in comparison to performance taken by our superior unimodal system, which is based upon EEG. Thus, the decision-level fusion is superior than the sole measurement.

In conclusion, consideration of multiple modalities is useful during the time that some modality features are lost or unreliable. This might happen, for instance, when the process of feature detection is critical due to noisy environmental factors, when corrupting the signals through transmission, or, if the system is incapable of recording one of the modalities. Therefore, the emotion recognition system must be robust enough to manage these real-life naturalistic scenarios.

4.7. Computational Time Comparisons

The total run time (in seconds) is computed over our corpus for the three modalities, by testing the two algorithms, i.e., proposed FCM-GA-NN and $FCM-GA-NN_{fixed}$. In terms of computational run time, Figure 15 depicts the comparative performance analysis of the two classifiers, during the CV. In comparison to the proposed FCM-GA-NN model, we can find that multi-modal emotion recognition using $FCM-GA-NN_{fixed}$ model reaches the highest total computational time of 3150 s. On contrary, the total computational time achieved by the proposed FCM-GA-NN model is reduced by about 1512 s.
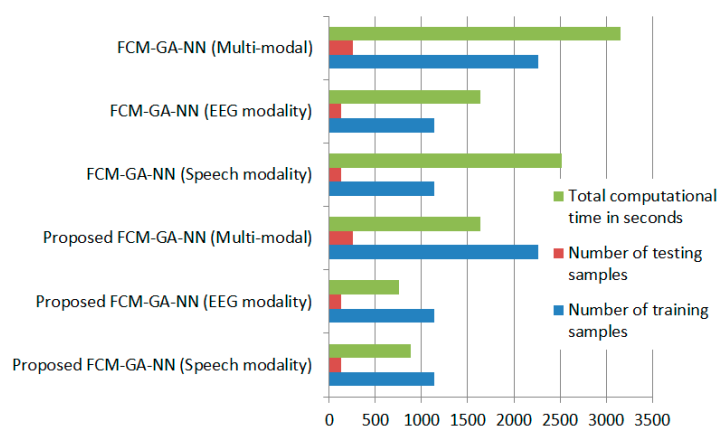
Figure 15. Computational time comparisons of the three modalities (speech, EEG, multi-modal), using the two algorithms i.e., the proposed FCM-GA-NN and $FCM-GA-NN_{fixed}$.

5. Conclusions and Future Work

This paper introduces a novel multi-modal emotion recognition framework by fusing speech and EEG information. The EEG signals were utilized as internal channel complementing of speech for more reliable emotions recognition. For classification of unimodal data (speech and EEG) and multi-modal data, a proposed hybrid FCM-GA-NN soft computing model was introduced. The fitness function of the algorithm checks the optimal number of fuzzy clusters that reduces the classification error. The proposed model was compared to another developed FCM-GA-NN model that relies on determination of a fixed number for fuzzy clusters, and the fitness function of the algorithm checks the optimal chromosome that reduces the classification error. Accordingly, the proposed model's errors were compared with the second model. The overall evaluation results of the two model, on emotion classes show the superiority of the proposed model. The algorithm was also implemented on two public databases for speech and EEG signals, namely SAVEE and MAHNOB, respectively. The overall performance of the proposed model reached 98.26%, 98.21%, and (98.06% and 97.28%) respectively for MAHNOB, SAVEE, and our dataset of its two modalities (EEG and speech, respectively). These results outperform the state-of-the-art results obtained using SAVEE and MAHNOB databases. Although the results taken by speech modality only (97.28%) are lower than those obtained with EEG only (98.06%), when applied to our dataset, this recommends fusing speech and EEG signals, because there are emotional states that are better recognized by the EEG than by speech. Therefore, after executing the automatic classification of every modality, the two modalities were fused on the decision-level. The comparative results of the fusion methods demonstrate that LWF performs better than EWF. By fusing speech and EEG modalities, the total computational time achieved by the proposed model also was reduced by about 1512 seconds than the $FCM-GA-NN_{fixed}$ model (with fixed number of centroids), when applied on our collected database. For the future work, we intend to recognize emotions through using a deep learning approach, which is optimized by the bio-inspired algorithms.

Author Contributions: Conceptualization, R.M.G. and A.D.A.; formal analysis, R.M.G. and A.D.A.; method and algorithms, R.M.G.; validation, R.M.G. and A.D.A.; visualization, R.M.G. and and A.D.A.; project management, R.M.G. and A.D.A.; writing—original draft: R.M.G.; writing—review and editing; R.M.G. and K.S.; supervision and paper revision, K.S.

Appendix A Algorithms

The pseudo-code of algorithms used in this paper is shown in Appendix A.

Algorithm A1: The proposed FCM-GA-NN model for emotion classification.

Inputs: (A) Parent population of each single modality M_p (speech and EEG), indicated as $P(M_p)$, $(p = 1, 2)$.

Output: Offspring population $P_{t+1}(M_p)$ where each offspring contains an optimal fuzzy cluster set that minimizes the classification error.

1.   **For** each single modality $M_p(p = 1, 2)$ **do**
2.   Generate fuzzy memberships of each training sample from M_p by one-step FCM clustering of Algorithm A2.
3.   Construct coded chromosome contain clusters centroids.
4.   Evaluate the fitness by Equations (11)–(17).
5.   Implement binary tournament selection on modality's population $P(M_p)$ to get the mating pool $P'(M_p)$.
6.   **For** each $p \rightarrow 1$ to p_{size}
7.   Decode two chromosomes from $P'(M_p)$ as matrices C_p and C_{p+1} using Equation (10).
8.   Compute fuzzy centroid sets F_p and F_{p+1} for C_p and C_{p+1}, respectively.
9.   **For** $k = 1$ **to** K // k indicates the index of kth centroid of the chromosome
10.   Apply crossover on $F_p(k)$ and $F_{p+1}(k)$ according to Algorithm A3 to get $F'_p(k)$ and $F'_{p+1}(k)$.
11.   Apply mutation on offsprings $F'_p(k)$ and $F'_{p+1}(k)$ according to Algorithm A4 to get $F''_p(k)$ and $F''_{p+1}(k)$ that represent updated centroid sets N_p and N_{p+1}.
12.   **End For**
13.   **End For**
14.   // Assigning object to a cluster
15.   **For** $j = 1$ to Z
16.   Compute distances among objects and every cluster centroid in N_j, then update fuzzy memberships of U_j by Equation (8) of Algorithm A2.
17.   Calculate the clustering indices by the maximum fuzzy memberships with regard to different clusters. **IF** empty cluster is found **THEN**
18.   Remove this empty cluster and then **Go To** Step 16.
19.   **ELSE**
20.   Assign fitness by Equations (11)–(17).
21.   **END**
22.   Get the resulting i-th offspring chromosome with its updated U_j and objectives.
23.   **End For**
24.   Apply Algorithm A5 to the resulting population $P_{t+1}(M_p)$ to compute the final solutions by clustering ensemble.
25.   Train the NN with solutions having the optimal cluster set that minimizes the classification error.
26.   **End For**

Algorithm A2: The clustering algorithm.

---

1. Determine the cluster number u.
2. Set the initial value of $\mu_{i,k}^{m}$ at time $m = 0$ satisfying:
3. $\sum_{i=1}^{u} \mu_{i,k}^{m} = 1, k = 1, 2, \ldots, K,$ (5)
4. $0 < \sum_{k=1}^{K} \mu_{i,k}^{m} < K, i = 1, 2, \ldots, u$, where: (6)

$$\begin{cases} \quad If \ \sum_{k=1}^{K} \mu_{i,k}^{m} = 0, empty \ cluster \\ If \ \sum_{k=1}^{K} \mu_{i,k}^{m} = K, feature vector \ belongs \ to \ i \end{cases}$$

5. Assume that K_a^m indicates the cluster index at time m and its initial value at $m = 0$ as 0.
6. Set the centroid of a cluster i at time m as R_i^m then compute the u cluster centroids for the partition as:

$$R_i^m = \frac{\sum_{k=1}^{K} \left(\mu_{i,k}^{m}\right)^a \cdot x_k}{\sum_{k=1}^{K} \left(\mu_{i,k}^{m}\right)^a}, i = 1, \ldots, u, 1 < a. \qquad (7)$$

7. Update membership degree for each emotion vector x_k:

$$\mu_{i,k}^{m+1} = \left[\sum_{h=1}^{u} \left(\|x_k - R_i^m\|^2 / \|x_k - R_h^m\|^2 \right)^{\frac{1}{a-1}} \right]^{-1} \qquad (8)$$

8. Calculate the clustering index
9. $K_a^{m+1} = \sum_{k=1}^{K} \sum_{i=1}^{u} \left[\left(\mu_{i,k}^{m+1}\right)^a \cdot \|x_k - R_i\|^2 \right],$ (9)
10. **IF** $\left| K_a^{m+1} - K_a^m \right| \geq \varepsilon$ **THEN**
11. $m = m + 1$
12. **Go to** Step 7
13. **END**

Algorithm A3: Crossover algorithm.

Inputs: Parents F_p and F_{p+1} assigned as R , and M , respectively, where $R = \{r_1, r_2, r_3, r_4\}$, and $M = \{m_1, m_2, m_3, m_4\}$
Output: Children D^1, and D^2

1.   **For** $i = 1$ **to** y
2.   **IF** $Random(0, 1) <= 0.5$ **THEN**
3. $C_i = r_i \& O_i = m_i$
4. $CROSS(C_i, O_i)$
5.   **Else**
6. $C_i = r_i \& O_i = m_i$
7.   **End IF**
8.   **End For**
9. $CROSS(C, O)$
10.   // random number between falls between 0 and 1
11. $g = Random(0, 1)$
12.   **IF** $g <= 0.5$ **THEN**
13. $beta = (2g)^{1/(eta_k+1)}$
14.   **Else**
15. $beta = (1/(2 * (1 - g)))^{1/(eta_k+1)}$
16.   **End IF**
17.   **For** $j = 1$ **to** x // x is the maximum dimension
18.   **IF** $C_j < O_j$ **THEN**
19. $Q_1 = C_j \& Q_2 = O_j$
20.   **Else**
21. $Q_1 = O_j \& Q_2 = C_j$
22.   **End IF**
23. $D_j^1 = \dfrac{[(1 + beta) * Q_1 + (1 - beta) * Q_2]}{2}$ (19)
24. $D_j^2 = \dfrac{[(1 - beta) * Q_1 + (1 + beta) * Q_2]}{2}$ (20)
25.   **End For**
26.   **IF** $Random(0, 1) <= 0.5$ **THEN**
27. $C = D^1 \& O = D^2$
28.   **Else**
29. $C = D^2 \& O = D^1$
30.   **End IF**
31.   **Return** D^1, and D^2

Algorithm A4: Mutation algorithm.

Inputs: Set of crossover-subjected chromosomes CR, $CR = \{cr_1, cr_2, cr_3, \ldots, cr_K\}$
Output: mutated chromosomes

1. // k is the index of *kth* centroid of the chromosome
2. **For** $k = 1$ **to** K
3.   **For** $j = 1$ **to** J // j is the index of chromosome point
4.   Generate a random number from 0 to : $B = random(0,1)$
5.   **IF** $B \leq P_T$
6.   Generate z_j random number $f_1, f_2, \ldots, f_{j,z}$

   from $[0,1]$ for the *jth* point of the centroid $E^k_{k,j}$

7.   Replace $E^k_{k,j}$ by $f_k / \sum^{zj}_{k=1} f_k$
8.   **End IF**
9.   **End For**
10. **End For**

Algorithm A5: Final solution computation by clustering ensemble.

Inputs: Solutions set $T = \{s_1, s_2, \ldots, s_n\}$ indicating fuzzy memberships.
Output: Final cluster label S^*.

1. **For** $j = 1$ **to** V // for each solution>
2.   Calculate the ranking vectors $R'(s_j)$.
3.   Calculate the aggregated ranks based upon the sum of each solution to get a vector:
   $\widetilde{R}_t = \left[R^{(1)}_j, .R^{(2)}_j, .., R^{(n)}_j \right]$.
   End For
4.   Sort T by the aggregated ranks (values of R) in ascending order and obtain a new solutions set:
   $T' = \{s'_1, s'_2, \ldots, s'_n\}$.
5.   Assume an ensemble of Z size.
6.   Select the first Z solutions to get a subset of non-dominated solutions: $T_{new} = \{s'_1, s'_2, \ldots, s'_Z\}$
7.   **For** $j = 1$ **to** Z // for each non-dominated solution s'_j in T_{new}
8.   Decode s'_j by Equation (10).
9.   Assign every object k of s'_j to a cluster i by the maximum of $\mu^m_{i,k}$, where $1 \leq k \leq K$.
10.   Get a vector x^j containing n cluster labels.
11.   Considering the size of T_{new} is Z , add the resulting vector x^j to a matrix E of size $Z \times n$.
12. **End**
13. Apply the clustering ensemble algorithm to the resulting matrix E, where each row represents one clustering solution.
14. Return a vector S^* comprising n cluster labels as the final output.

References

1. Kolodyazhniy, V.; Kreibig, S.D.; Gross, J.J.; Roth, W.T.; Wilhelm, F.H. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. *Psychophysiology* **2011**, *48*, 908–922. [CrossRef] [PubMed]
2. Liu, Y.-J.; Yu, M.; Zhao, G.; Song, J.; Ge, Y.; Shi, Y. Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals. *IEEE Trans. Affect. Comput.* **2018**, *9*, 550–562. [CrossRef]
3. Menezes, M.L.R.; Samara, A.; Galway, L.; Sant'Anna, A.; Verikas, A.; Alonso-Fernandez, F.; Wang, H.; Bond, R. Towards emotion recognition for virtual environments: an evaluation of EEG features on benchmark dataset. *Pers. Ubiquitous Comput.* **2017**, *21*, 1003–1013. [CrossRef]

4. Gharavian, D.; Bejani, M.; Sheikhan, M. Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks. *Multimed. Tools Appl.* **2016**, *76*, 2331–2352. [CrossRef]

5. Li, Y.; He, Q.; Zhao, Y.; Yao, H. Multi-modal Emotion Recognition Based on Speech and Image. *Adv. Multimed. Inf. Process. – PCM 2017 Lecture Notes Comput. Sci.* **2018**, 844–853.

6. Rahdari, F.; Rashedi, E.; Eftekhari, M. A Multimodal Emotion Recognition System Using Facial Landmark Analysis. *Iran. J. Sci. Tech. Trans. Electr. Eng.* **2018**, *43*, 171–189. [CrossRef]

7. Wan, P.; Wu, C.; Lin, Y.; Ma, X. Optimal Threshold Determination for Discriminating Driving Anger Intensity Based on EEG Wavelet Features and ROC Curve Analysis. *Information* **2016**, *7*, 52. [CrossRef]

8. Poh, N.; Bengio, S. How do correlation and variance of base-experts affect fusion in biometric authentication tasks? *IEEE Trans. Signal Process.* **2005**, *53*, 4384–4396. [CrossRef]

9. Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl.-Based Syst.* **2018**, *161*, 124–133. [CrossRef]

10. Adeel, A.; Gogate, M.; Hussain, A. Contextual Audio-Visual Switching For Speech Enhancement in Real-World Environments. *Inf. Fusion* **2019**.

11. Gogate, M.; Adeel, A.; Marxer, R.; Barker, J.; Hussain, A. DNN Driven Speaker Independent Audio-Visual Mask Estimation for Speech Separation. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018.

12. Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. *Knowl.-Based Syst.* **2019**, *167*, 26–37. [CrossRef]

13. Ahmadi, E.; Jasemi, M.; Monplaisir, L.; Nabavi, M.A.; Mahmoodi, A.; Jam, P.A. New efficient hybrid candlestick technical analysis model for stock market timing on the basis of the Support Vector Machine and Heuristic Algorithms of Imperialist Competition and Genetic. *Expert Syst. Appl.* **2018**, *94*, 21–31. [CrossRef]

14. Melin, P.; Miramontes, I.; Prado-Arechiga, G. A hybrid model based on modular neural networks and fuzzy systems for classification of blood pressure and hypertension risk diagnosis. *Expert Syst. Appl.* **2018**, *107*, 146–164. [CrossRef]

15. Engberg, I.; Hansen, A. Documentation of the Danish emotional speech database des 1996. Available online: http://kom.aau.dk/~{}tb/speech/Emotions/des (accessed on 4 July 2019).

16. Wang, K.; An, N.; Li, B.N. Speech emotion recognition using Fourier parameters. *IEEE Trans. Affect. Comput.* **2015**, *6*, 69–75. [CrossRef]

17. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; Andre, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]

18. Tahon, M.; Devillers, L. Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges. *IEEE/ACM Transact. Audio Speech Lang. Process.* **2016**, *24*, 16–28. [CrossRef]

19. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.

20. Liu, Z.-T.; Wu, M.; Cao, W.-H.; Mao, J.-W.; Xu, J.-P.; Tan, G.-Z. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing* **2018**, *273*, 271–280. [CrossRef]

21. Alonso, J.B.; Cabrera, J.; Medina, M.; Travieso, C.M. New approach in quantification of emotional intensity from the speech signal: Emotional temperature. *Expert Syst. Appl.* **2015**, *42*, 9554–9564. [CrossRef]

22. Cao, H.; Verma, R.; Nenkova, A. Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. *Comput. Speech Lang.* **2015**, *29*, 186–202. [CrossRef] [PubMed]

23. D Griol, D.; Molina, J.M.; Callejas, Z. Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances. *Neurocomputing* **2019**, *326-327*, 132–140. [CrossRef]

24. Yogesh, C.K.; Hariharan, M.; Ngadiran, R.; Adom, A.; Yaacob, S.; Polat, K. Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech. *Appl. Soft Comput.* **2017**, *56*, 217–232.

25. Shon, D.; Im, K.; Park, J.-H.; Lim, D.-S.; Jang, B.; Kim, J.-M. Emotional Stress State Detection Using Genetic Algorithm-Based Feature Selection on EEG Signals. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2461. [CrossRef] [PubMed]

26. Mert, A.; Akan, A. Emotion recognition based on time–frequency distribution of EEG signals using multivariate synchrosqueezing transform. *Digit. Signal Process.* **2018**, *81*, 106–115. [CrossRef]

27. Zoubi, O.A.; Awad, M.; Kasabov, N.K. Anytime multipurpose emotion recognition from EEG data using a Liquid State Machine based framework. *Artif. Intell. Med.* **2018**, *86*, 1–8. [CrossRef] [PubMed]

28. Zhang, Y.; Ji, X.; Zhang, S. An approach to EEG-based emotion recognition using combined feature extraction method. *Neurosc. Lett.* **2016**, *633*, 152–157. [CrossRef] [PubMed]

29. Bhatti, A.M.; Majid, M.; Anwar, S.M.; Khan, B. Human emotion recognition and analysis in response to audio music using brain signals. *Comput. Human Behav.* **2016**, *65*, 267–275. [CrossRef]

30. Ma, Y.; Hao, Y.; Chen, M.; Chen, J.; Lu, P.; Košir, A. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Inf. Fusion* **2019**, *46*, 184–192. [CrossRef]

31. Hossain, M.S.; Muhammad, G. Emotion recognition using deep learning approach from audio–visual emotional big data. *Inf. Fusion* **2019**, *49*, 69–78. [CrossRef]

32. Huang, Y.; Yang, J.; Liao, P.; Pan, J. Fusion of Facial Expressions and EEG for Multimodal Emotion Recognition. *Comput. Intell. Neurosci.* **2017**, *2017*, 1–8. [CrossRef]

33. Abhang, P.A.; Gawali1, B.W. Correlation of EEG Images and Speech Signals for Emotion Analysis. *Br. J. Appl. Sci. Tech.* **2015**, *10*, 1–13. [CrossRef]

34. MacQueen, J.B. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 21 June–18 July 1965 and 27 December 1965–7 January 1966; pp. 281–297.

35. Bezdek, J. Corrections for "FCM: The fuzzy c-means clustering algorithm". *Comput. Geosci.* **1985**, *11*, 660. [CrossRef]

36. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2001.

37. Ripon, K.; Tsang, C.-H.; Kwong, S. Multi-Objective Data Clustering using Variable-Length Real Jumping Genes Genetic Algorithm and Local Search Method. In Proceedings of the 2006 IEEE International Joint Conference on Neural Network, Vancouver, BC, Canada, 16–21 July 2006.

38. Karaboga, D.; Ozturk, C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Appl. Soft Comput.* **2011**, *11*, 652–657. [CrossRef]

39. Zabihi, F.; Nasiri, B. A Novel History-driven Artificial Bee Colony Algorithm for Data Clustering. *Appl. Soft Comput.* **2018**, *71*, 226–241. [CrossRef]

40. Islam, M.Z.; Estivill-Castro, V.; Rahman, M.A.; Bossomaier, T. Combining K-Means and a genetic algorithm through a novel arrangement of genetic operators for high quality clustering. *Expert Syst. Appl.* **2018**, *91*, 402–417. [CrossRef]

41. Song, R.; Zhang, X.; Zhou, C.; Liu, J.; He, J. Predicting TEC in China based on the neural networks optimized by genetic algorithm. *Adv. Space Res.* **2018**, *62*, 745–759. [CrossRef]

42. Krzywanski, J.; Fan, H.; Feng, Y.; Shaikh, A.R.; Fang, M.; Wang, Q. Genetic algorithms and neural networks in optimization of sorbent enhanced H2 production in FB and CFB gasifiers. *Energy Convers. Manag.* **2018**, *171*, 1651–1661. [CrossRef]

43. Vakili, M.; Khosrojerdi, S.; Aghajannezhad, P.; Yahyaei, M. A hybrid artificial neural network-genetic algorithm modeling approach for viscosity estimation of graphene nanoplatelets nanofluid using experimental data. *Int. Commun. Heat Mass Transf.* **2017**, *82*, 40–48.

44. Sun, W.; Xu, Y. Financial security evaluation of the electric power industry in China based on a back propagation neural network optimized by genetic algorithm. *Energy* **2016**, *101*, 366–379. [CrossRef]

45. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Trans. Affect. Comput.* **2012**, *3*, 42–55. [CrossRef]

46. Harley, J.M.; Bouchet, F.; Hussain, M.S.; Azevedo, R.; Calvo, R. A multi-componental analysis of emotions during complex learning with an intelligent multi-agent system. *Comput. Human Behav.* **2015**, *48*, 615–625. [CrossRef]

47. Ozdas, A.; Shiavi, R.; Silverman, S.; Silverman, M.; Wilkes, D. Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 1530–1540. [CrossRef] [PubMed]

48. Muthusamy, H.; Polat, K.; Yaacob, S. Particle Swarm Optimization Based Feature Enhancement and Feature Selection for Improved Emotion Recognition in Speech and Glottal Signals. *PLoS ONE* **2015**, *10*, e0120344. [CrossRef] [PubMed]

49. Jebelli, H.; Hwang, S.; Lee, S. EEG Signal-Processing Framework to Obtain High-Quality Brain Waves from an Off-the-Shelf Wearable EEG Device. *J. Comput. Civil Eng.* **2018**, *32*, 04017070. [CrossRef]

50. Ferree, T.C.; Luu, P.; Russell, G.S.; Tucker, D.M. Scalp electrode impedance, infection risk, and EEG data quality. *Clin. Neurophys.* **2001**, *112*, 536–544. [CrossRef]

51. Delorme, A.; Makeig, S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **2004**, *134*, 9–21. [CrossRef] [PubMed]

52. Ayadi, M.E.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.* **2011**, *44*, 572–587. [CrossRef]

53. Candra, H.; Yuwono, M.; Chai, R.; Handojoseno, A.; Elamvazuthi, I.; Nguyen, H.T.; Su, S. Investigation of window size in classification of EEG-emotion signal with wavelet entropy and support vector machine. In Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milano, Italy, 25–29 August 2015.

54. Ou, G.; Murphey, Y.L. Multi-class pattern classification using neural networks. *Pattern Recogn.* **2007**, *40*, 4–18. [CrossRef]

55. Yang, J.; Yang, X.; Zhang, J. A Parallel Multi-Class Classification Support Vector Machine Based on Sequential Minimal Optimization. In Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS06), Hangzhou, China, 20–24 June 2006.

56. Ghoniem, R.M.; Shaalan, K. FCSR - Fuzzy Continuous Speech Recognition Approach for Identifying Laryngeal Pathologies Using New Weighted Spectrum Features. In Proceedings of the 2017 International Conference on Advanced Intelligent Systems and Informatics (AISI), Cairo, Egypt, 9–11 September 2017; pp. 384–395.

57. Tan, J.H. On Cluster Validity for Fuzzy Clustering. Master Thesis, Applied Mathematics Department, Chung Yuan Christian University, Taoyuan, Taiwan, 2000.

58. Wikaisuksakul, S. A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering. *Appl. Soft Comput.* **2014**, *24*, 679–691. [CrossRef]

59. Strehl, A.; Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.

60. Koelstra, S.; Patras, I. Fusion of facial expressions and EEG for implicit affective tagging. *Image Vision Comput.* **2013**, *31*, 164–174. [CrossRef]

61. Dora, L.; Agrawal, S.; Panda, R.; Abraham, A. Nested cross-validation based adaptive sparse representation algorithm and its application to pathological brain classification. *Expert Syst. Appl.* **2018**, *114*, 313–321. [CrossRef]

62. Oppedal, K.; Eftestøl, T.; Engan, K.; Beyer, M.K.; Aarsland, D. Classifying Dementia Using Local Binary Patterns from Different Regions in Magnetic Resonance Images. *Int. J. Biomed. Imaging* **2015**, *2015*, 1–14. [CrossRef] [PubMed]

63. Gao, X.; Lee, G.M. Moment-based rental prediction for bicycle-sharing transportation systems using a hybrid genetic algorithm and machine learning. *Comput. Ind. Eng.* **2019**, *128*, 60–69. [CrossRef]

64. Liu, Z.-T.; Xie, Q.; Wu, M.; Cao, W.-H.; Mei, Y.; Mao, J.-W. Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing* **2018**, *309*, 145–156. [CrossRef]

65. Özseven, T. A novel feature selection method for speech emotion recognition. *Appl. Acoust.* **2019**, *146*, 320–326. [CrossRef]

66. Ghoniem, R.M. Deep Genetic Algorithm-Based Voice Pathology Diagnostic System. In Proceedings of the Natural Language Processing and Information Systems Lecture Notes in Computer Science, Salford, UK, 26–28 June 2019; pp. 220–233.

67. Ghoniem, R.M.; Shaalan, K. A Novel Arabic Text-independent Speaker Verification System based on Fuzzy Hidden Markov Model. *Procedia Comput. Sci.* **2017**, *117*, 274–286. [CrossRef]

68. Nakisa, B.; Rastgoo, M.N.; Tjondronegoro, D.; Chandran, V. Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors. *Expert Syst. Appl.* **2018**, *93*, 143–155. [CrossRef]

69. Munoz, R.; Olivares, R.; Taramasco, C.; Villarroel, R.; Soto, R.; Barcelos, T.S.; Merino, E.; Alonso-Sánchez, M.F. Using Black Hole Algorithm to Improve EEG-Based Emotion Recognition. *Comput. Intell. Neurosci.* **2018**, *2018*, 1–21. [CrossRef]

70. Munoz, R.; Olivares, R.; Taramasco, C.; Villarroel, R.; Soto, R.; Alonso-Sánchez, M.F.; Merino, E.; Albuquerque, V.H.C.D. A new EEG software that supports emotion recognition by using an autonomous approach. *Neural Comput. Appl.* **2018**. [CrossRef]