Kshirod Sarmah[1*],

Swapnanil Gogoi[2] ,

Hem Chandra Das[3],

Bikram Patir[4],

Mrinal Jyoti Sarma[5]

A State-of-arts Review of Deep Learning Techniques for Speech Emotion Recognition

Journal of Electrical Systems

Abstract: In sophisticated Human-Computer Interfaces (HCI), the emotional state of the user is becoming a crucial component that is closely linked to emotional speech recognition. Spoken expressions, which can be a part of human-machine interaction, are an important source of emotional information. Speech emotion recognition (SER) in deep learning (DL) continues to be a hot topic, especially in the field of emotional computing. Current deep learning (DL) and neural network methods are applied in this highly active field of research. This is as a result of its expanding potential, advancements in algorithms, and practical uses. Quantitative factors such as pitch, intensity, accent and Mel-Frequency Cepstral Coefficients (MFCC) can be employed to model the paralinguistic data contained in human speech. To achieve SER, three key procedures are usually followed: data processing, feature selection/extraction, and classification based on the underlying emotional qualities. The nature of these processes and the peculiarities of human speech lend support to the employment of DL techniques for SER implementation. A variety of DL methods have been used for SER tasks in recent affective computing research works; however, only a small number of them capture the underlying ideas and methodologies that can be used to facilitate the three main steps of SER implementation. With a focus on the three SER implementation processes, we provide a state-of-the-art assessment of research conducted over the last ten years that tackled SER tasks from DL perspectives in this work. Various issues are covered in detail, including the problem of low classification accuracy of Speaker-Independent experiments and the related remedies. The review offers principles for SER evaluation as well, emphasizing indicators that can be experimented with and common baselines.

Keywords: Speech Emotion Recognition, Machine Learning, Deep Learning, MFCC, Speaker-independent experiment, Classification.

1. Introduction

Speech Emotion Recognition (SER) is a subfield of Automatic Speech Recognition (ASR) [1][2][3]. Apart from employing same signal types, feature extraction procedures, and the possible utilization of distinct machine learning methodologies, such deep learning (DL) architectures, which are also employed in the Natural Language Processing (NLP) sector, SER and ASR share the sequential nature of the data [4].Over the course of more than 20 years, SER, a subfield of emotional computing [5], has produced a sizable body of published papers [6][7]. SER entails identifying the affective dimensions of speech regardless of its semantic content to assess the work of call center personnel [8]. According to Mekruksavanich this data can help businesses increase customer satisfaction and call center efficiency by enhancing service quality or offering focused training [10]. In industries like healthcare, smart homes, and smart entertainment, SER has emerged as a crucial component of many smart service systems [11]. Speech emotion analysis is a useful tool for emergency call centers to detect potentially dangerous or life-threatening situations [12]. An automobile's interactive voice response system could also employ SER to reduce accidents caused by tired drivers [13]. In clinical contexts, SER may improve mental health diagnosis [15] or foster telemental health [14], such as by identifying probable suicide ideation indicators [16].

In order to identify emotions inherent in speech signals, SER systems are often composed of various techniques for extracting, classifying, and isolating speech signals [6].Numerous real-world applications of SER exist, and some of them have shown how adding emotional characteristics to human-machine interactions can greatly enhance users' interaction experiences[9]. For instance, a SER system can identify client emotions like happiness or rage in order.

SER is a useful tool for online education services because it lets Professors use the emotional content of students' responses to gauge how well they have mastered new abilities. This can be applied to maximize learning outcomes and adjust the lesson plan [13]. Finding and extracting speech data that is most suited for computational emotion recognition and discrimination is one of SER's trickier jobs. Although there is a wealth of information in human speech, including linguistic and paralinguistic elements, this research will concentrate on the paralinguistic features.While linguistic aspects are associated with the qualitative trends in speech patterns of humans such as context and content, as well as paralinguistic elements, Anagnostopoulos and Zhao measures the differences in language pattern pronunciation [17][18]. These comprise

[1]Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (A Govt. Model College), Goalpara, 783124, Assam, India. kshirodsarmah@gmail.com
[2]GUCDOE, Gauhati University, Guwahati-781014, Assam, India. swapnanil@gauhati.ac.in
[3]Department of Computer Science and Technology, Bodoland University, Kokrajhar, 783370, Assam, India hemchandradas78@gmail.com
[4]Department of Computer Science, PDUAM, Dalgaon,784116 Assam, India. bikrampatir15@gmail.com
[5]Department of Computer Science and Engineering, Rajiv Gandhi University, Rono Hills, Doimukh, 791112, Arunachal Pradesh, India. mrinaljyoti.sarma@rgu.ac.in
Corresponding Author: Kshirod Sarmah
*Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (A Govt. Model College), Goalpara, 783124, Assam, INDIA, kshirodsarmah@gmail.com

spectral qualities as well as prosodic traits like pitch and intensity like Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictor Coefficients (LPC). Furthermore, more directly visual representations of the speech signals, like time-frequency spectrograms, are also possible [19]. The relationship between prosodic/spectral acoustic characteristics in speech and human emotions has been the subject of several SER studies.

The use of DL techniques for emotion recognition has grown significantly in recent years as a result of developments in digital signal processing, enhancements in human-computer interfaces [20], and quick advancements in ML[21]. It is necessary to concentrate on DL techniques at present due to the growing body of research on the topic, as well as the interest in and emphasis on emotion recognition in DL [22]. The ML pipeline techniques used in these studies, which include speech feature extraction, dimensionality reduction, emotion categorization based on underlying speech features, and speech signal isolation, were mostly used to complete SER tasks. The major goal is to employ machine learning (ML) to enhance user interaction with technology by better understanding users and communicating with them[23]. Other aspects of human speech and spectrum analysis along with the distinctive qualities of speech, facilitate the application of machine learning algorithms for voice recognition tasks. Traditional ML techniques involve identifying trends and deriving feature parameters from unprocessed data like audio, photos, videos, ECG, and graphics. A model that learns how to produce the required output label in a prediction or classification challenge is trained using these features.One can test many features, combine multiple features into a single feature vector, or use other feature selection approaches to determine which characteristics efficiently cluster data into classes.

Furthermore, more sophisticated approaches to avoiding the problem of an ideal feature selection are offered by more modern machine learning techniques, such as graphs and deep neural networks [24]. In order to train the ML model for SER, the audio data may be retrieved and used to represent various emotions in speech through the use of spectrograms and speech characteristics. Previous machine learning (ML)-based speech recognition (SER) research has examined acoustic speech characteristics and found associations between them and a speaker's emotions. Support Vector Machine (SVM) was used in the majority of the investigations [25][26][27],K-Nearest Neighbour (KNN) [29] and the Gaussian Mixture Model (GMM)[28] Neural networks (NN) and recurrent neural networks (RNN) [30][31]. Still, the advent of DL approaches has resulted in significant progress in this field. Improved emotion detection accuracy has resulted from the remarkable performance of CNNs and RNNs in capturing spatial and temporal relationships in emotional data [22].The three crucial steps are typically followed by the machine learning algorithms employed in these investigations: pre-processing [32], speech signal isolation, feature extraction and selection, and emotion classification from audio signals. There are some intrinsic difficulties in deducing the emotional states of speakers from their speech [33].First, it's unclear which aspects of speech are most useful for differentiating between different emotional states.

Additional levels of difficulty are created by the acoustic diversity brought about by the presence of various phrases, speakers, speaking rates, and speaking styles, all of which may have an immediate effect on the speech qualities that are recovered [33][34]. The SER performance may also be impacted by the speaker's reliance on specific emotional expressions as well as their environment, dialect, and culture. Secondly, it might be challenging to distinguish the borders between each unique emotional state due to emotion overlaps or the perception of numerous emotions in a single phrase. Despite the fact that several SER research projects have investigated a range of DL techniques using different combinations of speech characteristic.Furthermore, there is either little or no discussion of the difficulties posed by these techniques, such as the widespread poor accuracy of classification in Speaker-Independent SER systems and possible solutions. We have out a thorough analysis of DL-based SER systems to help with comprehending the incredibly varied applications of DL algorithms and their accessible approaches and strategies. Furthermore, we examine current approaches that tackle the problems of speaker reliance and The Speaker-Independent SER system's poor classification accuracy.This article is organized as follows: Section 2 reviews earlier studies on SER. The most recent SER approaches are provided in Section 3, the classic SER methodology is presented in Section 4, the DL techniques are shown in Section 5, and several significant SER databases are explained in Section 6 along with experimental findings of previous SER system, Future research scope and challenges are discussed in the Section 7 and finally Section 8 gives conclusions.

2. Review on Previous Research on Speech Emotion Recognition

The quickest and most natural form of human communication is speech [35]. The speech conveys both the speaker's emotional moods and the formal aspects of language expressions, such as phonology, morphology, syntax, and semantics. Using linguistic and paralinguistic clues, speech processing techniques can be utilized to extract affective information related to emotional expressions from it. [35][36]. Additionally, speech signal analysis can shed light on the language used, the speaker, the lexical material spoken, and the emotional component of the speech [37].

This affective factor is critical to comprehending human decision-making and can provide insight into a person's mental health [38][39][40]. Speech signals are a useful tool for human-machine communication because of these facts. While there are other modalities that can be investigated to identify human emotional states, such as facial expression[41][42], text[43], and Speech signal has certain intrinsic properties that compared to physiological data such skin conductivity, respiration, heart rate, and brain membrane, make it a more important source for emotional computing [42].For example, voice signals

are more easily and affordably obtained than physiological/biological signals. Additionally, Not every emotion is conveyed through facial expressions and syntactic and semantic difficulties make it harder to discern emotions from the text [44]. Physical exercise can cause confusion when utilizing heart rate [45]. Several machine learning techniques have been utilized to examine the highly valuable acoustic cues that the paralinguistic content of speech provides for encoding the speaker's emotional state [35][46]. All languages share these qualities, hence a general classification model can be used to all of them [47].

Furthermore, speech-based emotion recognition might help circumvent the privacy issues related to facial expression detection, which may be more acceptable in the eyes of patients utilizing digital health solutions [48]. Specific speech signal elements, like the emotional component, must be recognized while excluding other speech aspects, like linguistic and cultural information, in order to create a generalizable speech recognition system. More specifically, speaker/language independence—the ability of the SER system to support various speakers speaking various languages—must be supported. The speaker's difficulties and language dependence in SER implementations was thus the subject of some recent SER research [49][50][51][47][52]. Numerous recent SER research findings used machine learning (ML) approaches and techniques to address various implementation-related issues. Numerous non-linear classifiers, like the Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), are available for SER [53]. These are extensively employed in the classification of data originating from fundamental traits. Effective methods for recognizing emotions from speech include energy-based characteristics like Mel Energy-spectrum Dynamic Coefficients (MEDC), Mel-Frequency Cepstrum Coefficients (MFCC), Perceptual Linear Prediction Cepstrum Coefficients (PLP), and Linear Predictor Coefficients (LPC). For the purpose of recognizing emotions, additional classifiers such as Principal Component Analysis (PCA), K-Nearest Neighbor (KNN), and Decision trees are also used [54].

In recent years, deep learning has drawn more interest and is thought to be an emerging subject of machine learning research [55]. SER using DL algorithms has various benefits over traditional ML techniques. These include low-level feature extraction from the provided raw data, handling unlabeled input, and the capacity to identify complex structures and features without the need for manual feature extraction and tuning. Feed-forward topologies with one or more underlying hidden layers positioned between inputs and outputs form the foundation of DNNs. Feed-forward designs such as CNNs and DNNs yield good results for processing pictures and videos.On the other hand, recurrent architectures like RNNs and LSTM yield good results for voice-based classification applications like NLP and ASR [55]. These models do have certain limitations in addition to their usefulness in classification. For example, CNNs, have the advantage of being capable of extracting features from incoming data with high dimensions but, they also demand huge store capacity because they learn features from minor changes and distortion occurrences. Similarly, long-range sequential text data can be modeled by LSTM-based RNNs, which can also handle changeable input data.

Many research articles concentrate on a specific approach or method for accomplishing the ML objectives, even though nearly all of these studies explain the SER process in three steps: data pre-processing, feature extraction, and classification of emotions from audio signals. Similarly, only the $MFCC$ of speech characteristics were used to classify emotions using a CNN single-classifier architecture [56]. Adding more audio features or using ensembled model topologies like CNN and SVM-based models was not mentioned. As feature extractors, GMM and NN were used [57], however there was not much information available regarding feature selection and extraction techniques [58].To make finished the SER challenge using speech spectrograms and an ensembled model that based on CNN and RNN. Low-level descriptors ($LLDs$) were not taken into consideration here. The performance of multiple single-models such as attention model with CNN, $RNN - LSTM$, and SVM in the SER task was compared with some of the spectrograms and LLs. The data pretreatment methods, however, were not covered [59]. A highly particular ML method or phase in the ML process was the exclusive focus of a few additional studies. For instance, SER relied primarily on deep learning techniques [60][61], while focused heavily on emotion features and classifiers [34].

DBN has a considerably more complex structure and is built from cascading RBM structures [62]. DBN is an extension of RBMs, whereby layer by layer, RBMs are trained from the bottom up. DBNs are commonly used for SER because of their ability to learn recognition parameters rapidly, even with a large number of parameters. Furthermore, it avoids layer non-linearity [63]. DBNs use back propagation methods during training to address localized slow speed problems.
Moreover, a few of the research were limited to tackling a specific SER challenge [58]. The problems of data imbalance and insufficiency were resolved by a specific data augmentation technique [64] use a Histogram Equalization technique to mitigate the detrimental effects of speaker difference in audio data. Additionally, Kerkeni use a zero-crossing rate detection approach to lessen the negative influence of audio signal trends [42]. However, only a few number of studies offer a thorough analysis of the issues and solutions that exist in the SER task today. We acknowledge the existence of a few SER surveys conducted in the past ten years, including in the studies [6][30][34][35][47].

The majority of these publications encompass the essential elements of SER, such as database management, processing of data, feature extraction, selection, and classification. For instance, Fahad covered deep-learning as well as conventional ML methods for SER [37]. The writers discussed the benefits and drawbacks of each of the three types of databases—acted, evoked, and natural databases. There was a description of the difficulties and methods used to address problems with natural contexts, such as speaker, language, and textual dependencies. Speech emotion feature classes—acoustic and non-acoustic—were also covered. They go over several methods that can be used to extract features related to emotions. The report included common evaluation measures; nevertheless, it lacked details regarding evaluation methodologies, baselines, and data pre-processing. An end-to-end survey, from databases to evaluation methodologies, is presented by Schuller [65]. The authors treat SER tasks as problems involving both regression and classification. They go into great detail on SER elements such features and feature selection techniques. Their survey provides examples of several ways for evaluating SER models. Although their study does not delve into specific methodology, it does reflect the problems related to the resilience of SER in cross-corpus and noisy environments.

Databases, features, emotional models, preprocessing, supporting modalities, and emotion categorization were among the essential elements of SER that had been investigated by Akçay and Oguz [6]. Additionally, the trade-offs between two methods of modeling emotions were discussed [6]. Discrete models, based on the six categories of discrete emotions—happiness, sadness, fear, fury, surprise and disgust—are the foundation of the first method. These are fleeting emotions that are generally observed in people going about their daily business. However, distinct emotions are unable to fully represent a few of the most complex emotional states seen in interpersonal interactions.Comparatively speaking, dimensional models utilize a limited set of latent dimensions, terms like power, control, arousal, and valence to characterize human emotions. These feelings are systematic equivalents of each other. The writers also go over a number of newly developed deep learning classifiers that are ML-based, like auto-encoders and 3D-CNN. While this survey provides a range of preprocessing methods for speech data, it does not address feature selection, language-independent and speaker adaptation, or assessment procedures. Our review unifies and builds upon earlier studies [65][66], thereby filling in the knowledge gap about the state-of-the-art in SER.

3. State-of-art techniques in SER

Multimodal techniques have replaced unimodal ones in recent research, with a major emphasis on creating audiovisual models to guarantee greater accuracy. An end-to-end network that employs LSTM in addition to a CNN is proposed by F.Zhang[68]. G.Tang leverage an attention mechanism to boost the DL network's efficacy[69]. Although E. Ghaleb introduces a multimodal emotion recognition metric learning [70]. A correlation-based graph convolutional network (C-GCN) is described by W. Nie with the aim of audiovisual emotion recognition [71]. Z. Farhoudi presents an audio-visual fusion model of deep learning features that combines brain and emotional learning [72]. Both a CNN and a recurrent neural network (RNN) were employed in this technique.Nandi introduced an alternative federated learning multimodal emotion recognition model[73]. They presented a real-time emotional state classification technique using multimodal streaming and federated learning. They were mostly interested in using physiological data that was obtained via wearable sensors. With the exception of multimodal approaches from [73][74], which focused on certain objectives, there are very few federated learning-based methods for unimodal emotion detection systems that take into account the visual modality [75]or the auditory modality [76][77].We do a study centered on the multimodal audiovisual emotion identification challenge using federated learning, using relatively modest classification models that may be deployed at the edge. This is a positive development in the area of emotion identification while protecting privacy.

Novel approaches in SER are being investigated in recent works to enhance its computational complexity, dependability, and performance [78]. While some studies prioritize model performance over privacy protections, others investigate how Federated learning (FL) in SER can safeguard speech data from adversarial assaults [79]. Tsouvalas describes FL as a distributed machine learning paradigm that decentralizes the training of models using sensitive personal data [80]. FL is a useful tactic to avoid privacy infringement since it enables several users to collaboratively learn a common model without sharing their local data. The speech data for SER is stored on user devices in FL. Only locally constructed model parameters are sent to a central server, which aggregates updates from all users involved in order to train the model as a whole. As a result, only the original and updated model parameters are sent between the end user and the central server in terms of speech data [79]. Using Mel Filterbank characteristics and varying user interaction percentages to train CNN and RNN-based classifiers to identify four emotion classes, Latif examined the viability of FL for SER [79]. Nonetheless, the study draws attention to the resource and communication overhead in FL for SER at the client ends.

Furthermore, it is difficult to obtain enough label data for training at the user's end in SER applications, as conventional FL approaches expect [80]. To make use of both marked and unlabelled data for user device training, Tsouvalas investigated semi-supervised learning under FL conditions [80]. For unlabeled data, they compute pseudo-labels; only extremely

confident guesses are taken into account during the training phase, where the estimations are treated as target classes that are similar to the actual truth. As a SER classifier, an attention-based CNN architecture under FL is investigated [80]. A semi-supervised federated learning system is also offered by Amiriparian, which allows users to work with both labeled and unlabeled data by creating pseudo-labels for unlabeled data using a range of complementary perspectives [81]. An MLP-based SER model is used to conduct the experiments by Feng and Narayanan [82].

4. Traditional Techniques for SER

Three essential elements make up digital speech-based emotion identification systems: feature extraction, classification, and signal preprocessing [83]. Segmentation and other forms of acoustic preprocessing like denoising are used to identify the signal's meaningful units [84]. Finding the pertinent characteristics in the speech signal is done via feature extraction. Finally, the appropriate emotions are assigned to the gathered feature vectors. A thorough discussion and explanation of feature extraction, classification, and speech signal processing is suggested by A. Batliner [85]. Since they are pertinent to the subject, the distinctions between performed and spontaneous speech are also covered [86][87].
A simple speech-based system for recognizing emotions is shown in Figure 1. Speech-based signal processing starts with an initial stage that improves speech quality and eliminates noisy components. Feature extraction and feature selection are the two major components of the second stage. The selection method begins with the extraction of the preprocessed speech signal's key characteristics. The analysis of speech data in the temporal and frequency domains is typically the first step in this kind of feature extraction and selection. In the third stage, these features are categorized using a variety of classifiers, including GMM, HMM, and others. Lastly, to differentiate between various moods, feature categorization is employed.

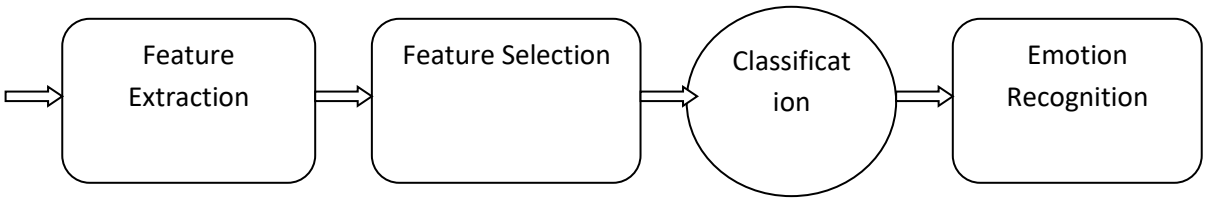

Figure 1. Basic SER System

5. SER Deep Learning Methodologies

On an audio stream, speech processing usually operates in a comprehensible manner [88]. For many speech-based applications, including as voice denoising techniques namely VAD, music classification, and SER, it is deemed necessary. Recent developments have led to a major growth in the significance of SER. However, in order to converse with people, effective methods of mimicking human behavior are still required [89]. As mentioned earlier, a SER system consists of several parts: language-based modeling, acoustic modeling, feature extraction and selection, feature categorization, and per-unit recognition. Conventional SER systems typically use several categorization models, including HMMs and GMMs. Many nonlinear components that execute computation in parallel make up deep learning algorithms [90].

To overcome the shortcomings of other strategies, these solutions, however, require deeper architectural levels. SER greatly enhances the overall performance of the system by utilizing fundamental DL techniques, such as Auto Encoder (AE), Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Networks (CNN), and Recurrent Neural Network (RNN).

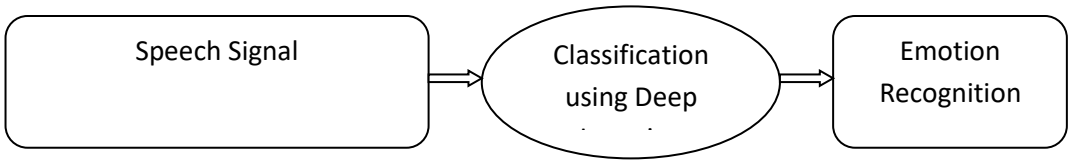

Figure 2. Deep Learning-Based SER System

In recent years, there has been a noticeable increase in interest in the machine learning study subject of deep learning. A few researchers trained their individual SER models using DNNs. The distinctions between the deep learning flow mechanisms for SER and the conventional machine learning flow are shown in Figure 2. Table 1 provides a detailed comparison between the Deep Convolutional Neural Network (DCNN) algorithm and traditional algorithms in the context of measuring and

recognizing different emotions, such as happiness, anger, and sorrow, using the IEMOCAP, Emo-DB, and SAVEE datasets [91]. It is found that deep learning algorithms outperform conventional methods in the identification of emotions.

In the part that follows, the study plans to go over lots of DL techniques in relation to SER. Despite being computationally intensive, these techniques produce results that are more accurate than those of traditional methods. Deep learning comes from the ML area, which is a more generic learning technique for describing emotions in data [92][93]. Deep learning can be done in three different ways: unsupervised, semi-supervised, or completely supervised. DL technique is a rapidly expanding field of study today because to its multi-layered structure and effective result delivery. Among these research areas are speech and picture identification, natural language processing, pattern recognition, and voice emotion detection [94][95]. Numerous deep learning algorithms, including CNNs, RNNs, RvNNs, AEs, DBMs, and DBNs, are covered in this section.DBMs are mainly developed from Markov Random fields [96], [97] and comprise several hidden layers. These layers mix variables chosen at random with stochastic entities. The main advantages of DBM are its quick learning curve and good representation of data and it uses layer-by-layer pre-training to accomplish it better way that explained by G.E.Hilton [98]. This explains why, when speech is used as input, DBM can produce results for emotion recognition that are more accurate. Furthermore, According to A.R. Mohamed DBMshave significant drawbacks as well, like limited efficacy under some particular conditions [99].

An RNN is a very popular neural network where inputs and outputs are interdependent and are based on sequential information [100]. This interdependency typically aids in forecasting the input's future state. Like CNNs, RNNs require memory to store all of the data gathered throughout the sequential deep learning modeling process. Frequently, they are only useful for a small number of back-propagation stages. Speech and other sequential data are a good fit for RNNs, especially the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants. The ability of these architectures to record speech feature temporal relationships throughout time is critical for identifying emotional emotions. It's common practice to employ stacked or bidirectional RNNs to record both forward and backward temporal contexts.RNN architectures now incorporate attention methods to concentrate on pertinent speech signal parts. The Transformer design and other self-attention techniques enable the model to adaptively weigh the various input sequence components according to their significance for emotion recognition.Tree-structured input sequences are not necessary for the hierarchical deep learning method known as RvNN [101]. Instead, the method is independent from the input sequences. It may find the parse tree of the supplied data more quickly if the input is divided into digestible chunks. While natural language processing is RvNN's main use case, speech recognition and other modalities such as SER can also be handled by its architecture.

The structure of DBN is much more complex and is built from cascading RBM structures [102]. DBN extends RBMs by training them from the bottom up, layer by layer. Due to its capacity to learn recognition parameters quickly, even with a high number of parameters, DBNs are frequently employed for speech emotion recognition. Moreover, the layers' non-linearity is avoided [103][104]. It uses back propagation methods during training to address localized slow speed problems. The first primary advantage of DBNs is their ability to perform pre-training procedures with large, unlabeled databases in an unsupervised manner [103][104]. Another benefit of DBNs is their ability to approximate the inference method in order to obtain the necessary output weight of the variables. Nevertheless, the bottom-up pass inference method of DBNs has certain limitations: the network's temporal states are stored in memory blocks between the recurrent connections, and each memory block has gated units to control the inflow of new input. The remaining connections, being typically rather deep, can help reduce the gradient problem [104][105].

CNN is an alternative DL approach that solely use feed-forward architecture for classification [106]. CNNs are widely utilized for pattern recognition and provide better data classification. These networks use small neurons found on each layer of the model design to process incoming data as receptive fields [107]. Because CNNs can detect local patterns in speech spectrogram representations, they are frequently utilized in SER. In order to extract hierarchical features, typical designs consist of many convolutional layers followed by max-pooling layers. To increase performance, CNNs can be enhanced with extra elements including batch normalization, dropout, and residual connections.Hybrid architectures combining CNNs and RNNs have demonstrated potential in collecting both global and local temporal characteristics. An RNN can be used for sequential modeling and emotion categorization, once a CNN has extracted high-level spectral information from raw audio.

Transformer designs have recently been adopted for SER, although they were originally created for challenges related to natural language processing. Transformers are useful for modeling emotional content over long sequences because they use self-attention mechanisms to capture long-range interdependence in speech features.Capsule Networks offer an alternative to traditional CNN architectures by representing entities as capsules with orientation and pose. These architectures aim to capture hierarchical relationships between speech features, potentially improving robustness to variations in emotion expression.Recently, Graph Neural Networks (GNNs) have attracted interest in SER, especially in modeling the links between conversational speech's linguistic context and acoustic characteristics. GNNs facilitate the more efficient

integration of contextual information for emotion recognition by capturing complex dependencies in graph-structured data. GNNs enhance the comprehension of emotional expressions expressed through speech in SER by capturing semantic linkages and contextual information from the linguistic context.

6. Important Databases for SER

A comprehensive list of all available datasets for SER is provided by several generic studies on emotion recognition. Based on their experience with related projects, Douglas-Cowie provide some recommendations on how to gather relevant datasets for SER [108]. During the dataset generation process, several factors must be taken into account, like the quality control, speaker selection process, recording setup, and recording prompts [109]. The most frequently cited datasets, which made use of the majority of the SER experiments, are succinctly described in this section.

Danish Emotional Speech (**DES)**[110][111] is a European Union-funded Danish database of emotions. It has four professional actor recordings—two male and two female—and is approximately thirty minutes long. The recordings consist of two distinct words, "yes" and "no," nine short phrases, and two paragraphs. The following five emotional states were simulated by each uttered utterance: neutral, surprise, happiness, sadness, and rage. Among the databases that scholars in this discipline have used the most is this one. Its uses include testing SVM techniques by themselves [112] and in conjunction with HMM [113]. It has also been used to innovative deep learning techniques [115] and gender-based emotion recognition [114].

EMODB [116] A German database that German actors recorded in high-fidelity audio for the German database EMODB execute ten German expressions representing the seven emotions namely happiness, sadness, anger, anger, boredom, , neutral, and disgust are presented in five short and five longer statements.The same author has recorded two versions of several emotive expressions. As a result, the database offers roughly 800 sentences. Typical sentences were employed to give the form a more organic feel. It's one of the most commonly used terms in technical writing. Along with bioinspired real-time speech emotion recognition [117], other works include ML classifiers that used features like MFCC [118][119], LPCC [120][121][122], and deep learning based approaches [123][124][125].

Survey Audio-Visual Expressed Emotion (**SAVEE)**[126] Here four English male actors in seven different emotional states are captured on camera and on audio (anger, contempt, fear, happiness, neutral, sadness, and surprise). 120 utterances were performed by each actor, for a total of 480 sentences. The performers' faces were painted with sixty markers for the visual elements. The recordings contain fifteen phonetically balanced sentences for each emotion: two emotion-specific, three common, and ten cliched phrases. Wavelet-based feature extraction techniques [127], DL strategies [128], gradient boosting classifiers [129], and multisource information fusion have all been tested with it [130][131].

Audiovisual Thai Emotion Database (Thai DB) [132**]**To construct this composition, six basic emotions are recorded namely, happiness, sadness, anger, disgust, fear, and surprise. Six kids who study drama made the statements. They were required to read one thousand of the most often used one- to seven-syllable Thai words. Recordings that were unrecognizable to human ears due to emotional content were removed. It has been used to evaluate traditional machine learning methods for emotion classification, like Support Vector Machines (SVM) [133].

Toronto Emotional Speech Set (TESS) [134**]**It has over 2,800 examples of two professional actresses' 200 isolated words. In this case seven emotions that are considered are happiness, pleasant surprise, sadness, anger, disgust, fear, and neutrality. Although it is also used with traditional ML techniques [137], it is commonly used for testing a type of DL models [135][136].

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[138**]**It is consists of 7356 audio and video clips combined. With the help of voice and song samples, models for the analysis of recorded music can be trained. In a set of 1440 speech audio samples, 24 experienced actors—12 men and 12 women—narrated two semantically neutral words in US English while displaying eight different emotions: calm, happy, neutral, sad, angry, fear, disgust, and surprise. It has been widely used to evaluate shallow neural networks and DL approaches, like the classical CNNs and LSTMs, as well as traditional ML techniques like SVM, gradient boosting, and the hybridization of feature extraction methods based on wavelets and spectral features[139][140].

Table 1. Some Important Results of SER using Deep Learning Techniques in Reviewed Papers

Features Selections	**Feature Vectors**		**ML and DL Algorithms**	**Recognition Rate**	**Reference Papers**
Sample: 16 kHz frequency Hamming Window framing	Prosodic (power, intensity); Spectral (LFPC, LPCC, MFCC); Acoustic (glottis, formant);	happiness, anger, anxiety,boredom, disgust, sadness, neural state	Autoencoder DBN Stacked	For the speaker-independent experiment, the best-case outcomes are as follows: 39.0% DBN SAE: 29.0%	[141]
Convert audio signals into two-dimensional graphs using STFT with a frame size of 256 and 50% overlap.	2-dimensional audio signal representations	happy , anger, boredom, neutral, fear, disgust, sadness,	CNNs with time-distributed RNNLSTMs Initials: CNNs; RNNLSTM	CNN distribution by TTime: 88.01%, 86.86%, 86.65% CNNs: 86.32%, 86.06%, and 87.74%79.87% / 78.83% / 78.31% RNN-LSTM	[142]
Framing: A 25 ms hammer window with a 10 ms stride is used. Filter banks based on the log Fourier transform with an efficiency of 40 on the Mel scale are gathered to produce a spectrogram for each frame.	Deep learning automatically extracts audio features	Happiness, sadness, anger,neutral	Each of the feed-forward and recurrent architectures, Utterance and Frame; Initials: Binary Decision Tree with Hierarchical Structure; DNN+ELM; SVM; ReplicatedSoftmax Models+SVM	Regarding the experiment without a speaker utilizing frame-based: 64.78% and 60.89%; DNN utilizing utterance-based: 57.74% and 58.28% A Hierarchical Binary Decision Tree It is -58.46% SVM with Softmax 57.39% of cases were repeated. 48.2% and 54.3% for DNN + ELM	[143]
Transformed into Spectrograms	Time-frequency spectrograms, or audio representation	melancholy, neutrality, fear, rage, contempt, and boredom	CNN	CNN: 84.3%	[144]
Pre-emphasis; Framing: 25 ms window with a 10-ms overlap	Spectral (MFCC); Acoustic (formant); Prosodic (pitch, energy, zero-crossing);	happiness, rage, fear, indifference, sadness, surprise, and mental state	SVM and DBN; Combining SVM with DBN	For gender-dependent DBN with SVM experimentation: 95.6% DBV stands for 94.6%. SVM: 84.8 percent.	[145]
Divided into syllabic segments	Both at the syllable (group 2) and utterance (group 1) levels are prosodic.	pleasure, rage, melancholy, neutral anxiety, boredom, and contempt	Group1 DNN Group2 DNN ,The fused model of Group1 and Group2 DNN	The fused DNNs: 61.68%, Group1 DNN :57.44%, Group2 DNN :58.88%	[146]
Extract spectral information from a speech signal with a 20 ms frame variation that is smaller than 20 ms.	Prosodic (pitch, intensity, jitter, shimmer); Spectral (MFCCs, formant); Statistical values (minimum, maximum, mean, standard deviation);	bliss, rage, terror, indifference, and melancholy	VQ ,KMeans; GMM and ANN	Regarding the MDB and IIT-KGP databases: GMM: 84% and 81% K-Means: 71% and 74% ANN: 72% and 79% VQ: 62% and 57%	[147]
Not stated	About the proposed method: F0, MFCC, energy boosted by delta and delta-delta, as shown by For baseline systems: the log-spectrogram feature with 128 filter-banks, the eGeMAPS and super-vector features	RAVDESS database: content, at ease, impartial, melancholy, irate, terrified, shocked, and disgusted	Baselines: TCapsNet; CNN with self-attention model; -RNN-LSTM; RNN-BLSTM; CNN; SVM	95.1% of the RAVDESS database's CNNs use the self-attention model. BLSTM: 63.9%; TCapsNet: 68.1 to 69.4. 51.3% for BLSTM with CNN 36.3% SVM 35.4% for CapsuleNet with BLSTM CNN: 34.6 %	

7. Future of Research Scope in SER

Multi-Modal Fusion: Continued exploration and refinement of multi-modal fusion techniques, integrating information from audio, text, and visual modalities, to capture richer emotional cues and improve recognition accuracy.

Context-aware Models: Development of context-aware SER models that consider situational context, speaker traits, and conversational dynamics to enhance the understanding of emotional expressions in real-world interactions.

Transfer Learning and Few-shot Learning:Advancements in transfer learning and few-shot learning techniques to address data scarcity and domain adaptation challenges, enabling SER models to generalize better across diverse datasets and conditions.

Interpretability and Explainability: Focus on designing interpretable and explainable SER systems to provide insights into model predictions, enhancing transparency, trust, and user acceptance in applications such as mental health monitoring and human-computer interaction.

Robustness and Generalization: Research efforts aimed at improving the robustness and generalization of SER models to environmental factors (e.g., noise, speaker variability) and demographic diversity, ensuring equitable performance across different populations and contexts.

Self-supervised Learning: Exploration of self-supervised learning techniques for pre-training SER models on large-scale unlabeled speech corpora, leveraging auxiliary tasks such as contrastive learning and reconstruction to learn robust representations of emotional content.

Incremental and Lifelong Learning: Development of incremental and lifelong learning approaches in SER, allowing models to adapt and evolve over time with new data and tasks while retaining previously learned knowledge, facilitating continual improvement and adaptation in real-world applications.

Ethical and Societal Implications: Increased attention to ethical considerations and societal implications of SER technology, including fairness, bias mitigation, privacy preservation, and user consent, to ensure responsible development and deployment in diverse socio-cultural contexts.

Real-world Applications: Translation of SER research into practical applications across domains such as affective computing, mental health monitoring, virtual assistants, education, entertainment, and customer feedback analysis, with a focus on usability, user experience, and societal impact.

Cross-disciplinary Collaborations: Collaboration between researchers from diverse fields such as psychology, linguistics, computer science, and neuroscience to advance our understanding of emotions, foster interdisciplinary innovation, and develop more holistic and human-centered SER solutions.

By addressing these future trends and challenges, researchers aim to push the boundaries of SER, creating more accurate, robust, and ethically informed systems capable of understanding and responding to human emotions in a wide range of contexts and applications.

As researchers continue to advance Speech Emotion Recognition (SER) technology,We have observed several challenges lie ahead in SER research that given below:

Real-world Data Variability: SER systems often struggle with variability in real-world data, including diverse accents, speech styles, recording conditions, and cultural differences. Developing models robust to such variability remains a significant challenge.

Subjectivity and Ambiguity: Emotions are inherently subjective and context-dependent, leading to ambiguity in labeling emotional states. Developing models capable of understanding and interpreting nuanced emotional expressions is challenging.

Contextual Understanding: Emotions are influenced by various contextual factors such as conversational context, speaker characteristics, and situational cues. Capturing and integrating contextual information into SER systems remains a challenge.

Limited Labeled Data: Annotated emotional speech datasets are often limited in size and diversity, hindering the training of robust SER models. Developing effective techniques for learning from limited labeled data is crucial for advancing the field.

Cross-cultural Variability: Emotion expression can vary across cultures and languages, leading to cultural biases in SER models trained on specific datasets. Addressing cross-cultural variability and developing culturally sensitive SER systems is essential.

Temporal Dynamics: Emotions evolve over time, and their expression in speech may exhibit complex temporal dynamics. Modeling and capturing these temporal dynamics effectively pose challenges for SER systems, particularly in real-time applications.

Multi-modal Integration: Integrating information from multiple modalities such as audio, text, and visual cues is challenging but crucial for improving SER accuracy. Developing effective fusion strategies and architectures for multi-modal SER remains an active area of research.

Ethical Considerations: SER technology raises ethical concerns related to privacy, bias, fairness, and the potential misuse of emotional analysis. Ensuring ethical development, deployment, and use of SER systems requires careful consideration and regulation.

Interpretability and Explainability: SER models often lack interpretability and explainability, making it difficult to understand their decision-making process. Developing transparent and interpretable SER systems is important for building trust and understanding in human-computer interactions.

User-centric Design: Designing SER systems that are user-centric, intuitive, and responsive to user needs and preferences poses a challenge. Considering user feedback and preferences in the development of SER technology is essential for its adoption and acceptance.

Addressing these challenges will require interdisciplinary collaboration, innovative research methodologies, and a deep understanding of human emotions and communication. Despite these challenges, ongoing advancements in machine learning, signal processing, and affective computing offer promising opportunities for overcoming barriers and advancing the state-of-the-art in Speech Emotion Recognition.

8. Conclusions

Lastly, in light of multiple research that attain high SER results in Speaker-Independent tasks. This paper offers potential explanations for the poor performance seen in the great majority of earlier research. It is easy to discover that Speaker-Dependent or Speaker-Independent trials have a significant influence on the SER performance. Previous studies that have been published have noted that speaker-dependent trials had a greater accuracy range than speaker-independent trials. Nevertheless, several studies are still achieving good speaker-independent trial SER accuracy. The unique methodologies used in the studies have the best chance of contributing to this outcome, barring any common techniques or methods. Examining these specific approaches or procedures can support future studies and applications on different datasets. Furthermore, these methods could be able to pinpoint the precise causes of why Speaker-Independent studies consistently yield worse performance when compared to Speaker-Dependent experiments. Thus, it is advised that future studies concentrate more on the developments and methods in the DL sector over the last four years and take into account the involvement of major global corporations like Google and Microsoft. Owing to the SER task's constrained space and complex dimensionality, certain discoveries are helpful in other fields. Furthermore, certain research that go into further detail about the methods for resolving the highly particular problems in SER cannot be included due to the high generalization of the search keywords. These restrictions may guide future research endeavors. These deep learning architectures can be adapted and combined in various ways to address specific challenges in SER, such as modeling temporal dynamics, capturing long-range dependencies, and integrating multi-modal information. Researchers continue to explore novel architectures and training strategies to improve the accuracy and robustness of SER systems in real-world applications.

Reference

[1] L. Deng, Deep learning: from speech recognition to language and multimodal processing, APSIPA Transactions on Signal and Information Processing 5 (2016).

[2] Panda, S.P.: Automated speech recognition system in advancement of human-computer interaction. In: Proc. IEEE 2017 International Conference on Computing Methodologies and Communication. pp. 302–306 (2017).

[3] A. Wali, Z. Alamgir, S. Karim, A. Fawaz, M.B. Ali, M. Adan, M. Mujtaba, Generative adversarial networks for speech processing: A review, Computer Speech & Language 72 (2022)

[4] A. D. Dileep and C. C. Sekhar, ‘‘GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines,’’ IEEE Trans. neural Netw. Learn. Syst., vol. 25, no. 8, pp. 1421–1432, Aug. 2014.

[5] Picard, R. W. (2000). Affective computing. MIT Press.

[6] Akçay, M. B., & Oguz, ˘ K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication, 116, 56–76

[7] Gadhe, R. P., & Deshmukh, R. R. (2015). Emotion recognition from isolated Marathi speech using energy and formants. International Journal of Computer Applications, 125.

[8] Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. Frontiers of Computer Science, 2, 14

[9] Mustafa, M. B., Yusoof, M. A., Don, Z. M., & Malekzadeh, M. (2018). Speech emotion recognition research: An analysis of research focus. International Journal of Speech Technology, 21, 137–156

[10] Mekruksavanich, S., Jitpattanakul, A., & Hnoohom, N. (2020). Negative emotion recognition using deep learning for Thai language. In 2020 Joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON) (pp. 71–74). IEEE.

[11] Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. Sensors, 17, 1694

[12] Ahmad, J., Muhammad, K., Kwon, S-i., Baik, S. W., & Rho, S. (2016). Dempster-Shafer fusion based gender recognition for speech analysis applications. In 2016 international conference on platform technology and service (PlatCon) (pp. 1–4). IEEE.

[13] Zhou, X., Guo, J., & Bie, R. (2016). Deep learning based affective model for speech emotion recognition. In 2016 Intl IEEE conferences on ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, Internet of people, and smart world congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld) (pp. 841–846). IEEE.

[14] Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mirza, F., Mathew, S., & Schneider, S. (2022). Automatic speech emotion recognition using machine learning: Digital transformation of mental health
[15] Rawat, A., & Mishra, P. K. (2015). Emotion recognition through speech using neural network. International Journal of Advanced Research in Computer Science and Software Engineering, 5, 422–428
[16] France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Transactions on Biomedical Engineering, 47, 829–837.
[17] Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. Artificial Intelligence Review, 43, 155–177
[18] Zhao, J., Mao, X., & Chen, L. (2019b). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47, 312–323.
[19] Alu, D., Zoltan, E., & Stoica, I. C. (2017). Voice based emotion recognition with convolutional neural networks for companion robots. Science and Technology, 20, 222–240.
[20] Costantini, G., Parada-Cabaleiro, E., & Casali, D. (2021). Automatic emotion recognition from DEMoS Corpus by machine learning analysis of selected vocal features. In Biosignals (pp. 357–364).
[21] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, 59, 103–126.
[22] Sarmah, K., Das, H.C., Rajbongshi, S.K and Gogoi,S.(2023) A Comprehensive Study on State-of-Art Learning Algorithms in Emotion Recognition, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321-8169, Volume: 11 Issue: 11, 717-732.
[23] Czerwinski, M., Hernandez, J., & McDuff, D. (2021). Building an AI that feels: AI systems with emotional intelligence could learn faster and be more helpful. IEEE Spectrum, 58, 32–38.
[24] Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. Frontiers of Computer Science, 2, 14
[25] Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R. K., et al. (2020). Speech emotion recognition using support vector machine. arXiv preprint, arXiv:2002.07590.
[26] Bhavan, A., Chauhan, P., Shah, R. R., et al. (2019). Bagged support vector machines for emotion recognition from speech. Knowledge-Based Systems, 184, Article 104886.
[27] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., & Mahjoub, M. A. (2018). Speech emotion recognition: Methods and cases study. In ICAART (2) (pp. 175–182)
[28] Vondra, M., & Vích, R. (2009). Recognition of emotions in German speech using Gaussian mixture models. In Multimodal signals: Cognitive and algorithmic issues (pp. 256–263). Springer.
[29] Umamaheswari, J., & Akila, A. (2019). An enhanced human speech emotion recognition using hybrid of PRNN and KNN. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon) (pp. 177–183). IEEE
[30] Yadav, S. P., Zaidi, S., Mishra, A., & Yadav, V. (2021). Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). Archives of Computational Methods in Engineering, 1–18
[31] Li, Y., Zhao, T., & Kawahara, T. (2019). Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In Interspeech (pp. 2803–2807).
[32] [32] Yogesh, C., Hariharan, M., Ngadiran, R., Adom, A. H., Yaacob, S., Berkai, C., & Polat, K. (2017b). A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. Expert Systems with Applications, 69, 149–158.
[33] El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44, 572–587
[34] Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. International Journal of Speech Technology, 21, 93–120
[35] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2021). Survey of deep representation learning for speech emotion recognition. IEEE Transactions on Affective Computing
[36] Liu, Z.-T., Wu, M., Cao, W.-H., Mao, J.-W., Xu, J.-P., & Tan, G.-Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing, 273, 271–280
[37] Fahad, M. S., Ranjan, A., Yadav, J., & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. Digital Signal Processing, 110, Article 102951.
[38] Harati, S., Crowell, A., Mayberg, H., & Nemati, S. (2018). Depression severity classification from speech emotion. In 2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 5763–5766). IEEE.
[39] Miner, A. S., Haque, A., Fries, J. A., Fleming, S. L., Wilfley, D. E., Wilson, G. T., Milstein, A., Jurafsky, D., Arnow, B. A., Agras, W. S., et al. (2020). Assessing the accuracy of automatic speech recognition for psychotherapy. npj Digital Medicine, 3, 1–8
[40] Mitsuyoshi, S., Nakamura, M., Omiya, Y., Shinohara, S., Hagiwara, N., & Tokuno, S. (2017). Mental status assessment of disaster relief personnel by vocal affect display based on voice emotion recognition. Disaster and Military Medicine, 3, 1–9
[41] Ali, H., Hariharan, M., Yaacob, S., & Adom, A. H. (2015). Facial emotion recognition using empirical mode decomposition. Expert Systems with Applications, 42, 1261–1277.
[42] Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. Speech Communication, 114, 22–35
[43] Madanian, S., Rasoulipanah, H., & Yu, J. (2023b). Stress detection on social network: Public mental health surveillance: Public mental health surveillance. In Proceedings of the 2023 Australasian computer science week ACSW '23 (pp. 170–175). New York, NY, USA: Association for Computing Machinery.
[44] Koolagudi, S. G., Murthy, Y. S., & Bhaskar, S. P. (2018). Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. International Journal of Speech Technology, 21, 167–183

[45] Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mirza, F., Mathew, S., & Schneider, S. (2022). Automatic speech emotion recognition using machine learning: Digital transformation of mental health
[46] Konar, A., & Chakraborty, A. (2015). Emotion recognition: A pattern analysis approach. John Wiley & Sons.
[47] Fahad, M. S., Ranjan, A., Yadav, J., & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. Digital Signal Processing, 110, Article 102951.
[48] Madanian, S., Nakarada-Kordic, I., Reay, S., & Chetty, T. (2023a). Patients' perspectives on digital health tools. PEC Innovation, 2, Article 100171. https://doi.org/ 10.1016/j.pecinn.2023.100171
[49] Kalhor, E., & Bakhtiari, B. (2021). Speaker independent feature selection for speech emotion recognition: A multi-task approach. Multimedia Tools and Applications, 80, 8127–8146.
[50] Sun, Y., & Wen, G. (2015). Emotion recognition using semi-supervised feature selection with speaker normalization. International Journal of Speech Technology, 18, 317–331.
[51] Liu, Z.-T., Xiao, P., Li, D.-Y., & Hao, M. (2019). Speaker-independent speech emotion recognition based on CNN-BLSTM and multiple SVMs. In International conference on intelligent robotics and applications (pp. 481–491). Springer.
[52] Abdelwahab, M., & Busso, C. (2017). Ensemble feature selection for domain adaptation in speech emotion recognition. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5000–5004). IEEE
[53] A. D. Dileep and C. C. Sekhar, ''GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines,'' IEEE Trans. neural Netw. Learn. Syst., vol. 25, no. 8, pp. 1421–1432, Aug. 2014.
[54] L. Deng and D. Yu, ''Deep learning: Methods and applications,'' Found. Trends Signal Process., vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.
[55] J. Schmidhuber, ''Deep learning in neural networks: An overview,'' Neural Network., vol. 61, pp. 85–117, Jan. 2015.
[56] Alu, D., Zoltan, E., & Stoica, I. C. (2017). Voice based emotion recognition with convolutional neural networks for companion robots. Science and Technology, 20, 222–240.
[57] Tashev, I. J., Wang, Z.-Q., & Godin, K. (2017). Speech emotion recognition based on Gaussian mixture models and deep neural networks. In 2017 information theory and applications workshop (ITA) (pp. 1–4). IEEE.
[58] Niu, Y., Zou, D., Niu, Y., He, Z., & Tan, H. (2017). A breakthrough in speech emotion recognition using deep retinal convolution neural networks. arXiv preprint, arXiv:1707.09917
[59] Jalal, M. A., Moore, R. K., & Hain, T. (2019). Spatio-temporal context modelling for speech emotion classification. In 2019 IEEE automatic speech recognition and understanding workshop (ASRU) (pp. 853–859). IEEE.
[60] Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. Electronics, 10, 1163.
[61] Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recognition Letters, 146, 1–7.
[62] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, ''Random deep belief networks for recognizing emotions from speech signals,'' Comput. Intell. Neurosci., vol. 2017, Mar. 2017, Art. no. 1945630.
[63] A. Mohamed, G. E. Dahl, and G. Hinton, ''Acoustic modeling using deep belief networks,'' IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 14–22, Jan. 2012.
[64] Shih, P.-Y., Chen, C.-P., & Wang, H.-M. (2017). Speech emotion recognition with skewrobust neural networks. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2751–2755). IEEE.
[65] Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53, 1062–1087.
[66] Torres-Carrión, P. V., González-González, C. S., Aciar, S., & Rodríguez-Morales, G. (2018). Methodology for systematic literature review applied to engineering and education. In 2018 IEEE global engineering education conference (EDUCON) (pp. 1364–1373). IEEE.
[67] Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. Health Information and Libraries Journal, 26, 91–108.
[68] Ma, F.; Zhang, W.; Li, Y.; Huang, S.-L.; Zhang, L. An End-to-End Learning Approach for Multimodal Emotion Recognition: Extracting Common and Private Information. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1144–1149.
[69] Tang, G.; Xie, Y.; Li, K.; Liang, R.; Zhao, L. Multimodal emotion recognition from facial expression and speech based on feature fusion. Multimed. Tools Appl. 2023, 82, 16359–16373.
[70] Ghaleb, E.; Popa, M.; Asteriadis, S. Metric Learning-Based Multimodal Audio-Visual Emotion Recognition. IEEE MultiMedia 2020, 27, 37–48.
[71] Nie, W.; Ren, M.; Nie, J.; Zhao, S. C-GCN: Correlation Based Graph Convolutional Network for Audio-Video Emotion Recognition. IEEE Trans. Multimed. 2021, 23, 3793–3804.
[72] Farhoudi, Z.; Setayeshi, S. Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition. Speech Commun. 2021, 127, 92–103.
[73] Nandi, A.; Xhafa, F. A federated learning method for real-time emotion state classification from multi-modal streaming. Methods 2022, 204, 340–347.
[74] Chhikara, P.; Singh, P.; Tekchandani, R.; Kumar, M.; Guizani, M. Federated Learning Meets Human Emotions: A Decentralized Framework for Human–Computer Interaction for IoT Applications. IEEE Internet Things J. 2021, 8, 6949–6962.
[75] Salman, A.; Busso, C. Privacy Preserving Personalization for Video Facial Expression Recognition Using Federated Learning. In Proceedings of the ICMI '22: 2022 International Conference on Multimodal Interaction, Bangalor, India, 7–11 November 2022; pp. 495–503.
[76] Chang, Y.; Laridi, S.; Ren, Z.; Palmer, G.; Schuller, B.W.; Fisichella, M. Robust Federated Learning Against Adversarial Attacks for Speech Emotion Recognition. arXiv 2022, arXiv:2203.04696.

[77] Zhang, T.; Feng, T.; Alam, S.; Lee, S.; Zhang, M.; Narayanan, S.S.; Avestimehr, S. FedAudio: A Federated Learning Benchmark for Audio Tasks. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5
[78] Stappen, L., Baird, A., Rizos, G., Tzirakis, P., Du, X., Hafner, F., Schumann, L., MallolRagolta, A., Schuller, B. W., Lefter, I., Cambria, E., & Kompatsiaris, I. (2020). MuSe 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In Proceedings of the 1st international on multimodal sentiment analysis in real-life media challenge and workshop MuSe'20 (pp. 35–44). New York, NY, USA: Association for Computing Machinery.
[79] Latif, S., Khalifa, S., Rana, R., & Jurdak, R. (2020). Poster abstract: Federated learning for speech emotion recognition applications. In 2020 19th ACM/IEEE international conference on information processing in sensor networks (IPSN) (pp. 341–342).
[80] Tsouvalas, V., Ozcelebi, T., & Meratnia, N. (2022). Privacy-preserving speech emotion recognition through semi-supervised federated learning. In 2022 IEEE international conference on pervasive computing and communications workshops and other affiliated events (PerCom workshops (pp. 359–364).
[81] S. Amiriparian, L. Christ, A. König, E.-M. Meßner, A. Cowen ,E. Cambria, & B.W. Schuller. Muse 2022 challenge: Multimodal humour, emotional reactions, and stress. In Proceedings of the 30th ACM international conference on multimedia MM '2022 (pp. 7389–7391). New York, NY, USA: Association for Computing Machinery
[82] T. Feng and S. Narayanan, Semi-FedSER: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling. In Proc. interspeech 2022, pp. 5050–5054
[83] T. Vogt and E. André, ''Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition,'' in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2005, pp. 474–477.
[84] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, ''Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011,'' Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177, 2015.
[85] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, ''The automatic recognition of emotions in speech,'' in Emotion-Oriented Systems. Springer, 2011, pp. 71–99.
[86] E. Mower, M. J. Mataric, and S. Narayanan, ''A framework for automatic human emotion classification using emotion profiles,'' IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1057–1070, Jul. 2011.
[87] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, ''Prediction-based learning for continuous emotion recognition in speech,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5005–5009.
[88] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' Nature, vol. 521, no. 7553, p. 436, 2015.
[89] W. Wang, Ed., Machine Audition: Principles, Algorithms and Systems. Hershey, PA, USA: IGI Global, 2010.
[90] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, ''Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,'' Speech Commun., vol. 53, nos. 9–10, pp. 1062–1087, Nov./Dec. 2011.
[91] M. Sidorov, S. Ultes, and A. Schmitt, ''Emotions are a personal thing: Towards speaker-adaptive emotion recognition,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2014, pp. 4803–4807.
[92] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, ''Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling,'' in Proc. INTERSPEECH, Makuhari, Japan, 2010, pp. 2362–2365.
[93] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, ''A breakthrough in speech emotion recognition using deep retinal convolution neural networks,'' 2017, arXiv:1707.09917. [Online]. Available: https://arxiv. org/abs/1707.09917
[94] Y. Bengio, ''Learning deep architectures for AI,'' Found. Trends Mach. Learn., vol. 2, no. 1, pp. 1–127, 2009.
[95] M. W. Bhatti, Y. Wang, and L. Guan, ''A neural network approach for human emotion recognition in speech,'' in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), vol. 2, May 2004, p. II-81.
[96] K. Poon-Feng, D.-Y. Huang, M. Dong, and H. Li, ''Acoustic emotion recognition based on fusion of multiple feature-dependent deep Boltzmann machines,'' in Proc. IEEE 9th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP), Sep. 2014, pp. 584–588.
[97] V. Nair and G. E. Hinton, ''Rectified linear units improve restricted Boltzmann machines,'' in Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010, pp. 807–814.
[98] G. E. Hinton, S. Osindero, and Y.-W. Teh, ''A fast learning algorithm for deep belief nets,'' Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.
[99] A.-R. Mohamed, G. Dahl, and G. Hinton, ''Deep belief networks for phone recognition,'' in Proc. NIPS Workshop Deep Learn. Speech Recognit. Rel. Appl., Vancouver, BC, Canada, 2009, vol. 1, no. 9, p. 39.
[100] J. Lee and I. Tashev, ''High-level feature representation using recurrent neural network for speech emotion recognition,'' in Proc. 16th Annu. Conf. Int. Speech Commun. Assoc., 2015, pp. 1537–1540.
[101] Y. Kamp and M. Hasler, Recursive Neural Networks for Associative Memory. Hoboken, NJ, USA: Wiley, 1990.
[102] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, ''Random deep belief networks for recognizing emotions from speech signals,'' Comput. Intell. Neurosci., vol. 2017, Mar. 2017, Art. no. 1945630.
[103] A. Mohamed, G. E. Dahl, and G. Hinton, ''Acoustic modeling using deep belief networks,'' IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 14–22, Jan. 2012.
[104] C. Huang, W. Gong, W. Fu, and D. Feng, ''A research of speech emotion recognition based on deep belief network and SVM,'' Math. Problems Eng., vol. 2014, Aug. 2014, Art. no. 749604.
[105] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, ''Greedy layerwise training of deep networks,'' in Proc. Adv. Neural Inf. Process. Syst., 2007, pp. 153–160.
[106] W. Q. Zheng, J. S. Yu, and Y. X. Zou, ''An experimental study of speech emotion recognition based on deep convolutional neural networks,'' in Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII), Sep. 2015, pp. 827–831.
[107] Y. Kim, ''Convolutional neural networks for sentence classification,'' 2014, arXiv:1408.5882. [Online]. Available: https://arxiv.org/abs/1408.5882

[108] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: Towards a new generation of databases, Speech Communication 40 (2003) 33–60.
[109] Bhutekar, S.D., Chandak, M.B.: Designing and recording emotional speech databases. In: National Conference on Innovative Paradigms in Engineering & Technology (NCIPET-2012). pp. 6–10 (2012
[110] Engberg, I.S., Hansen, A.V.: Documentation of the Danish emotional speech database. Tech. rep., Center for Person Kommunilation, Denmark (1996).
[111] I.S. Engberg, A.V. Hansen, O. Andersen, P. Dalsgaard, Design, recording and verification of a Danish emotional speech database, in: Proc. 5th European Conf. Speech Communication and Technology, 1997, pp. 1695–1698.
[112] V.M. Chavan, V.V. Gohokar, Speech emotion recognition by using SVMclassifier, Int. J. Engineering and Advanced Technology 1 (5) (2012) 11–15
[113] Y.L. Lin, G. Wei, Speech emotion recognition based on HMM and SVM, Proc. Fourth IEEE Int. Conf. on Machine Learning and Cybernetics. (2005) 4898– 4901.
[114] M. Kotti, C. Kotropoulos, Gender classification in two emotional speech databases, in: Proc. 19th Int. Conf. on Pattern Recognition, 2008, pp. 1–4.
[115] Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. pp. 80–804 (2013)
[116] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, in: Proc. 9th European Conf. Speech Communication and Technology, 2005, pp. 1517–1520
[117] Lotfidereshgi, R., Gournay, P.: Biologically inspired speech emotion recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing. pp. 5135–5139 (2017).
[118] Garg, V., Kumar, H., Sinha, R.: Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers (2013).
[119] F. Daneshfar, S.J. Kabudian, A. Neekabadi, Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier, Applied Acoustics 166 (2020).
[120] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M.A. Mahjoub, C. Cleder, in: A. Cano (Ed.), Automatic speech emotion recognition using machine learning, Social Media and Machine Learning. IntechOpen, 2019
[121] Y. Gao, B. Li, N. Wang, T. Zhu, Speech emotion recognition using local and global features, Int. Conf. Brain Informatics (2017) 3–13.
[122] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M.A. Mahjoub, C. Cleder, Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO, Speech Communication 114 (2019) 22–35.
[123] T. Anrarjon, Kwon Mustaqeem, S.: Deep-net: A lightweight CNN-based speech emotion recognition system using deep system using deep, Sensors 20 (2020) 5212.
[124] Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network (2017).
[125] Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. pp. 80–804 (2013).
[126] Haq, S., Jackson, P.J.B.: Multimodal emotion recognition. In: Wang, W. (ed.) Machine audition: Principles, algorithms and systems, pp. 398–423. IGI Global (2010).
[127] H.K. Palo, M.N. Mohanty, Wavelet based feature combination for recognition of emotion, Ain Shams Engineering Journal 9 (4) (2018) 1799–1806
[128] K. Manohar, E. Logashanmugam, Hybrid deep learning with optimal feature selection for speech emotion recognition using improved meta-heuristic algorithm, Knowledge-Based Systems 246 (2022)
[129] Iqbal, A., Barua, K.: A real-time emotion recognition from speech using gradient boosting. In: Proc. Int. Conf. Electrical, Computer and Communication Engineering. pp. 1–5 (2019).
[130] X. Li, M. Akagi, Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model, Speech Communication 110 (2019) 1–12
[131] L. Chen, K. Wang, M. Li, M. Wu, W. Pedrycz, K. Hirota, K-means clusteringbased kernel canonical correlation analysis for multimodal emotion recognition in human-robot interaction, IEEE Transactions on Industrial Electronics (2022).
[132] Stankovic´, T., Karnjanadecha, M., Delic´, V.: Improvement of Thai speech emotion recognition by using face feature analysis. In: Int. Symposium Intelligent Signal and Communication Systems. pp. 1–5 (2011).
[133] Seehapoch, T., Wongthanavasu, S.: Speech emotion recognition using support vector machines. In: Int. Conf. Knowledge and Smart Technology. pp. 86–91 (2013).
[134] K. Duouis, M.K. Pichora-Fuller, Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set, Canadian Acoustics - Acoustique Canadienne 39 (3) (2011) 182–183
[135] M. Gokilavani, H. Katakam, S.A. Basheer, P. Srinivas, Ravdness, crema-d, tess based algorithm for emotion recognition using speech, in: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2022, pp. 1625–1631
[136] Slimi, A., Hamroun, M., Zrigui, M., Nicolas, H.: Emotion recognition from speech using spectrograms and shallow neural networks. In: ACM Int. Conf. Advances in Mobile Computing & Multimedia. pp. 298–301 (2020).
[137] K.V. Krishna, N. Sainath, A.M. Posonia, Speech emotion recognition using machine learning, in: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), 2022, pp. 1014–1018.
[138] S.R. Livingstone, F.A. Russo, The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facieal and vocal expressions in North American English, PLoS ONE 13 (5) (2018).
[139] D. Issa, M. Faith-Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control 59 (2020).
[140] Huang, A., Bao, P.: Human vocal sentiment analysis. arXiv, 1905.08632 (2019)

[141] Zhou, X., Guo, J., & Bie, R. (2016). Deep learning based affective model for speech emotion recognition. In 2016 Intl IEEE conferences on ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, Internet of people, and smart world congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld) (pp. 841–846). IEEE
[142] Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA) (pp. 1–4).
[143] Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92, 60–68.
[144] Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In 2017 international conference on platform technology and service (PlatCon) (pp. 1–5). IEEE
[145] Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. Sensors, 17, 1694
[146] Koolagudi, S. G., Murthy, Y. S., & Bhaskar, S. P. (2018). Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. International Journal of Speech Technology, 21, 167–183
[147] Jalal, M. A., Moore, R. K., & Hain, T. (2019). Spatio-temporal context modelling for speech emotion classification. In 2019 IEEE automatic speech recognition and understanding workshop (ASRU) (pp. 853–859). IEEE