International Research Journal of
SCIENCE, TECHNOLOGY, EDUCATION, AND MANAGEMENT
P-ISSN: 2799-063X | E-ISSN: 2799-0648
ISO 9001:2015 Certified

Volume 3, No. 2 | June 2023

Deep learning techniques for speech emotion recognition: A review

Silviana Widya Lestari[1], Saliyah Kahar[2], Trismayanti Dwi[3]
[1,2]*Management & Science University, Malaysia*
[3]*State Polytechnic of Jember, Indonesia*
Corresponding email: silvianawidya46@gmail.com

ABSTRACT

Speech emotion recognition is gaining significant importance in the domains of pattern recognition and natural language processing. In recent years, there has been notable progress in voice emotion detection within this field, primarily attributed to the successful application of deep learning techniques. Some research in this area lacks a thorough comparative study of different deep learning models and techniques related to speech emotion detection. This makes it difficult to identify the best performing approaches and their relative strengths and weaknesses. Therefore, the purpose of this work is to provide a comprehensive overview and provide a detailed overview of deep learning methods for speech emotion detection. The method used is a comparative literature analysis of previous articles that are relevant to the topic, which are related to both the methods of deep learning and the collections of data. The datasets that to be analyzed include the EMO-DB, RAVDESS, TESS, CREMA-D, IEMOCAP, and Danish Emotional Speech Databases. The language that used in the dataset is English, except for EMO-DB which used German language and Danish Emotional Speech Database that used Danish language. Most of the emotion types extracted from these datasets included basic emotions such as happiness, sadness, neutrality, disgust, surprise, and anger. The results of this review show that the application of deep learning techniques has made significant progress in the introduction of speech emotion detection. Complex deep learning models, for instance the CNN-RNN combination, can extract relevant acoustic features and produce accurate results in recognizing emotion from speech. This advancement has significant implications for various applications, including human computer interaction, affective computing, call center analytics, psychological research, and clinical diagnosis.

ARTICLE INFO

Received : May 23, 2023
Revised : June 17, 2023
Accepted : June 30, 2023

KEYWORDS

Convolutional neural network, Deep learning, Deep neural network, Recurrent neural network, Speech emotion recognition

Suggested Citation (APA Style 7th Edition):

Lestari. S.W., Kahar, S., & Dwi, T. (2023). Deep learning techniques for speech emotion recognition: A review. *International Research Journal of Science, Technology, Education, and Management,* 3(2), 78-91. https://doi.org/10.5281/zenodo.8139722

INTRODUCTION

The field of the identification of emotions from speech has transitioned from evolving into a specialized area to a crucial component in Human Computer Interaction (Chen et al., 2017). One major factor that contributed to this transition is the improvement in speech recognition accuracy. Another factor is the rise of virtual assistants and voice-controlled devices. Examples of technologies include Amazon's Alexa, Apple's Siri, Microsoft's Cortana, and Google Assistant have gained widespread adoption, bringing speech-based interactions into people's everyday lives. These virtual assistants can understand and respond to voice commands, enabling users to perform various tasks, such as playing music, sending messages, creating reminders, and managing smart home devices, solely through voice commands.

Research published by Lane & Georgiev (2015) facilitates seamless and intuitive communication between humans and machines through direct speech interaction, eliminating the need for traditional input devices and allowing human listeners to effortlessly respond. A system has been proposed that aims to allow Therefore, the aim of this study is to provide a comprehensive overview of deep learning techniques used to recognize emotions from speech. The method used is a comparative literature analysis of previous articles related to deep learning techniques and datasets on this subject. The composition of this work is as follows. The purpose of this study is described in Section II. Section II provides an overview of the methodology. Section III of this document provides an overview of deep learning techniques used to recognize emotion in speech. Section IV focuses on the datasets required for speech emotion detection. In addition, Section V presents findings and discussion, and Section VI presents conclusions.

Various applications are emerging, such as implementation of spoken dialogue systems for different scenarios, such as conversations in call centers, vehicle driving systems, and the use of emotional patterns from speech in medical applications (Nassif et al., 2019). However, despite advancements, there are still numerous challenges that HCI systems must address as they transition from laboratory testing to real-world implementation (Balomenos et al., 2005). One of the most challenging aspects of dealing with issues related to robustness, variability in speech signals, environmental noise, and the need for real-time processing is the ability to maintain a high level of accuracy and efficiency.

Speech signals can be incredibly diverse, with variations in tone, pitch, and pronunciation, making it difficult to develop algorithms that can accurately interpret and understand the intended message. Additionally, environmental noise further complicates the process, as it introduces additional disturbances that can affect the quality of the input signal. Moreover, real-time processing adds another layer of complexity, requiring systems to provide prompt and accurate responses without sacrificing performance. Successfully addressing these challenges is crucial for developing robust and reliable speech processing systems that can effectively handle the complexities of everyday communication. Therefore, it is imperative to make concerted efforts in effectively resolving these issues to achieve improved machine-based emotion recognition. Identifying the emotional condition of individuals is a unique undertaking and can serve as a benchmark for any model designed for recognizing emotions. Among the wide range of models used to classify these emotions, discrete emotion models are an essential approach. The emotion categories covered are boredom, anger, surprise, disgust, fear, happiness, neutrality, sadness, and joy (Vogt et al., 2008). Another significant model employed is a continuous three-dimensional space that encompasses arousal, valence, and potency as parameters.

Deep learning, a field that has gained increasing attention in machine learning, is recognized as a developing area of research (Schmidhuber, 2015). Deep Learning methods offer various benefits for Speech Emotion Recognition (SER) when compared to conventional approaches. These advantages include the ability to identify intricate patterns and characteristics without relying on manual feature extraction or tuning. Moreover, deep learning techniques can extract basic features from raw data and excel at processing unlabelled data. A deep neural network (DNN) is equipped with a feedforward structure containing a hidden layer between the input and output. These architectures, including DNNs and CNNs, yield efficient results in image and video processing. On the other hand, iterative architectures such as RNNs and LSTMs are particularly effective for language-related classification tasks such as NLP and speech emotion recognition (Schmidhuber, 2015). Although these models are effective classifiers, they have certain limitations. For example, CNNs excel at learning features from high-dimensional input data, but they also tend to learn features from subtle variations and biases and require large amounts of storage space. Similarly, LSTM-based RNNs excel at handling variable input data and modelling long-range continuous text data.

OBJECTIVES OF THE STUDY

The objective of this study is to conduct a comprehensive review of deep learning techniques for SER. The review aims to offer valuable information on the latest cutting-edge techniques and progressions in the fields, as well as identify potential challenges and future research directions.

The study aims to achieve the following specific goals, include:

1. To examine different deep learning approaches employed in the identification of emotions in speech, including CNNs, RNNs, and their variations.
2. To assess the effectiveness of various deep learning technique in terms of accuracy and generalization capabilities.
3. To identify limitations and challenges faced by existing deep learning technique in SER.

The main aim of this comprehensive overview is to offer researchers a thorough understanding of the current state of deep learning techniques in speech emotion detection. By achieving this objective, the study intends to establish a strong foundation for future advancements in this rapidly evolving field.

METHODS

The method that used for this study is literature analysis of articles relevant to the topic. This method involves review and evaluation of existing literature regarding the use of deep learning techniques in emotion recognition from speech. This method includes systematic steps to collect relevant articles, understand and analyse the content of each article, and evaluate the reliability and validity of the sources used. In this review, the focus is on recent studies using deep learning techniques to recognize emotion in speech. Relevant articles are collected from various sources, such as scientific journals, conferences, and other publications. After collecting the articles, the content of each article was carefully studied to understand the methods used, the data used, and the results achieved. The reliability and validity of the articles were also evaluated considering factors such as the research methodology, sample size and statistical analysis performed. Through this method, it is hoped that a comprehensive understanding of employing Deep Learning methodologies in recognizing emotions from speech can be obtained, as well as finding the latest findings that can be used to develop further research in this field.

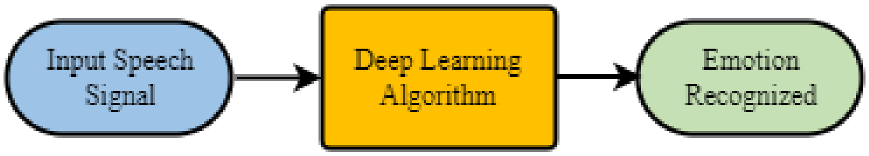

Fig. 1. Deep Learning Flow Mechanism

Input Speech Signal

Voice Input uses deep learning to recognize emotions in speech sound signal that is used as input in the process of recognizing voice emotions using the Deep Learning method. Thos sound signal can be in the form recordings of human voices speaking or making sounds with various emotional expressions. According to Nassif et al., (2019), speech signal offer diverse types of information including speech recognition for understanding the components encompassed in spoken signals, such as speaker recognition for discerning the speaker's identity, emotion detection to detect the emotional state of the speaker, Health detection to assess the patient's health, Voice detection to determine the spoken word, Accent detection to identify the accent of the speaker, Age detection to estimate the age of the speaker, Speaker Gender-discriminating gender detection.

Speaker recognition can be divided into two main component categories. There are two categories of speaker recognition: speaker identification and speaker verification (also called authentication). The primary objective of the speaker identification phase is to determine the registered speaker to whom a particular utterance belongs. This aspect finds practical applications in various setting such as public facilities and media environments. Examples of such scenarios include interactions with governmental or district institutions, phone conversations directed to radio stations, communications with insurance agencies, or documented exchanges (Reynolds, n.d., 2002).

Emotion recognition involves the task of identifying an unfamiliar emotion by analysing information embedded in speech signals. This field encompasses two branches, there emotion identification and emotion verification are the two branches of emotion detection. The goal of the first branch is to identify unknown emotions. The second branch, on the other hand, is to verify the presence of a particular emotion by comparing the input audio signal to different emotion models to find the best match. The aim of emotion detection is to determine whether a particular emotion belongs to a known category or represents an unknown emotion. Emotion detection has many applications, such as capturing the emotional state of speakers during telephone conversations in call centers and providing feedback to operators for monitoring purposes (Petrushin, 2020). Other applications include voicemail classifying her messages based on the expressed emotion of the caller, identifying individuals exhibiting emotional speech patterns (such as happiness) in environments where distrustfulness needs to be recognized This includes doing.

According to Swain et al., (2018), voice signals have become mankind's most advanced and fastest means of communication. It involves complex signal processing systems, networking, and diverse signalling units that convey information about the message, speaker, and language. In recent decades, much research has been devoted to converting human speech into word strings. Despite these advances, there remains a significant disparity between and machines, as machines lack the ability to comprehend the speaker’s emotional state, thus failing to interpret individual emotions The emergence of a new research field called speech emotion recognition focuses on comprehending and extracting specific emotions from the speech signal. The aim is to improve speech as the most efficient means of interaction between humans and machines. However, a significant challenge lies in the limited intelligence of machines to accurately recognize human voices. Researchers are working towards overcoming this obstacle and enhancing the capability of machines to accurately perceive and interpret human speech.

Deep Learning Algorithm

Deep learning algorithm, also known as deep neural networks, are primarily based on artificial neural networks. The term "deep" refers to the presence of multiple hidden layers within these networks, in contrast to traditional neural networks with few hidden layers, which typically extend to hundreds of layers. In recent years, deep learning algorithms have shown superior performance compared to traditional machine learning algorithms, and research focus has shifted to deep learning algorithms. This trend can also be observed in his SER region. The advantage of these algorithms is that the relevant features are automatically determined by the deep learning process, eliminating the need for separate feature extraction and selection steps. As Akçay & Oğuz (2020) point out, CNNs and RNNs are among the most commonly used deep learning algorithms in SER research.

Recurrent Neural Networks (RNNs)

RNNs such as LSTM and GRU have been used to capture the sequential nature of data and classify emotions. In particular, LSTM-RNN has been applied to speech emotion detection research. RNNs have limited short-term memory, but the integration of the LSTM architecture allows RNNs to effectively access long-term memory. LSTM RNN, a type of gate RNN, addresses the challenge of managing long-term dependencies. They use special "LSTM cells" that have inner repeats in addition to the outer repeats of the RNN. In addition to the standard input/output mechanism, LSTM RNNs include gate units with additional parameters and sigmoidal nonlinearities. The gating units in LSTM (Long Short-Term Memory) networks play a crucial role in regulating the flow of information by determining whether to store, allow input, output, or delete information. An LSTM cell typically includes three types of gates: the main gate, forget gate, and remember gate. These gates, through their opening and closing actions, enable LSTM cells to make decisions regarding information storage and control the timing of input, output, and erasure processes (Akçay & Oğuz, 2020).

In a study that conducted by Lee and Tashev (2015), a system called RNN (prop.)-ELM is introduced, showcasing significant advancements in the field of speech emotion detection. The suggested system exhibits noteworthy performance improvements, with a substantial 12% increase in accuracy, elevating it from 52.13% to 63.89%. Additionally, in two different scenarios, the system demonstrates a commendable 5% increase, from 57.91% to 62.85%. The paper introduces an original framework for speech emotion recognition that utilizes a recurrent neural network (RNN) and integrates a proficient learning method. This novel approach successfully addresses the difficulties related to capturing extensive contextual influences in emotional speech and effectively manages the uncertainty inherent in emotional labelling. By integrating recurrent neural networks and employing a maximum likelihood-based learning strategy, the proposed approach not only provides valuable insights into emotion recognition but also delivers remarkable advancements in the field.

According to Ho et al. (2020), accurate emotion recognition systems play an important role in the advancement of human behavior informatics and the design of efficient systems for human-machine interaction. These systems play a key role in efficiently and reliably processing human behavioral data, enabling seamless and natural communication. In this study, we propose a multimodal approach to speech emotion recognition that combines recurrent neural networks and self-versus-multiple-headed attention mechanisms. This framework includes her two types of linguistic representations: Word embedding of MFCC audio signals and text data. By training these functions simultaneously in the time domain, we achieve excellent performance on the IEMOCAP, MELD, and CMU-MOSEI datasets. However, there is room for improvement in this situation. Instead of merging modalities at a later stage, you can consider synchronizing audio and text data at a lower level of presentation to create a stronger relationship between the two. Additionally, we may consider integrating other audio features such as perceptual linear prediction (PLP), chroma, and prosody. For text data, she focused on selecting emotionally relevant words in utterances and filtering out information not emotionally relevant to utterances. Moreover, we see an emerging trend of applying domain-her matching techniques to transfer knowledge from speech recognition techniques to emotion recognition through a pre-training and subsequent refinement process.

One year after that, Byun & Lee (2021), conducted to develop a Korea emotional speech database for the specific purpose of analysis in speech emotion. The researcher introduced a feature combination technique aimed at improving emotion recognition performance using recurrent neural network techniques. This study focused on classifying emotions from linguistic data and used the Korean Emotion Language Database as a dataset. Studies have found that speech is fragmented into shorter intervals and recognizing emotions in those intervals yielded a higher recognition accuracy of 83.81%. This was compared to the accuracy achieved by recognizing speech emotions using a single long LSTM model, which achieved an accuracy of 75.51%. Thus, the study highlighted the effectiveness of utilizing shorter intervals to enhance the recognition of emotions in speech analysis.

Deep Neural Networks (DNNs)

One of way to applied Deep Neural Networks (DNNs) in speech emotion recognition studies is using Attention mechanisms. Attention mechanisms have been incorporated into DNN architectures to focus on relevant parts of the speech signal for emotion classification. In a study by Lieskovská et al. (2021), the researchers primarily reviewed recent advances in SER and sought to analyze the effects of different attentional mechanisms on her SER system's performance. was intended. To conduct the study, they used a DNN model that incorporates attentional mechanisms. The researchers used three widely used datasets: IEMOCAP, Emo-DB, and RECOLA to assess how different attentional mechanisms affected the performance of her SER system. The study's findings indicated that the results demonstrated that the integration of attentional mechanisms significantly improved the performance of the SER system. This suggests that attentional mechanisms play an important role in improving the accuracy and efficacy of his SER.

Scheidwasser-Clow et al., (2022) introduced the SER Adaption Benchmark (SERAB), a framework designed to assess the effectiveness and adaptability of various approaches using Deep Neural Networks (DNNs) for utterance-level SER. The purpose of SERAB (SER Analysis Benchmark) is to assess the performance and generalization capabilities of various methods in the field of SER. Given the rapid progress in DNN-based representations for speech and emotion, SERAB serves as a comprehensive platform for comparing different techniques. It includes tasks that span multiple languages, dataset sizes, and sentiment categories, and provides reliable estimates of performance and ability to generalize. to evaluate the effectiveness of SERAB, a range of recent baselines was assessed. Notably, approaches based on BYOL (Bootstrap Your Own Latent)-based methods demonstrated superior performance across all metrics considered. Surprisingly, when BYOL-A models were pre-trained solely on speech samples from Audio Set (BYOL-S), there was an impressive 3% increase in accuracy in comparison to the initial approach. These evaluation outcomes can establish fundamental benchmarks for the advancement of new methodologies, such as the exploration of CvT-based methods discussed in this study.

Convolutional Neural Networks (CNNs)

A CNN, or Shift-Invariant Artificial Neural Network (SIANN), is a specific type of neural network architecture that incorporates dedicated filters or regions within its hidden layer. This design allows the CNN to have a shift-invariant property, meaning it can effectively recognize patterns and features in input data regardless of their spatial location. These filters or regions are specifically designed to identify and capture specific features present in the input signal. The design of their system is based on (Hubel & Wiesel, 1968), which introduced the concept visual neural cortex is a specialized structure that plays a crucial role in processing visual information. That responds to specific characteristics of the input signal. CNNs play a crucial role in speech emotion recognition by extracting features, analyzing time-frequency patterns, modelling spatial information, leveraging transfer learning, incorporating multi-modal data, utilizing attention mechanisms, and participating in ensemble modelling. These applications contribute to advancing the accuracy and effectiveness of emotion recognition systems. One advantageous aspect of Convolutional Neural Networks (CNNs) has the capability to acquire features from intricate input data. However, this feature has a large memory footprint during development as it learns features from even the slightest changes and distorted appearances. Therefore, CNNs usually include a convolutional layer followed by a down sampling mechanism to address this issue. A convolutional layer consists of multiple filter banks, whose weights are adjusted by backpropagation, as proposed by Weng et al. (1993).

Bertero & Fung (2017), in their work, presented a convolutional neural network that can identify emotions such as anger, happiness, and sadness, and achieve 66.1% accuracy. The researchers performed a comparison between the trained network and the basic feature based His SVM. To train and evaluate this approach, we used a dataset of his TED talks, manually labelled by a student, and collected from a crowd. The CNN implementation was done using the Theano toolkit. In addition, we used features from the INTERSPEECH 2009 Emotion Challenge to train a linear SVM to enable comparative analysis.

In their study, Zhao et al., (2019) focused on developing a technique for acquiring deep emotional features for accurately recognizing speech emotion. To accomplish this, they employed a combined of CNNs and LSTM models. Through rigorous experimentation, the researchers observed that designed networks yielded exceptional performance in the task of speech emotion recognition. These encouraging findings underscore the efficacy of their method in capturing and comprehending the subtle emotional cues present in spoken language. The research conducted by Zhao et al. constitutes a noteworthy contribution to the field of emotion recognition and sets the stage for future advancements in this area.

In a study conducted by Mekruksavanich et al. (2020), they conducted an experiment utilizing a one-dimensional convolutional neural network. Their findings revealed an impressive classification accuracy of 96.60% when it came to identifying negative emotions in Thai language datasets. Furthermore, the researchers applied their developed method to various datasets, including SAVEE, RAVDESS, TESS, Cream-D, and Thai datasets.

In the same year, Anvarjon et al. (2020) conducted research aimed at introducing a novel and efficient approach to speech emotion detection (SER) by developing a lightweight model with a balance of low computational complexity and high detection accuracy. Did. To achieve this goal, they chose his CNN approach. To evaluate the effectiveness of the proposed system, researchers used two known datasets, IEMOCAP and Emo DB language records. The research results show that the CNN-based His-SER system they developed outperforms the current state-of-the-art His-SER system in terms of detection accuracy. This work has the potential to advance the SER field and provide more efficient and effective solutions for real-world applications.

Auto Encoder

Autoencoders can be used in speech emotion recognition as a feature learning and dimensionality reduction technique. Autoencoders offer advantages in speech emotion recognition by leveraging unsupervised learning, performing feature learning, reducing dimensionality, and providing denoising capabilities. These advantages make Autoencoders a valuable tool for extracting informative features and improving the effectiveness of emotion recognition systems, particularly in situations with limited labelled data. A study that conducted by Aouani & Ayed (2020) proposed a novel characteristic called the Harmonic to Noise Rate (HNR) to enhance their emotion recognition system. This new characteristic was combined with other existing characteristics, including MFCC coefficients, ZCR, and TEO. To facilitate this integration, they employed an Auto-Encoder, a powerful dimension reduction technique. To evaluate the effectiveness of their system, they utilized the RML emotion database, a widely used benchmark in emotion recognition research. The results obtained from their approach demonstrated remarkable success when compared to other existing systems for emotion recognition. Additionally, their findings indicated that the application of the auto-encoder dimension reduction technique significantly improved the identification rate. Overall, in that study showcased the potential of incorporating the Harmonic to Noise Rate, in conjunction with other characteristics and the utilization of an Auto-Encoder, to achieve notable advancements in the field of SER. Their research serves as a significant contribution of value to the ongoing efforts in developing robust and accurate systems for emotion analysis and understanding.

Building on Zhang & Xue (2021), the researchers evaluated his approach using two widely accepted datasets, namely IEMOCAP and EMODB. The research findings demonstrated that the proposed model exhibited substantial performance improvement when compared to other methods employed in speech emotion recognition (SER). Additionally, the study revealed that the proposed algorithm led to a significant enhancement in classification accuracy. achieving a classification rate of 71% for the IEMOCAP dataset and 95.6% for his HE on the EMODB dataset. These results highlight the potential for improving the accuracy of spoken emotion recognition systems that use both latent expressions and acoustic features.

The Combination of Deep Learning Technique

According to Tarunika et al., (2018) the researchers focused on the application of DNNs and k-NN techniques for recognizing emotional content derived from speech, specifically in the context of a scary state of mind. They employed deep learning techniques and utilized their own sound database to train and test their models. The results presented in their paper provided a substantial and valuable contribution to the field of palliative care systems, providing insights and advancements in the recognition of emotions, which can aid in improving the well-being of individuals in need of palliative care.

Pandey et al. (2019) conducted a study examining different deep learning techniques for detecting and classifying emotional states from spoken language. That study use CNNs and LSTMs to capture emotion using different standard representations of speech, such as mel and magnitude spectrogram, and MFCC make evaluated the effect. To conduct the experiments, researchers used the EMODB and IEMOCAP datasets. As a result, we find that the CNN+LSTM architecture gives good results using the log-mel spectrogram.

Yao et al., (2020) conduct a study that aimed to develop a comprehensive framework by integrating three distinct classifiers, namely DNNs, CNNs, and RNNS. To accomplish this, they utilized the IEMOCAP databases as a valuable resource. The incorporation of a weighted pooling method in the neural networks, utilizing an attention mechanism, allowed for a focus on emotionally significant segments within the data. As a result, this approach led to higher accuracies in performing the given task. This innovative framework holds promising potential for enhancing the understanding and analysis of emotional content in various applications.

Datasets For SER

To conduct speech emotion recognition (SER), like other deep learning tasks, a well-constructed training dataset is crucial. This involves the manual annotation of samples by human agents, but due to the subjective nature of perceiving emotions, there can be variations in how individuals interpret and label emotional voices. For instance, what one person may perceive as anger, another might identify as excitement. To eliminate this ambiguity, we need to implement a reliable system that allows multiple personnel to review and label samples and confidently determine the appropriate label for each sample. There are three main types of databases that designed for voice emotion detection, including a collection of simulated, semi-natural, and natural languages. A simulated dataset is generated by having trained speakers read the same text with different emotional expressions. Naturalistic collections ask individuals to see the scenarios with different emotions. In contrast, the natural datasets are extracted from sources such as television stations, YouTube videos, and call centers and emotionally annotated by human listeners (Douglas-cowie et al., 2000).

Table 3.1 The classifications of databases that containing emotional speech

No.	Database	Language	Emotion
1.	EMO-DB Dataset	Use German Language	1. Sadness 2. Happiness 3. Boredom 4. Disgust 5. Anger 6. Neutral
2.	Danish Emotional Speech Databases	Use Danish Language	1. Sadness 2. Anger 3. Neutral 4. Surprise 5. joy

3.	RAVDESS Dataset	Use English Language	1. Sad 2. Happy 3. Angry 4. Surprised 5. Fearful 6. Calm 7. Disgusted 8. Neutral
4.	TESS Dataset	Use English Language	1. Pleasantly 2. Angry 3. Surprised 4. Happy 5. Fearful 6. Disgusted 7. Sad 8. Neutral
5.	CREMA-Dataset	Use English Language	1. Sadness 2. Happiness 3. Anger, 4. Disgust 5. Neutral 6. Fear
6.	IEMOCAP Dataset	Use English Language	1. Happiness 2. Anger, 3. Sadness 4. Frustation

From some of the datasets above, there are some of the datasets that have same emotions, so it can see the comparison in the fig.2 below:

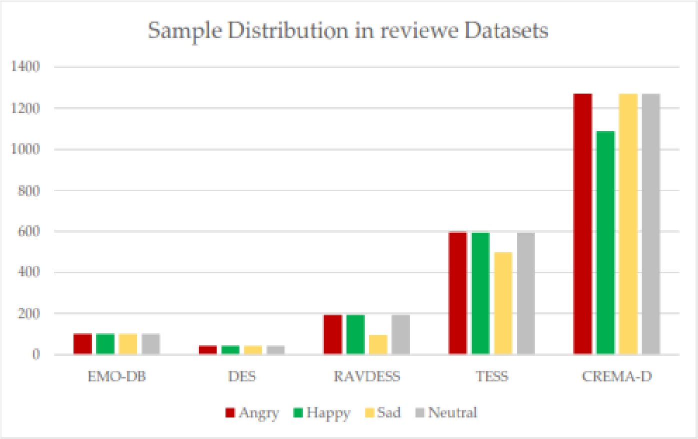

Fig. 2. Comparison the distribution of comparable emotions across the analyzed simulated datasets
Source : (Abbaschian et al., 2021)

In the overview of the evaluated algorithms, the paragraph summarizes the maximum accuracy achieved for each dataset, the encompassed features utilized for system training, the techniques employed, and, when relevant, the layer count in each approach are listed in the table below:

Table 3.2 A comparison result of literature review (Abbaschian et al., 2021)

Research Title	Methodology and Number of Layers	Features	Dataset and Accuracy
Article by Zhao et al. "Speech Emotion Recognition Using Deep 1D and 2D CNN LSTM Networks" (2019) focuses on deep learning models, especially the application of his 1D and 2D CNN LSTM networks to speech emotion recognition tasks.	This paper proposes using DCNN with sequence length 5 and LSTM networks for the task at hand.	The paper explores the use of Pulse Code Modulation (PCM) and Log-Mel Spectrogram representations for the analysis of speech data.	In this study, the EMO-DB dataset achieved 95.33% accuracy and the IEMOCAP dataset achieved 86.16% accuracy.
An article by Xie et al. titled "Speech Emotion Classification Using Attention-Based LSTMs" (2019) focuses on the application of attention-based long-short-term memory (LSTM) models to speech emotion classification tasks.	LSTM, DNN/5	The study examined various acoustic features for speech analysis, including MFCC and Zero Crosing	In this study, the eENTERFACE dataset achieved 89.6% accuracy and the GEMEP dataset achieved 57% accuracy. On the other hand, the CASIA dataset achieved 92.8% accuracy.
Article by Chatziagapi et al. The book entitled Data Augmentation Using His GANs for Speech Emotion Recognition (2019) focuses on the application of generative adversarial networks (GANs) for data augmentation in the context of speech emotion recognition.	This paper uses deep convolutional neural network (DCNN) architectures, specifically VGG19 model and GAN/19, for the present task.	The feature extraction process involves extracting 128 MFCCs	In this study, the IEMOCAP dataset achieved 53.6% accuracy and the Feel-25k dataset achieved 54.6% accuracy.
Article by Sahu et al. "Enhancing Speech Emotion Recognition Using Generative Adversarial Networks" (2018) explores the use of generative adversarial networks (GANs) to improve the performance of speech emotion recognition systems.	In the current task, the research uses a combination of Generative Adversarial Networks (GANs) and SVMs	This work uses the 1582-dimensional openSMILE feature space.	In this study, the IEMOCAP dataset achieved an accuracy of 60.29%.

Article titled "Adversarial Machine Learning and Speech Emotion Recognition": "Using Generative Adversarial Networks for Robustness" Latif et al. (2018) describes the application of generative adversarial networks (GANs) to improve the robustness of speech emotion recognition systems in the context of adversarial machine learning. are investigating.	The paper suggests the utilization of LSTM networks and GANs with a factor of 2 for the task at hand.	The paper focuses on the use of eGeMAPS features for speech analysis.	In this study, the aibo dataset achieved an accuracy rate of 64.86%, while the IEMOCAP dataset achieved an accuracy rate of 53.76%.
Article entitled "Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition" by Eskimez et al. (2018) consider an unsupervised learning approach to feature analysis in the context of automatic speech emotion recognition.	The paper proposes use dof Convolutional Neural Networks (CNN) in combination with Variational Autoencoders (VAE) with different configurations, including sequence lengths of 5, 6, 4, 10, and 5, for the task at hand.	Studies analyzing speech data use logmel spectrograms.	In this study, the IEMOCAP dataset achieved an accuracy of 48.54%.
Article titled "A Variable Autoencoder for Learning Speech Emotion Latent Expressions": A preliminary study by Latif et al. (2017) explores the application of variational autoencoders (VAEs) to learning latent representations of vocal emotion.	The paper investigates the utilization of Variational Autoencoders (VAEs) in combination with two layers of LSTM networks with a sequence length of 4 for the task at hand.	The Log-Mel Spectrogram representation is utilized in the study for analyzing speech data.	In this study, the IEMOCAP dataset achieved an accuracy of 64.93%.

FINDINGS AND DISCUSSION

This study details the emphasis on the importance of HCI in the sector of emotion recognition from speech. We carefully examine the various applications in which voice emotion detection plays a central role, and discuss the complex challenges encountered in implementing it. This paper then develops a set of overarching goals, including investigating deep learning techniques for detecting emotion from speech, performing analyses, carefully evaluating their performance, and identifying their inherent limitations and challenges is focused on to achieve these goals, this methodology includes an extensive literature analysis of relevant papers to obtain a comprehensive understanding. This paper provides an overview and evaluation about the current state of deep learning technology that used in SER field.

Deep learning techniques are very promising for SER because they can automatically identify complex patterns in audio data without requiring manual feature extraction or extensive parameter tuning. Among these techniques, DNNs, including CNNs, RNNs, and their diverse variants, have showcased notable effectiveness not only in image and video processing but also in speech-based classification tasks. Nonetheless, it is crucial to acknowledge that these models are not devoid of limitations. For instance, CNNs tend to learn features from minor variations and distortions, which can pose challenges in certain scenarios, while RNNs based on LSTM architecture require substantial storage capacity to be effectively employed.

The paper meticulously expounds upon the fundamental principles underpinning deep learning algorithms, with a particular emphasis on RNNs and CNNs, accentuating their distinct advantages within the field of speech emotion recognition. RNNs, augmented by LSTM cells, are adept at accessing and effectively utilizing long-term memory, while CNNs excel in extracting salient features from intricate and complex input data. Numerous studies of notable significance are cited within the paper, underscoring the substantial performance improvements achieved through the utilization of deep learning algorithms within the realm of SER.

Overall, the paper underscores the remarkable promise showcased by deep learning algorithms, specifically RNNs with LSTM architecture and CNNs, in the realm of SER. Incorporating attentional mechanisms is believed to be a highly effective approach to improve speech emotion recognition system performance. A multitude of studies have unequivocally demonstrated the efficacy of deep learning models in accurately recognizing emotions from speech, prompting the development of benchmark frameworks meticulously designed to comprehensively evaluate their performance and generalization capabilities.

CONCLUSION

This article provides a comprehensive analysis of deep learning techniques used in SER. Deep learning such as DNNs, RNNs, CNNs, AEs, and their combinations have been extensively explored in recent work. These techniques are briefly described in terms of their ability to categorize different natural emotions such as sadness, happiness, neutrality, anger, disgust, fear and boredom. These methods use shared weights to make the model easier to train and more efficient. However, deep learning techniques have certain limitations such as their complex internal architecture consisting of multiple layers, reduced efficiency in handling input data that changes over time, and the risk of over-learning by memorizing layer-wise information. The complex internal architecture of deep learning models, such as DNNs and RNNs, presents challenges in terms of training, fine-tuning, and optimization. The multiple layers and interconnected nodes require careful hyperparameter tuning, extensive computational resources, and longer training times.

Managing and optimizing these complex architectures can be a demanding task. These models operate under the assumption that the training and test data exhibit a comparable distribution. However, significant changes in the input data distribution, such as the introduction of new classes or variations in input characteristics, can hinder the models' ability to adapt effectively. Consequently, deep learning models may struggle to generalize well to handle novel or unseen instances, resulting in reduced efficiency and performance degradation. To overcome this limitation, the model should be fine-tuned or retrained with updated data.

This study serves as a basis for evaluating the effectiveness and limitations of current deep learning techniques and identifying potential opportunities for improving SER systems. The field of speech emotion recognition has wider implications in areas such as human-computer interaction, affective computing, psychology, and healthcare. It enables natural interactions, personalized interventions, and improved emotional well-being assessment. Future research directions include multimodal fusion, transfer learning, explain ability, modelling long-term dynamics, real-world applications, and ethical considerations. These directions aim to integrate multiple modalities, address domain shift, enhance transparency, capture temporal dynamics, validate practical utility, and ensure responsible development. Advancing in these areas will improve system performance and maximize the potential of SER in diverse applications.

REFERENCES

Abbaschian, B.J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. In *Sensors (Switzerland),* 21(4), 1-27. MDPI AG. https://doi.org/10.3390/s21041249

Akçay, M.B. & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*(June 2019), 56–76. https://doi.org/10.1016/j.specom.2019.12.001

Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors (Switzerland)*, *20*(18), 1–16. https://doi.org/10.3390/s20185212

Aouani, H. & Ayed, Y.B. (2020). Speech Emotion Recognition with deep learning. *Procedia Computer Science*, *176*, 251–260. https://doi.org/10.1016/j.procs.2020.08.027

Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., & Kollias, S. (2005). Emotion analysis in man-machine interaction systems. *Lecture Notes in Computer Science*, 3361, 318–328. https://doi.org/10.1007/978-3-540-30568-2_27

Bertero, D. & Fung, P. (2017). A first look into a convolutional neural network for speech emotion detection. In Proceedings of the IEEE *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) (pp. 5115-5119). Clear Water. http://ieeexplore.ieee.org/document/7953131/

Byun, S.W. & Lee, S.P. (2021). A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences (Switzerland)*, *11*(4), 1–15. https://doi.org/10.3390/app11041890

Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., & Narayanan, S. (2019). Data augmentation using GANs for speech emotion recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2019-Septe*, 171–175. https://doi.org/10.21437/Interspeech.2019-2561

Chen, M., Zhou, P., & Fortino, G. (2017). Emotion Communication System. *IEEE Access*, *5*, 326–337. https://doi.org/10.1109/ACCESS.2016.2641480

Douglas-cowie, E., Cowie, R., & Schröder, M. (2000). A New Emotion Database: Considerations, Sources and Scope. *In*, 39–44.

Eskimez, S. E., Duan, Z., & Heinzelman, W. (2018). Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In Proceedings of the 2018 IEEE *International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 5099-5103). http://ieeexplore.ieee.org/document/8462417/

Ho, N. H., Yang, H. J., Kim, S. H., & Lee, G. (2020). Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network. *IEEE Access*, *8*, 61672–61686. https://doi.org/10.1109/ACCESS.2020.2984368

Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Latif, S., Rana, R., & Qadir, J. (2018). *Adversarial Machine Learning And Speech Emotion Recognition: Utilizing Generative Adversarial Networks For Robustness*. 1–7. http://arxiv.org/abs/1811.11402

Latif, S., Rana, R., Qadir, J., & Epps, J. (2017). *Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study*. http://arxiv.org/abs/1712.08708

Lee, J. & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-Janua*, 1537–1540. https://doi.org/10.21437/interspeech.2015-336

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics (Switzerland),* 10(10). https://doi.org/10.3390/electronics10101163

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, 7, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Pandey, S.K., Shekhawat, H.S., & Prasanna, S.R.M. (2019). Deep learning techniques for speech emotion recognition: A review. *2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA 2019 - Microwave and Radio Electronics Week, MAREW* 2019. https://doi.org/10.1109/RADIOELEK.2019.8733432

Petrushin, V.A. (2000). Emotion recognition in speech signal: Experimental study, development, and application. *6th International Conference on Spoken Language Processing, ICSLP* 2000, *Icslp*, 6–9. https://doi.org/10.21437/icslp.2000-791

Reynolds, D.A. (n.d.). *2002 Reynolds D - An overview of automatic speaker recognition.pdf*.

Sahu, S., Gupta, R., & Espy-Wilson, C. (2018). *On Enhancing Speech Emotion Recognition using Generative Adversarial Networks*. http://arxiv.org/abs/1806.06626

Scheidwasser-Clow, N., Kegler, M., Beckmann, P., & Cernak, M. (2022). Serab: a Multi-Lingual Benchmark for Speech Emotion Recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2022-May*, 7697–7701. https://doi.org/10.1109/ICASSP43922.2022.9747348

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1), 93–120. https://doi.org/10.1007/s10772-018-9491-z

Tarunika, K., Pradeeba, R.B., & Aruna, P. (2018, October 16). Applying Machine Learning Techniques for Speech Emotion Recognition. *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*. https://doi.org/10.1109/ICCCNT.2018.8494104

Vogt, T., André, E., & Wagner, J. (2008). Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4868 LNCS*, 75–91. https://doi.org/10.1007/978-3-540-85099-1_7

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech Emotion Classification Using Attention-Based LSTM. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 27(11), 1675–1685. https://doi.org/10.1109/TASLP.2019.2925934

Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Communication*, *120*, 11–19. https://doi.org/10.1016/j.specom.2020.03.005

Zhang, C. & Xue, L. (2021). Autoencoder with emotion embedding for speech emotion recognition. *IEEE Access*, *9*, 51231–51241. https://doi.org/10.1109/ACCESS.2021.3069818

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035