Emotion recognition using machine learning: Opportunities and challenges for supporting those with autism or depression

Xianshan Jin

Shanghai SMIC Private School, Shanghai, 201210, China

alicek6110@gmail.com

Abstract. Facial recognition technology, driven by advancements in machine learning, has become integral to various applications, from mobile security to access control. Recent developments in deep learning, particularly through convolutional and recurrent neural networks, have paved the way for enhanced emotion recognition capabilities. This paper explores the potential of facial emotion recognition (FER) and voice emotion recognition technologies, assessing their current state and future prospects. While FER techniques have evolved from traditional methods to deep learning approaches, and voice emotion recognition has shown promising results, challenges remain due to the inherent variability in human facial expressions and vocal characteristics. The paper further examines the impact of emotion recognition technologies on mental health, highlighting their potential to provide significant support to individuals with mental health disorders. By addressing the accuracy and application of these technologies, the research underscores the importance of continued innovation and careful implementation to maximize their benefits and mitigate potential risks. Ultimately, emotion recognition provides promising avenues for supporting those with mental illnesses. Aimed at fostering more personalized and effective interventions, these technologies hold the potential to revolutionize mental health care and improve the overall well-being of individuals.

Keywords: Machine learning, Emotional support, Emotion recognition, Mental illness, Mental support.

1. Introduction

It is now common for us to unlock our phones, pass through security checks, or enter a company simply by looking at a device. Continuous advancements in technology have significantly enhanced the quality of our lives, providing us with greater convenience and efficiency, and a substantial portion of these advancements can be attributed to the rapid development of machine learning techniques. Machine learning has significantly advanced facial recognition technologies, which are now widely used in mobile devices for identity verification and security purposes. The success and prevalence of face recognition in our society today prompted many to explore potential applications beyond facial recognition.

Ongoing developments in deep learning, particularly convolutional and recurrent neural networks, are driving more accurate detection and classification of human emotions, offering a promising direction for emotion recognition technologies.

Emotions are a universal aspect of human experience, universally present at all times and in all places. Unlike facial recognition, which relies on identifying specific features unique to each individual, emotion recognition delves deeper. It requires machines to accurately detect emotions based on facial expressions or tone of voice. This advancement not only enhances human-computer interactions but also holds promise for revolutionize fields such as mental health, security, and customer service through the provision of more sophisticated and compassionate reactions. This paper aims to highlight the significance of emotion recognition technology and its potential to create substantial societal effects.

This paper employs a systematic review of the current literature on facial emotion recognition (FER) and voice emotion recognition technologies. The research methodology involves a comprehensive analysis of existing models, techniques, and frameworks, with a focus on deep learning approaches such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The review is structured to evaluate the performance of these technologies across different datasets, assessing accuracy rates, generalizability, and real-world applicability. Additionally, this study examines the integration of emotion recognition technologies in mental health applications, highlighting the opportunities and challenges associated with their deployment. Through this multifaceted approach, the research aims to provide a detailed understanding of the current state of emotion recognition technologies and their potential impact on society.

2. Machine learning

Machine learning is a subset of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit instructions [1]. Instead of programming specific rules, these systems learn patterns from input data and improve their performance over time as they process more information [1].

2.1. Definition of supervised and unsupervised learning

Supervised learning is a type of machine learning where the goal is to learn a function that maps inputs to corresponding outputs using example input-output pairs [2]. This method involves inferring a function from labeled training data, where each training example includes both the input data and the desired output [2]. In supervised learning, algorithms require external guidance, which means that the input data is divided into a training set and a test set. The training set contains the output labels that need to be predicted or classified. Through this process, the algorithms acquire knowledge of patterns from the training data, which they subsequently utilize to make predictions or classifications on the test data [2].

Unsupervised learning, on the other hand, involves using algorithms to analyze and cluster data sets that lack labels [3]. Unlike supervised learning, these algorithms operate without predefined output labels, instead independently identifying hidden patterns or natural groupings within the data , without any human intervention [3].

2.2. Supervised and unsupervised emotional recognition

In supervised emotion recognition, algorithms are trained on labeled datasets where each input (such as an image of a face or a segment of speech) is associated with a specific emotion label (e.g., happiness, sadness, anger). The model learns to map these inputs to the corresponding emotions, and once trained, it can predict the emotion in new, unseen data. This approach is commonly used in applications where the goal is to accurately classify or predict specific emotional states.

Unsupervised learning can be used in emotion recognition to discover patterns or groupings in data without predefined labels. For example, an unsupervised approach might cluster similar facial expressions or vocal tones together, potentially identifying different emotional states without prior labeling. This can be valuable for exploratory analysis, especially in cases where labeled data is limited.

3. The latest research on emotion recognition

3.1. Facial Emotion Recognition (FER)

In our daily communications, linguistic cues such as spoken words convey only 7% of the information [4]. Paralinguistic cues, including vocal elements like tone, pitch, and volume, account for 38% of the information conveyed [4]. The remaining 55% of the information comes from non-verbal cues, primarily facial expressions [4]. This highlights the important role that facial expressions play in real-life conversations. A single facial expression can greatly alter the the intended meaning of a message, underscoring the importance of non-verbal communication in understanding and interpreting messages accurately.

Driven by the advancement of artificial intelligence, detecting the emotion one possesses through machine learning has been studied and researched. FER research can be divided into two categories: approaches using handcrafted features and those using deep neural networks. Traditional FER methods involve three steps: detecting faces and facial components, extracting features, and classifying expressions. In contrast, deep learning approaches, particularly using convolutional neural networks (CNNs), enable "end-to-end" learning directly from input images, reducing reliance on pre-processing techniques and yielding state-of-the-art results [5].

FER can also be divided into static (frame-based) and dynamic (video-based) approaches. Static FER uses single frames to analyze expressions, while dynamic FER captures expression changes over time, offering higher recognition rates but facing challenges with temporal normalization [5].

There have been several research and models developed on FER for the past decade. One of the earliest research done in 2012 by Aung D.M, the model classified facial expression based on feature vectors. The four main tasks of the model were: pre-processing of images, segmentation of mouths, extraction of features, and histogram-based categorization. This model can identify five different human emotions: neutral, surprise, anger, sorrow, and happiness, with an average accuracy rate of 81.6% [6].

Another research done in 2015 by Perikos, I. and his team developed a hybrid two-stage classification schema. In this model, a Multilayer Perceptron Neural Network determines each expression's emotional content based on Ekman's emotional categories after SVM is used to determine if the expressions carry emotional content or are neutral. Their performance is reported by the system to be 85%, which is higher than the model proposed in 2012 by another researcher [7].

3.2. Voice emotion recognition

Just like facial expression, the complexity of the human voice allows people to change the meaning of words simply by altering the tone of their voice. The main characteristics for machines to catch and therefore detect emotions are mainly by pitch, rate, intensity, and articulation [8]. A table is shown below of the parameters each emotion possesses.

Table 1. Key Voice Features of the Five Distinct Emotions [8]

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Rate	Slightly faster	Faster or slower	Slightly slower	Must faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly narrower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

In 2017, Eduard Frant and his group developed a CNN model and experimented with it. The process begins with preprocessing the audio data using PRAAT software to extract Mel-Frequency Cepstral Coefficients (MFCCs) [9]. These coefficients are structured into a 2D array, resembling an image, to be fed into the CNN. The CNN architecture includes: a convolutional layer with 200 filters of size 5x5, activated by ReLU; a max-pooling layer; a flattening layer; a fully connected layer with 1000 neurons; an output layer with 6 neurons for classifying emotions [9].

The network was implemented in Python using TensorFlow and Keras libraries, designed to handle input dimensions suitable for emotion classification from voice data.

The CNN was trained using a dataset of voice recordings from 30 Romanian speakers, each recording being 5 seconds long. The recordings were divided into six emotion categories: happiness, fear, sadness, disgust, anger, and surprise. The model was trained over 25 epochs with the dataset containing a balanced distribution of these emotions.

The trained model was evaluated on a test set of 30 voice samples, achieving a mean accuracy of 71.37%. This performance is comparable to other speech recognition methods. The experimental results showed varying accuracy for different emotions, with happiness at 71%, fear at 75%, and disgust at 67% [9].

4. Support for disordered individuals

4.1. Prevalence of mental health

The main purpose of FER (facial emotion recognition) is to the support we can give to individuals with a disability. As society progresses, the prevalence of mental health issues is becoming more and more serious regardless of sex, gender, and ethnicity. Mental health illness can be a very broad term, covering anxiety disorder, autism, depression, etc [10].Below is a study on mental illness conducted in 2021 by the National Institute of Mental Health [11].

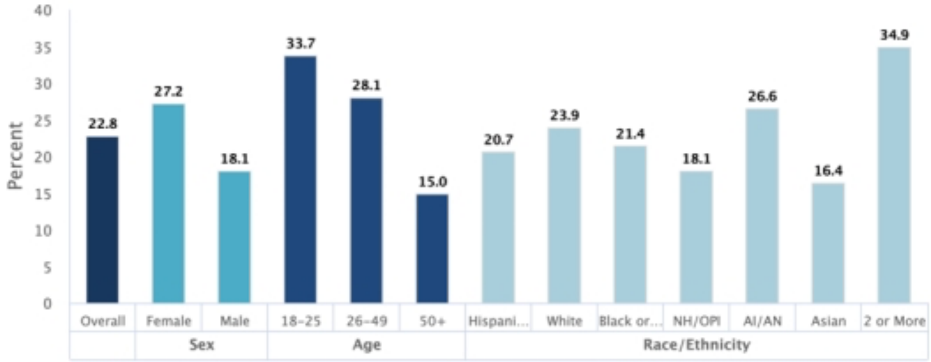

Figure 1. 2021 Prevalence of Mental Illness Among U.S. Adults [10]

Based on the figure above, we see a trend that younger generations (age 18-25) seemed to possess the greatest prevalence for mental illness with a 33.7%. Still, overall, 22.5% or one in every five U.S. adults have issues with their mental health.

Despite these high rates of mental health, the most up-to-date treatment is the usage of medicine as well as mental therapy. However, the rise of recent research on FER should be mainly aimed at the support for these individuals.

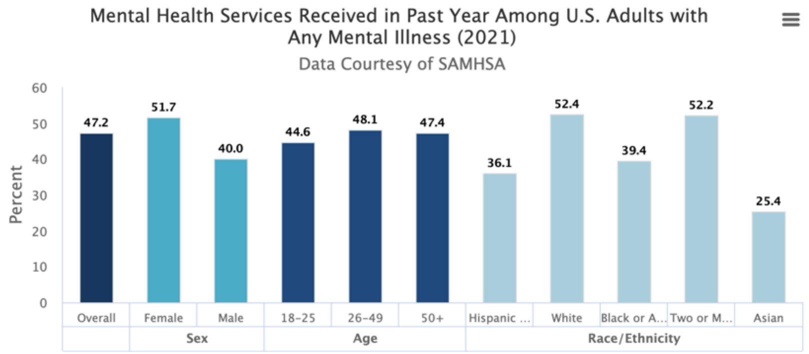

Figure 2. Mental Health Services Received Among U.S. Adults with Mental Illness in 2021 [10]

Based on the statistics on top, only around half of the mental health patients received treatments. This may be due to various factors such as convenience, cost, or other unknown reasons. In order to lower the rates in Figure 1 and increase the rate in Figure 2, implementing such emotion recognition into our technology today would help mitigate these data.

4.2. Expressing emotions

Individuals experiencing hardships or illness may express their emotions in various ways. Despite the advances in current technology, talking to an AI assistant often falls short of providing meaningful help. AI assistants today are not yet capable of accurately detecting human emotions, which confines their ability to offer emotional support effectively.

Robert Plutchik, an American psychologist, developed a widely recognized classification of fundamental emotions [12]. He identified eight primary emotions: joy, trust, fear, anticipation, sadness, disgust, anger, and surprise [12].

For individuals with mental illness, any of these emotions can manifest at any time. Emotion recognition models aim to identify the specific emotion an individual is experiencing and estimate its intensity. By systematically identifying and evaluating these emotions, AI systems can be designed to respond with appropriate emotional support, potentially offering a significant benefit to those in need.

5. Potential opportunities or challenges

One of the most significant challenges in the realm of emotion detection through facial recognition technology lies in the inherent variability of human faces. Despite the universal presence of fundamental facial features—such as the eyes, nose, and mouth—these features exhibit considerable diversity across individuals. This variability is influenced by a range of factors including age, gender, sex, ethnicity, and individual physiological differences. Additionally, facial expressions, which are crucial for interpreting emotions, can be uniquely expressed by each person even when experiencing the same emotional state. This diversity introduces considerable complexity for facial recognition systems, making it challenging for these systems to accurately detect and interpret emotional states.

Current research in this domain should primarily focus on enhancing the accuracy of emotion detection models. Given the profound implications of these technologies, achieving high accuracy is crucial. Researchers and developers should strive to reach or exceed a 90% accuracy threshold before these systems are widely implemented. This level of accuracy is critical not only to ensure the reliability of the technology but also to mitigate potential risks associated with misinterpretation of emotional cues. Inaccurate emotion detection could lead to erroneous conclusions, which in turn could affect decision-making processes and interpersonal interactions.

Despite these challenges, the potential benefits of successful emotion detection technology are substantial. Advanced emotion recognition systems could serve as valuable tools for aiding individuals with mental health conditions by providing more accurate assessments of their emotional states. Such technology could offer new insights for mental health professionals, facilitate early detection of emotional disturbances, and support the development of personalized treatment plans. The integration of emotion detection technology into therapeutic settings could enhance the efficacy of interventions and contribute to improved patient outcomes.

6. Conclusion

In conclusion, emotion recognition technologies, including facial and voice-based systems, represent a significant leap forward in enhancing human-computer interactions and providing support to individuals with mental health conditions. While the progress in these technologies is promising, achieving high accuracy remains a crucial challenge. The variability in facial expressions and vocal characteristics across different individuals poses a significant hurdle for accurate emotion detection. Continued research and development are essential to address these challenges and improve the reliability of these systems. The potential benefits of advanced emotion recognition technologies are substantial, particularly in the realm of mental health, where they could offer valuable insights and support for personalized treatment plans. As these technologies evolve, careful consideration of their implementation and accuracy will be critical in ensuring their positive impact and mitigating risks associated with misinterpretation.

While the technology for emotion recognition is still in its developmental stages, there is no denying its potential to significantly contribute to mental health care. Within the next decade, there is the total possibility of its potential doubling or even tripling. With emotion recognition implemented into our technology just like the facial recognition now, it will undoubtedly benefit those with mental illness. Going further, these technologies have the capacity to not only improve individual outcomes but also to foster a deeper understanding of human emotions in various societal contexts.

References

[1] GeeksforGeeks. “Machine Learning Algorithms.” GeeksforGeeks, 17 Aug. 2023, www. geeksforgeeks.org/machine-learning-algorithms/. Accessed 14 July 2024.

[2] Mahesh, Batta. “Machine Learning Algorithms -A Review.” International Journal of Science and Research (IJSR) ResearchGate Impact Factor, vol. 9, no. 1, 2018, www.ijsr.net/archive/v9i1/ ART20203995.pdf, https://doi.org/10.21275/ART20203995. Accessed 8 Aug. 2024.

[3] IBM. “What Is Unsupervised Learning? | IBM.” Www.ibm.com, 2023, www.ibm.com/topics/ unsupervised-learning. Accessed 8 Aug. 2024.

[4] Mehrabian, A.: Communication without words. Psychol. Today 2(4), 53–56 (1968)

[5] Giannopoulos, Panagiotis, et al. “Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013.” Advances in Hybridization of Intelligent Methods, 15 Oct. 2017, pp. 1–16, https://doi.org/10.1007/978-3-319-66790-4_1.

[6] Aung, D.M., Aye, N.A.: Facial expression classification using histogram based method. In: International Conference on Signal Processing Systems (2012)

[7] Perikos, I., Ziakopoulos, E., & Hatzilygeroudis, I.: Recognize emotions from facial expressions using a SVM and neural network schema. In: Engineering Applications of Neural Networks, pp. 265–274. Springer International Publishing, (2015)

[8] Murray and Arnott, 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. Journal of the Acoustical Society of America. v93 i2. 1097-1108.

[9] Franti, Eduard, et al. “Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots.” ROMANIAN JOURNAL of INFORMATION SCIENCE and TECHNOLOGY, vol. 20, 3 Nov. 2017.

[10] National Institute of Mental Health. “Mental Illness.” National Institute of Mental Health, Mar. 2023, www.nimh.nih.gov/health/statistics/mental-illness. Accessed 16 July 2024.

[11] National Institute of Mental Health. “NIMH» Health Topics.” Www.nimh.nih.gov, 2024, www. nimh.nih.gov/health/topics. Accessed 19 July 2024.

[12] Plutchik, Robert, The nature of emotions, American Scientist 89 (2001), page 344.