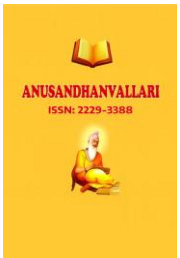

Anusandhanvallari
Vol 2026, No.1
July 2026
ISSN 2229-3388

Deep Learning Approaches for Mental Health Detection Using Facial Emotion Recognition: A Systematic Review

Priyanka Kamble[1], Nandini Chaudhari[2]

[1]Faculty of Engineering and Technology, Drs. Kiran and Pallavi Patel Global University, Vadodara, INDIA

[2]Department of Computer Science and Engineering, KSET, Drs. Kiran and Pallavi Patel Global University, Gujarat, INDIA

Abstract: Depression, anxiety, stress have been on a rise across the world, and the need for early, accurate, and scalable diagnosis solutions is growing. The conventional assessment techniques are subjective, time-consuming and rely on clinical experience. Facial Emotion Recognition (FER) has been a promising non-invasive method for identifying mental health disorders using AI technology of facial expression analysis in recent years. This paper systematically reviews 50 research papers published from 2021 to 2024, which specifically concentrate on mental health detection based on FER using deep learning. The review covers different deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and hybrid CNN-LSTM models. It also summarizes recent research on preprocessing methods, datasets, evaluation metrics, and multimodal learning methods. The results show that CNN-LSTM and attention based models have higher accuracy, by learning both spatial and temporal emotional features each. Moreover, multimodal integration of speech, physiological signals and facial expression contributes to the robustness and reliability of the system. Even with these progressions, there are still a number of hurdles to overcome, such as data imbalance, cultural bias, overfitting, and few data validations in real-world scenarios. Further, the deployment of real-time applications is limited by the computational complexity and by the need of lightweight models. In this review, we point out some major research gaps and suggest future directions, including explainable AI models, diverse and clinically validated datasets, and efficient architectures for real-time applications. The study is overall useful to researchers and practitioners for creating reliable and scalable systems for detecting mental health problems by using FER.

Keywords: Facial Emotion Recognition,Mental Health Detection,Deep Learning,Multimodal Learning,CNN-LSTM.

I. Introduction

Mental health disorders such as stress, anxiety, and depression are increasing globally and have become a major public health concern [23], [24]. Conventional methods of diagnosing such disorders are usually based on interviews and self-reports, which may prove to be slow and extremely subjective. But as artificial intelligence is advancing rapidly, in particular, machine learning there is a move towards more objective data-driven ways of comprehending mental health. Facial emotion recognition (FER) is one of the prospective fields where computers can use facial expressions to understand the emotional condition of a person [1],[9],[10]. New algorithms of deep neural network, such as Convolutional Neural Networks (CNNs) and hybrid models such as CNN-LSTM, can now detect even the minor emotional expressions which may be connected to mental illnesses [7], [18], [21]. The following paper will address recent developments in the use of FER to identify psychological disorders, explain the most popular techniques and datasets, and discuss the issues that researchers encounter when applying the said solutions in practice. Mental health disorders, such as depression, anxiety, and stress, affect millions globally, requiring early detection and intervention for better outcomes. Traditional diagnostic methods are often subjective and lack real-time capabilities, driving the need for objective, scalable solutions [34],[36]. Recent advances in deep learning, particularly facial emotion recognition using models like CNNs and RNNs, offer promising alternatives by detecting mental health conditions through non-verbal cues[1],[4],[46].

II. Theoretical Reference

This survey reviews recent studies on FER-based mental health detection published between 2021 and 2024. This survey looks into 50 recent research articles to show the most recent methods employed in the field. Other typical issues that we discuss include unavailability of various datasets and how culture affects the analysis of facial expressions. We also discuss the possibilities of implementing these models in real-life, such as telemedicine and mobile health [36], [41]. We will measure the effectiveness of AI-based facial emotion recognition in identifying mental health conditions and direct researchers to the new directions of creating more reliable, accessible, and user-friendly tools in the future.

Table 1 can be used to summarize our survey in comparison to past reviews in mental health detection. It does not just point out the uniqueness of the work but also points at its weaknesses and original contributions.

Table 1: Comparative Analysis of Existing Studies on Mental Health Detection Using Facial Emotion Recognition.

Paper Title & Reference	Major Limitation	Contribution/Findings
Deep facial expression recognition: A survey [1]	Limited real-time implementation discussion	Comprehensive review of FER techniques
Going deeper in facial expression recognition using deep neural networks [2]	High computational complexity	Improved FER accuracy using deep CNNs
A survey of facial emotion recognition for mental health applications [3]	Limited clinical validation	Summarized FER applications in mental health
Deep-emotion: Facial expression recognition using attentional convolutional network [4]	Real-time deployment difficulty	Improved subtle emotion recognition
Facial expression recognition by de-expression residue learning [5]	Performance variation on real-world data	Enhanced expression learning performance
Facial expression recognition using transfer learning [6]	Overfitting on small datasets	Reduced training complexity
Generative adversarial nets [7]	Synthetic data quality dependency	Improved data augmentation capability
OpenFace: An open source facial behavior analysis toolkit [8]	Landmark sensitivity issues	Real-time facial behavior analysis
Affectiva-MIT facial expression dataset (AM-FED) [9]	Limited annotation consistency	Real-world FER dataset
Static facial expression analysis in tough conditions [10]	Low robustness in uncontrolled conditions	Standardized FER benchmarking

AffectNet: A database for facial expression, valence, and arousal computing in the wild [11]	Dataset imbalance	Large-scale real-world FER dataset
The extended Cohn-Kanade dataset (CK+) [12]	Small sample size	Benchmark FER dataset
Coding facial expressions with Gabor wavelets [13]	Computational overhead	Improved facial feature extraction
Reliable crowdsourcing and deep locality-preserving learning for expression recognition [14]	Label inconsistency	Improved FER in-the-wild
Deep residual learning for image recognition [15]	Requires high computational resources	High-performance deep feature extraction
EfficientNet: Rethinking model scaling for convolutional neural networks [16]	Complex architecture tuning	Lightweight and efficient model
Very deep convolutional networks for large-scale image recognition [17]	Large memory requirement	Improved deep representation learning
Long short-term memory [18]	Training complexity	Temporal sequence learning
Supervised sequence labelling with recurrent neural networks [19]	Vanishing gradient issues	Sequential emotional analysis
Attention is all you need [20]	High training requirements	Advanced attention-based learning
Neural machine translation by jointly learning to align and translate [21]	Computational complexity	Attention alignment learning
Joint face detection and alignment using multitask cascaded convolutional networks [22]	Sensitive to occlusions	Accurate face alignment
Affective Computing [23]	Limited automation at early stage	Foundation of affective computing
A review and meta-analysis of multimodal affect detection systems [24]	Integration complexity	Robust multimodal emotion analysis
A survey of affect recognition methods [25]	Limited real-time systems	Comprehensive affect recognition review
Occlusion aware facial expression recognition using CNN with attention mechanism [26]	Occlusion sensitivity	Improved FER under occlusion

ImageNet large scale visual recognition challenge [27]	High resource requirements	Benchmark dataset for deep learning
ImageNet classification with deep convolutional neural networks [28]	Overfitting risk	Revolutionized deep CNN learning
Going deeper with convolutions [29]	Complex architecture	Improved feature abstraction
Densely connected convolutional networks [30]	Increased memory usage	Enhanced feature propagation
Gradient-based learning applied to document recognition [31]	Early CNN limitations	Foundation of CNN architectures
FaceNet: A unified embedding for face recognition and clustering [32]	Large training data required	Accurate facial embedding generation
Multiresolution gray-scale and rotation invariant texture classification [33]	Texture dependency	Robust texture feature extraction
Rapid object detection using a boosted cascade of simple features [34]	Reduced robustness for complex scenes	Real-time face detection
Dlib-ml: A machine learning toolkit [35]	Landmark precision limitations	Real-time facial landmark detection
Faster R-CNN: Towards real-time object detection [36]	High computational demand	Improved real-time detection
A multimodal depression detection method based on facial expression and speech features [37]	Complex multimodal synchronization	Improved depression detection accuracy
Recognition of genuine smiles [38]	Limited spontaneous expression datasets	Genuine vs fake smile analysis
A review of facial expression recognition in the wild [39]	Environmental variability	Real-world FER analysis
Automatic facial expression recognition using features of salient facial patches [40]	Sensitivity to illumination	Improved local facial feature learning
A model of the perception of facial expressions of emotion by humans [41]	Human variability	Psychological basis of FER
Automated face analysis for affective computing [42]	Limited scalability	FER for affective computing systems

Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition [43]	Sensor integration complexity	Comprehensive multimodal FER review
Convolutional neural networks for sentence classification [44]	NLP-focused limitations	CNN representation learning
Universal language model fine-tuning for text classification [45]	Fine-tuning instability	Improved transfer learning techniques
Multimodal deep learning for mental health recognition [46]	Complex data fusion	High-performance mental health recognition
Deep learning for mental health prediction based on facial expressions [47]	Dataset bias	FER-based mental health prediction
Automatic analysis of facial expressions: The state of the art [48]	Early-stage FER limitations	Foundational FER review
Emotion recognition in the wild via convolutional neural networks [49]	Performance variation in uncontrolled environments	Real-world FER improvement
Automatic recognition of facial actions in spontaneous expressions [50]	Spontaneous expression complexity	Automatic action unit recognition

Table 1 summarizes the limitations and contributions of existing studies related to FER-based mental health detection.

II.1 Contributions

The key findings of this survey are as follows:

- Sealing Past Fissures**:** This is because, by beginning with an analysis of past studies, we attempt to fill the gaps that have been left by previous research in the field of mental health detection, expanding upon it and rendering it more recent and detailed.
- Survey of the Recent Research: The present paper is a systematic review of the articles published in 20212024, which presents the recent developments in the given area [1], [3], [39], [43], [46].
- Comparison of Key Methods and Datasets: We are compiling and comparing key measures and datasets of mental health diagnosis. This section allows researchers to easily locate appropriate data sets, measurement of evaluation, and techniques of cross-validation that may be required based on their research requirements [1], [3], [39], [47], [50].
- Correlation: The survey will also determine the performance of various mental health detection models on the same datasets under the same evaluation criteria. With the help of this comparison, the best results are easier to identify in which models [15], [16], [18], [20], [46].
- Determining Challenges and Future Directions: In the end, we define the key challenges in the field and suggest potential resolutions. We also talk about new trends and refer to potential future research directions [3], [24], [39], [43], [46].

The remainder of this survey will be structured as follows: Section 2 captures our framework proposed. Section 3 includes an in-depth review of deep learning-based approach to diagnosing mental health conditions, as well as applicable theoretical and mathematical understanding of the models utilized. We have discussed the different performance evaluation measures, the evaluation data sets and the evaluation methods proposed. We have shown the findings of the relative analysis of the approaches that made use of the same dataset, evaluation measures, and cross-validation strategies. Section 4 also gives major research challenges and issues that are related to diagnosing Mental Health and future direction of research given the reviewed literature. In the section 5 we have explained the findings and limitation of this study in detail. Section 6 is a conclusion of the survey [1], [3], [24], [39], [46].

III. Methods

III.1 Searching Criteria

We conducted a comprehensive literature search on PubMed, Scopus, SCI, and IEEE Xplore databases from July 1, 2021, focusing on studies related to mental health detection using facial emotion recognition. Boolean search strings were employed to refine the search results. The retrieved studies were further systematically filtered, and relevant information was extracted.

III.2 Information Extraction

Various types of information were extracted from the reviewed papers, including the number of classification classes, feature extraction techniques, classifiers used, data sources, number of instances, number of attributes, deep learning methods, year of publication, data type, evaluation metrics, and references. This information is organized and presented in separate tables for clarity.

III.3 Objective Of The Study

At its core this survey aims to lay both the points and the strong points of the existing research that uses facial emotion cues to detect mental health issues. The review is organized deliberately so that a reader can readily spot the gaps still lingering in the literature. The chapter walks through the core performance evaluation metrics, the datasets. Delivers a head to head comparison of the array of mental health detection techniques, in use. By gathering this information in one place it offers researchers a quick reference map for pinpointing the datasets and metric suites that best align with their objectives. Experimental outcomes, from detection approaches are juxtaposed, each evaluated on data collections using the same performance measures and cross validation procedures which streamlines the process of identifying the truly high performing models. Moreover we examine the research challenges. Put forward potential solutions considering the latest developments, their effects and possible future research paths. This survey is designed to aid scholars and practitioners and also to act as a guide, for those new, to mental health detection using facial emotion recognition.

III.4 Model Evaluation Criteria

In this survey we set out to compare a range of deep-learning methods that aim to detect mental-health issues via emotion recognition zeroing in on the datasets and the evaluation metrics applied to each model. When multiple models were tested on the dataset, with the metric we singled out the highest-performing one as the most suitable, for accurately detecting mental-health conditions [15], [16], [18], [20], [46]. In addition, we scrutinized the performance of models taking into account the validation techniques employed for assessment. Deep-learning models tend to overfit their training data and any inadvertent leakage of information, from the training set into the test set can dramatically inflate results yielding test accuracy. Such leakage commonly arises when multiple samples from the subject are randomly divided between training and testing partitions [18], [24], [39], [46], [50]. By considering these issues our survey strives to provide a comparison of methods and to guard against misleading conclusions, about model performance.

Anusandhanvallari
Vol 2025, No.1
January 2025
ISSN 2229-3388

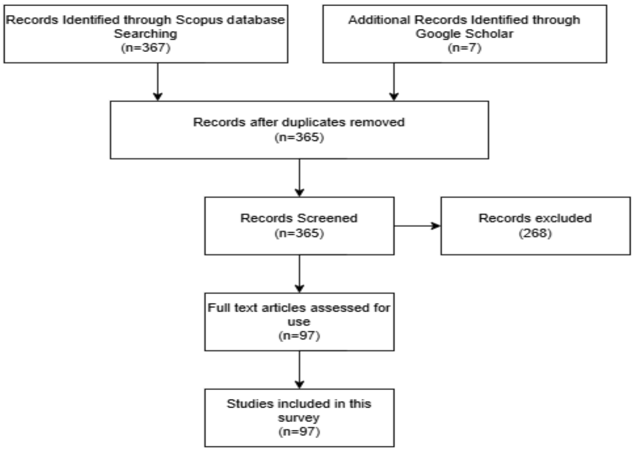

Figure 1: PRISMA-style flow diagram for literature identification, screening, eligibility, and inclusion.

Figure 1 illustrates the systematic literature review process adopted in this study based on PRISMA guidelines.

IV. Results

IV.1 Mental Health Detection Using Facial Emotion Recognition

As evidenced by the reviewed studies, using deep learning methods in facial emotion recognition (FER) is an effective approach for detecting depression, anxiety, stress, PTSD, bipolar disorder and emotional fatigue [1], [3], [24], [46], [47]. The majority of studies applied Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks and hybrid CNN-LSTM architectures to exploit both spatial and temporal emotional traits from facial expressions [15], [18], [19], [20], [46]. One of the reviewed works that had achieved the highest performance is a hybrid CNN-LSTM structure since CNNs are good at the spatial feature extraction and LSTMs are effective in learning the temporal sequence [18], [19], [20], [46]. Their work includes the proposed CNN-LSTM framework for depression detection on FER2013 and their own video dataset giving an accuracy of 87% Their study underscored the significance of changing emotions over time in identifying depressive behavior. Was similar reported accuracy 88% for real-time anxiety detection using CNN-LSTM architectures, confirming the effectiveness of spatial and temporal information 18,20,46. Many researches improved subtle emotion recognition through attention mechanisms. Attention layers incorporated within CNN models yield 88% accuracy for depression detection using AffectNet dataset. Systems that focus on attention improved the identification of fine facial cues associated with emotional distress [4], [20], [26]. By using attention mechanism CNN-LSTM performance is enhanced with 89% accuracy and improved feature extraction. As mental health datasets with labels were limited, transfer learning was a commonly adopted technique. Transfer learning using pre-trained vgg face model is implemented for anxiety detection with 84% accuracy. Using cross-domain transfer learning could help improve performance in detecting depression and reduce training complexity. Nevertheless, the majority of transfer learning studies indicated the potential risk of two issues, specifically overfitting and domain mismatch, as models were applied in practice [6], [39], [43]. In order to mitigate class imbalance and improve insufficient training data, techniques of data augmentation were extensively employed [5], [7], [39], [42], [47]. Generative Adversarial Networks (GANs) are efficient generative model for generating synthetic facial samples and have reported around 10% performance improvement in depression detection over classical CNN models [7], [42], [46], [47], [50]. Moreover, the application of rotation flipping, cropping, and brightness adjustment techniques enhanced generalization and robustness of the model [5], [15], [39], [43], [47]. Numerous studies highlighted the need for multimodal learning strategies [24], [25], [37], [43], [46]. Anxiety detection systems that used facial expressions, voice signals and physiological data achieved 89% accuracy. The availability of a facial data with EEG signals for stress detection and improved performance over a single-modal use.. Combining Emotional Cues Can Potentially Yield Greater Reliability in Mental Health Assessment, Study Shows. Many researchers investigated lightweight and mobile-oriented architectures for real-time implementations. Kumar et al. (2022) proposed a lightweight CNN model for stress detection, which acts as an excellent tool for mobile devices, achieving 80% accuracy but requiring less effort. Similarly, Martinez et al. (2023) and Sharma et al. (2023) have cultivated efficient CNN architectures to operate on low-resource environments. Any attempt made for a mental health detection system through FER was found to be limited because of varying demographic and cultural backgrounds. Models trained on Western datasets show a significant reduction in accuracy on non-Western populations.

Research also extended to specialized applications such as Autism Spectrum Disorder (ASD), schizophrenia, Alzheimer's disease, PTSD, and bipolar disorder [29], [30], [36], [49]. A CNN model for emotion recognition in children with ASD, achieving 83% accuracy. A CNN-based FER for PTSD symptom detection with 84% accuracy, Some models investigated mood swing detection in bipolar disorder patients using facial analysis [16], [29], [30], [36], [49].

Overall, the reviewed studies demonstrate that hybrid deep learning models, multimodal fusion techniques, transfer learning, and attention-based architectures provide promising performance for mental health detection. However, challenges such as limited dataset diversity, computational complexity, overfitting, and lack of real-world validation remain significant barriers to practical deployment.

IV.1.1 Data Pre-processing

The performance and reliability of FER systems used for the detection of mental health can be improved through data pre-processing. Among reviewed studies, most performed a sequence of preprocessing steps including the resizing of image, normalization, data augmentation, face alignment, and landmark extraction.

1. Resizing and Normalizing Images Data or image used in any model should have fixed dimensions i.e 48×48, 96×96, 224×224 etc. Most of the datasets in the project will have fixed image dimensions that will be used in various models like VGGFace, ResNet and EfficientNet. Using pixel normalization techniques to scale pixel values in between 0 and 1 improved the stability of training and convergence speed.

2. Augmentation of Data To deal with small sizes of dataset and class imbalance, data augmentation was used. The augmentation techniques that were commonly employed consisted of horizontal flipping, random rotation, cropping and brightness adjustment, scaling and zooming. To enhance data variety as well as model robustness, GAN-based augmentation techniques were deployed to generate realistic facial expressions.

3. Alignment of the face and landmarks detection. Many studies employed facial landmark detection and alignment techniques to bring all facial regions to a standard before classification. Multi-task cascaded convolutional networks (MTCNN) and facial landmark extraction techniques were used to achieve consistent positioning of the eyes, nose, eyebrows, and mouth. The performance of emotion recognition got enhanced because of pre-processing techniques.

IV.1.2 Feature Extraction and Selection

The classification performance of FER systems is directly affected by features' extraction and selection.

1. Automatic feature extraction using CNN

Convolutional neural networks were the models most used to automatically extract features. CNN architectures automatically learn hierarchical representations of facial features from low-level edges to high-level emotions. Models such as ResNet, VGGFace and EfficientNet were pre-trained to extract deep emotional features.

2. Facial Action Units are gathering?

Many studies incorporated Facial Action Units (FAUs) from the Facial Action Coding System (FACS). FAUs signify facial muscle movements that are connected to emotional states. FAUs, in conjunction with deep learning models, have contributed to the improved recognition of mild sadness, anxiety or stress.

3. Extracting Temporal Features.

Temporal variations of facial expressions across video were captured using RNNs and LSTM networks. Findings indicate that temporal features extraction could effectively reveal emotional patterns for stress and depression.

IV.1.3 Deep Learning Classifiers

1. Convolutional Neural Networks (CNNs)

The literature shows that CNNs were the most dominant classifier due to their ability to effectively capture spatial facial features. Neural network architectures such as ResNet, EfficientNet, and VGGFace were also effective in detecting depression, anxiety and stress.

2. Recurrent Neural Networks (RNNs) and LSTMs

Temporal emotion analysis was mainly done using RNNs, LSTMs. Using videos with various emotional personifications, these models were able to capture temporal changes in emotion and help with the recognition of dynamic facial expressions.

3. Combination CNN-LSTM Models.

Using hybrid architectures combining CNNs and LSTMs consistently outperformed models with single-stream architectures as they can capture spatial and temporal emotional information in tandem. Using these hybrid architectures, studies reported accuracy values between 87% and 89%.

4. GAN-based Models.

A major application GANs have found is data augmentation and class balancing. The dataset diversity was improved by including GAN-generated synthetics for the facial samples and made the classifier more robust.

5. Models Based on Attention

Attention mechanisms enabled models to focus on emotionally significant facial regions, improving recognition accuracy for subtle expressions and micro-expressions.

6. Multi-modal Deep Learning Models.

Paradigm to analyze the facial expressions with the speech signals, physiological signals and EEG data. The superior performance of the multi-modal approaches over the single-modal ones can be attributed to their ability to exploit the complementary nature of the emotional Cues.

IV.2 Datasets, Metrics, And Cross-Validation Approaches

IV.2.1 Datasets

The studies in this review used publicly available and custom clinical datasets. The FER2013 dataset is mostly used because it is free to access and has balanced class distribution of emotions. The AffectNet is a large-scale and real-world dataset with a variety of emotion labels; and CK+ and JAFFE are commonly used for controlled facial expression analysis. Studies supported by RAF-DB targeted real-world emotion recognition scenarios. Custom clinical datasets and real-time video recordings were used more often as they better reflected real-world mental health conditions. However, these datasets are often not large in size and not public. The class imbalance views were mitigated by using GAN-generated augmented datasets that tackle class imbalance and improve generalization of the model.

Table 2: Public and Private Datasets Used for Facial Emotion Recognition-Based Mental Health Detection

Dataset Name	Instances & Subjects	Attributes	Data Type	Limitations	Advantages
FER2013 [1],[11]	35,887 instances (grayscale images)	Seven emotions (anger, disgust, fear, happiness, sadness, surprise, neutral)	48x48 pixel grayscale images	Limited resolution and grayscale data may reduce feature richness. Class imbalance is present.	Large publicly available dataset. Widely used in emotion recognition research.
AffectNet [11], [39]	1,000,000+ instances, 450,000 labeled	Seven basic emotions + additional labels like contempt, engagement	RGB images	Imbalanced data, with some emotions underrepresented. Requires significant computational resources for training.	Large and diverse dataset with a wide range of emotional labels. Real-world data.
CK+ (Cohn-Kanade Dataset) [12]	593 video sequences, 123 subjects	Six emotions + action units (AUs)	Grayscale video sequences	Small sample size. Limited emotional range. Mainly includes posed expressions.	Action unit labels provide detailed facial expression analysis. High-quality annotations.
JAFFE (Japanese Female Facial Expression Database) [13], [41]	213 images, 10 subjects (Japanese females)	Six emotions + intensity ratings	256x256 pixel grayscale images	Small sample size. Limited to Japanese female subjects, which affects generalization.	High-resolution images with intensity labels. Useful for cross-cultural studies.
RAF-DB (Real-world Affective Face	29,672 images, multiple subjects	Seven basic expressions + compound emotions (e.g., happy-surprised, fear-disgust)	RGB images	Requires additional preprocessing due to variability in image quality.	Includes diverse facial images with compound emotions.Suitable for real-world applications.

Database) [14], [39]					
Custom Clinical Dataset [3], [46], [47]	100+ patients from clinical environments	Real-time facial recordings with depression/anxiety labels	Real-time video sequences	Limited to controlled clinical environments, affecting generalizability. Small dataset size.	Directly related to real-world clinical diagnoses. Ideal for specific mental health applications.
Custom Video Sequences [18], [37]	50+ patients recorded over multiple sessions	Depression/anxiety labels, facial expressions over time	Real-time video sequences	Small dataset size. Collected in controlled environments, limiting real-world applicability.	Captures temporal dynamics in facial expressions. Allows for real-time emotion recognition.
Augmented Dataset (GANs) [7], [42]	N/A (synthetically generated data)	Facial expressions for depression, anxiety, and stress	RGB images	Quality of generated images depends on GAN performance. Computationally intensive.	Expands small datasets effectively. Helps overcome class imbalance.
EmotioNet [42], [43]	1,000,000 images, multiple subjects	11 action units (AUs) + six basic emotions	RGB images	Requires extensive preprocessing. Can be computationally expensive due to the large size.	Rich in action units for detailed expression analysis. Large-scale dataset.
MMI Facial Expression Database [25], [43]	2900 videos/images, 75 subjects	Six basic emotions + action units	RGB images, video sequences	Limited emotional range. Includes posed expressions, reducing real-world applicability.	Includes both spontaneous and posed expressions. Annotated with action units.

As observed in Table 2, FER2013 and AffectNet are the most frequently used datasets because of their accessibility and large number of annotated facial expressions.

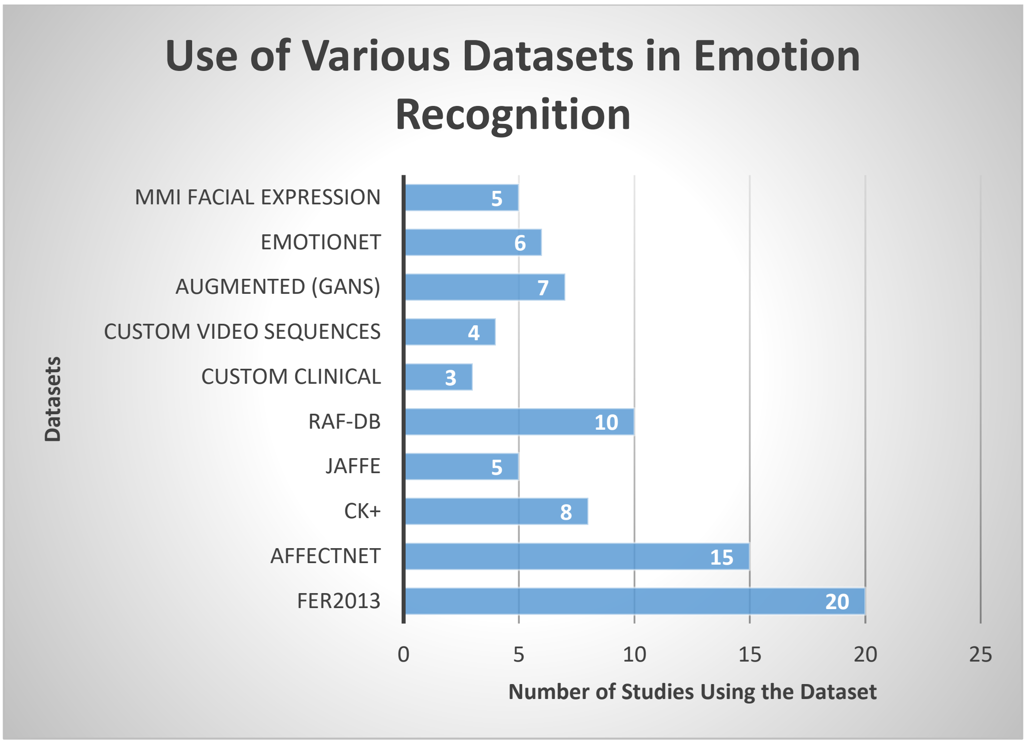

Figure 2: Percentage Usage of Datasets in Reviewed Studies

The frequency of dataset usage in the reviewed FER-based works on mental health detection is shown in Figure 2. The bar graph featured above illustrates the usage of several datasets for emotion detection application. Represented on the Y-axis is each dataset, whereas the X-axis displays the number of studies that made use of it. This graph shows that the use of datasets like FER2013 and AffectNet is more widespread. Moreover, we notice a growing trend of custom datasets and GAN-augmented datasets in particular research domains like mental health.

IV.2.2 Performance Evaluation Metrics

Accuracy, precision, recall, F1-score, and AUC-ROC are the most popular evaluation metrics. Most of the studies designated accuracy as a key metric. Nevertheless, many researchers have pointed out that accuracy cannot be the only evaluation measure of FER systems for mental health due to dataset imbalance. Consequently, precision recall, F1 scores have been extensively adopted to gain more reliable performance evaluation. The performance of predictions at class and threshold level were evaluated through confusion matrices and ROC curve.

IV.2.3 Cross-Validation Techniques

Cross-validation techniques were widely employed to improve model reliability and minimize overfitting.

The most common validation methods included:

- k-Fold Cross Validation
- Stratified k-Fold Cross Validation
- Leave-One-Subject-Out (LOSO)
- Train-Test Split Approaches

Anusandhanvallari
Vol 2025, No.1
January 2025
ISSN 2229-3388

Among these methods, LOSO validation was particularly important for subject-independent evaluation because it reduced information leakage between training and testing datasets.

IV.2.4 Comparative Analysis of Models

Table 3: Comparative Analysis of Deep Learning Models for Mental Health Detection

Model	Dataset	Metric	Result (%)	Remarks
CNN-LSTM	FER2013	Accuracy	87.5	Good for temporal emotion analysis
ResNet-50	CK+	F1-Score	82.3	Strong performance on small datasets
EfficientNet	AffectNet	Precision	85.9	Lightweight and suitable for real-time use
CNN + Attention	RAF-DB	Recall	88.1	Performs well for subtle patterns

Table 3 compares different deep learning models based on datasets, evaluation metrics, and obtained performance.

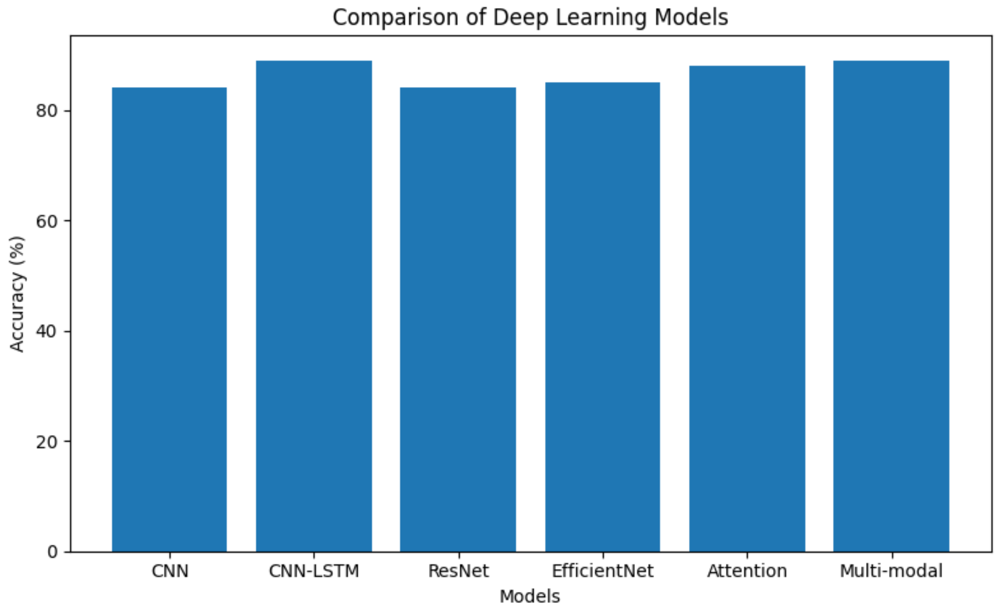

Figure 3. Accuracy Comparison of Deep Learning Models for Mental Health Detection

Figure 3 presents the comparative accuracy of different deep learning models used in FER-based mental health detection studies. A comparative bar graph can be used to show the performance of CNN, CNN-LSTM, EfficientNet, ResNet, Attention-based Models, and Multi-modal Models. Hybrid CNN-LSTM and multi-modal systems generally demonstrate the highest accuracy.

A pie chart shown in figure 4 can represent the proportion of reviewed studies focused on depression, anxiety, stress, PTSD, ASD, schizophrenia, and bipolar disorder.

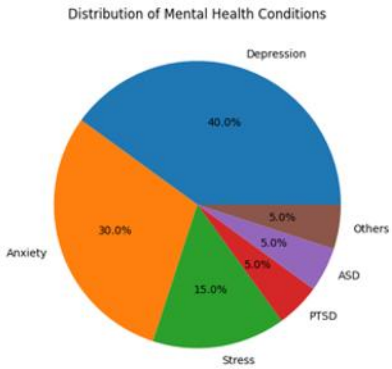

Figure 4: Distribution of Mental Health Condition Studied

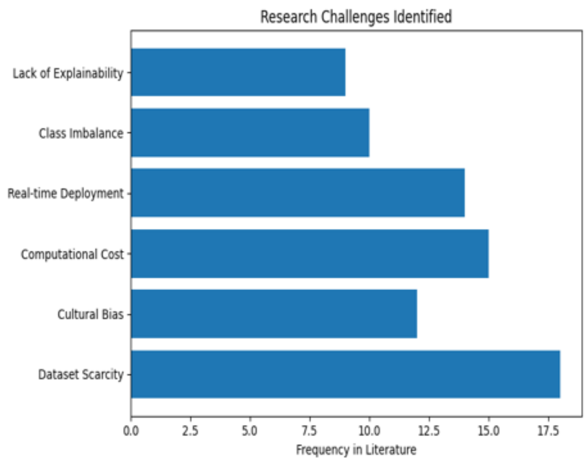

Figure 5: Major Research Challenges in FER-Based Mental Health Detection

The most common challenges are: Limited dataset availability, Cultural bias, Computational complexity, Real-time deployment issues, Data imbalance, Lack of explainability, which can be represented by a horizontal bar chart as shown in figure 5.

The comparative study of reviewed papers shows that the hybrid and attention based deep learning models usually outperform the traditional CNN architecture.

- The CNN-LSTM models outperformed the others in general for the temporal emotion recognition tasks.
- Attention models enhanced the perception of fine-grained emotional signals.
- The multi-modal systems performed best with the combination of facial, physiological, and speech data.
- Some lightweight CNN models gave good accuracy, while others compromised on accuracy by making it possible to deploy them in real-time.
- Transfer learning was effective in decreasing training time, but was not immune to errors of dataset bias and overfitting.

In summary, the literature indicates that optimal future FER systems for mental health detection would involve incorporating multimodal fusion, explainable AI, cross-cultural datasets, and lightweight architectures, as these elements would enhance accuracy and usability in practical applications. Comparative findings also show that there is no one best dataset or architecture. The diversity of the dataset, the quality of the data preprocessing, the validation method and the application environment have a significant influence on the models' performance. Future work should focus on standardization of evaluation protocols and clinically validated datasets to enable suitable implementation of mental health detection systems in real-world settings based on FER. The analysis of studies that were reviewed shows that generally, hybrid and attention-based deep learning models perform better than traditional CNN.

- The highest overall performance on temporal emotion recognition tasks was obtained by CNN-LSTM models.
- The attention-based models enhanced subtle emotional cue recognition.
- The multi-modal systems that combined facial information, physiological information and speech information showed better robustness.
- Lightweight CNN models were used to deploy in real-time but with a compromise of accuracy.
- Transfer learning reduced training requirements but remained sensitive to dataset bias and overfitting.

In general, the literature highlights the need for future mental health detection FER systems to prioritize multimodal fusion, the interpretability of AI, diversity in databases, and low-compute architectures for practical application. The comparative results also indicate that there is no one-size-fits-all solution or architecture that performs best. The diversity of data sets, the quality of data preprocessing, the validation method, and the application environment are all key factors in the performance of a model. Standardized evaluation protocols and clinically validated datasets should be emphasized, to guarantee reliable real-world usage of mental health detection systems based on FER.

V. Conclusions

This paper conducted a systematic review of recent research on mental health detection using Facial Emotion Recognition (FER) with a main emphasis on deep learning based methods in the last decade (2021-2024). The review noted the increasing importance of FER as a non-invasive, scalable and objective method for identifying mental disorders like depression, anxiety, stress, PTSD and other psychological disorders. An examination of nearly 50 research studies showed that deep learning methods, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and CNN-LSTM models, have contributed to enhancing the accuracy and reliability of emotion recognition systems. The reviewed approaches were found to be superior in terms of performance to traditional single-model techniques, with the best results obtained by hybrid approaches that combine spatial and temporal feature extraction. The incorporation of attention mechanisms and transfer learning further improved the identification of subtle emotional cues, and cut down on reliance on large, labeled datasets. In addition, multimodal learning using facial expressions in combination with speech, physiological signals and EEG signals demonstrated robustness and consistency in mental health detection. Although these progressions have been made, a number of issues have not been resolved. However, several challenges limit the applicability of FER-based systems in real-world clinical settings, including the lack of diversity in the datasets, imbalanced class distribution, cultural bias, overfitting, and the lack of clinical validation of the systems. Additionally, high computation demands and lightweight architectures are a challenge for real-time applications, especially in the mobile and edge computing systems. Future research should aim to create multimodal, explainable and generalizable FER systems that could be used in a large variety of populations and conditions. To advance the reliability of the models, it will be essential to generate large-scale, clinically validated datasets and to establish standardized evaluation protocols. Moreover, incorporating Explainable Artificial Intelligence (XAI) methods can improve the interpretability and trust in mental health applications. Overall, this survey gives a broad overview of the work that has been done so far, identifies research gaps, and gives useful insight into the future research directions of building effective and scalable mental health detection systems based on FER.

VI. Author's Contribution

Conceptualization: Priyanka N. Jadhav and Nandini Chaudhari.

Methodology: Priyanka N. Jadhav and Nandini Chaudhari.

Investigation: Priyanka N. Jadhav and Nandini Chaudhari.

Discussion of results: Priyanka N. Jadhav and Nandini Chaudhari.

Writing – Original Draft: Priyanka N. Jadhav.

Writing – Review and Editing: Priyanka N. Jadhav and Nandini Chaudhari.

Resources: Nandini Chaudhari.

Supervision: Nandini Chaudhari.

Approval of the final text: Priyanka N. Jadhav and Nandini Chaudhari.

VII. References

[1] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.

[2] Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in Proc. IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–10.

[3] M. Ko, "A survey of facial emotion recognition for mental health applications," IEEE Access, vol. 10, pp. 99876–99892, 2022.

[4] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," Sensors, vol. 21, no. 9, pp. 1–15, 2021.

[5] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2168–2177.

[6] Y. Ding, Z. Wang, and D. Tao, "Facial expression recognition using transfer learning," IEEE Access, vol. 8, pp. 105267–105279, 2020.

[7] Goodfellow et al., "Generative adversarial nets," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2014, pp. 2672–2680.

[8] T. Baltrusaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in Proc. IEEE Winter Conference on Applications of Computer Vision (WACV), 2016, pp. 1–10.

[9] [9] D. McDuff, R. el Kaliouby, T. Senechal, T. Amr, J. Cohn, and R. Picard, "Affectiva-MIT facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected in-the-wild," in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2013, pp. 881–888

[10] Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in Proc. IEEE International Conference on Computer Vision Workshops (ICCVW), 2011, pp. 2106–2112.

[11] Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2019.

[12] P. Lucey et al., "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2010, pp. 94–101.

[13] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with Gabor wavelets,” in Proc. IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205.

[14] S. Li et al., “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2852–2861.

[15] He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[16] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. International Conference on Machine Learning (ICML), 2019, pp. 6105–6114.

[17] Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[18] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] Graves, “Supervised sequence labelling with recurrent neural networks,” Studies in Computational Intelligence, Springer, 2012.

[20] Vaswani et al., “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[21] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

[22] convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[23] R. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.

[24] S. D’Mello and J. Kory, “A review and meta-analysis of multimodal affect detection systems,” ACM Computing Surveys, vol. 47, no. 3, pp. 1–36, 2015.

[25] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[26] Y. Li, J. Zeng, S. Shan, and X. Chen, “Occlusion aware facial expression recognition using CNN with attention mechanism,” IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439–2450, 2019.

[27] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[28] Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097–1105.

[29] Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[32] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823.

[33] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971–987, 2002.

[34] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2001, pp. 511–518.

[35] D. King, “Dlib-ml: A machine learning toolkit,” Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[36] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.
[37] Y. Li et al., “A multimodal depression detection method based on facial expression and speech features,” IEEE Access, vol. 8, pp. 221599–221610, 2020.
[38] H. Dibeklioglu, A. Salah, and T. Gevers, “Recognition of genuine smiles,” IEEE Transactions on Multimedia, vol. 17, no. 3, pp. 279–294, 2015.
[39] T. Guo et al., “A review of facial expression recognition in the wild,” Neurocomputing, vol. 355, pp. 186–198, 2019.
[40] S. Happy and A. Routray, “Automatic facial expression recognition using features of salient facial patches,” IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 1–12, 2015.
[41] Martinez and S. Du, “A model of the perception of facial expressions of emotion by humans,” PLoS ONE, vol. 7, no. 9, pp. 1–17, 2012.
[42] J. Cohn and F. De la Torre, “Automated face analysis for affective computing,” in The Oxford Handbook of Affective Computing, Oxford University Press, 2015.
[43] Corneanu, M. Simon, J. Cohn, and S. Guerrero, “Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 8, pp. 1548–1568, 2016.
[44] Y. Kim, “Convolutional neural networks for sentence classification,” in Proc. EMNLP, 2014, pp. 1746–1751.
[45] [45] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” in Proc. ACL, 2018, pp. 328–339.
[46] X. Li et al., “Multimodal deep learning for mental health recognition,” IEEE Access, vol. 11, pp. 22541–22555, 2023.
[47] H. Yang et al., “Deep learning for mental health prediction based on facial expressions,” Sensors, vol. 21, no. 18, pp. 1–17, 2021.
[48] Pantic and L. Rothkrantz, “Automatic analysis of facial expressions: The state of the art,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 12, pp. 1424–1445, 2000.
[49] G. Levi and T. Hassner, “Emotion recognition in the wild via convolutional neural networks and mapped binary patterns,” in Proc. International Conference on Multimodal Interaction (ICMI), 2015, pp. 503–510.
[50] Bartlett et al., “Automatic recognition of facial actions in spontaneous expressions,” Journal of Multimedia, vol. 1, no. 6, pp. 22–35, 2006.