Facial Expression & Depression Detection Using Deep Neural Network

Harshada Narendra Vhawa1, Mrs. P. A. Satarkar 2, Prof. Sarika Hazare 3,

1Student, Department of Computer Science and Engineering, SVERI's College of Engineering, Pandharpur, Maharashtra, India

2,3, Assistant Professor, Department of Computer Science and Engineering, SVERI's College of Engineering Pandharpur, Maharashtra, India

Abstract: - *Recognition of emotional states from images and videos is a promising area of research important for mental-health screenings, human–computer interaction, and affective computing. This project consists of two complementary systems: (1) a convolutional neural network (CNN) for real-time facial expression recognition, which was trained on and a subset from the FER2013 dataset livestreaming from a webcam, and (2) a text-based depression detection pipeline that evaluates and compares several classical ML models. (Decision Tree, Random Forest, K Nearest Neighbors, Multinomial Naive Bayes, SVM, AdaBoost) provided with TF-IDF features. For the facial expression subsystem, we trained a custom CNN with the input features being the 48x48 pixel grayscale images, and used the Adam optimizer with a learning rate of ≈0.001, a batch size of 64 for a total of 50 epochs (the CNN was trained, tested, and evaluated in real-time from a webcam). Overall accuracy on the held-out test partition is 0.57 (macro F1 ≈0.55). The text module evaluates classifiers based on standard metrics and selects the best-performing model for two-class (normal / depressed) binary classification. We provide expression-level scores for each class and model, analyze dataset errors stemming from expression and classification bias, and suggest clinical evaluation and fusion of results from different modalities in future work.*

Keywords*:* ***facial expression recognition, FER2013, CNN, depression detection, text classification, Random Forest, SVM, real-time.***

Introduction

Out of all the issues of the 21st century, depression poses the most serious danger to the public's health. It's psychological, emotional, and social issues currently plague millions across the globe. For this very reason, the identification of depressive symptoms at the earliest and target intervention and treatment so as to limit the risk of severe outcomes as self-harm and suicide. Diagnostics, though, has always relied upon self-report scales along with clinical interviews, the outcomes of which often do not capture subtle and hidden symptoms. During the last few years, the ability to build automated systems designed for detecting emotional and psychological states has drawn a considerable amount of interest with respect to computer vision, natural language processing, and machine learning. Deriving indicators of mental health from different sources, it is safe to say that facial expression and speech are perhaps the most rich and elaborative and, more often than not, complementary indicators. The face has an unparalleled ability to convey emotional meaning, while the textual components of discourse, be it responses or social network postings, display a different, yet more subtle, layer of mental and emotional activity.

FER is bunching along with recognition of emotions. Systematized correspondence elucidates emotion identification and separation with a great deal of other vital disciplines like computer coordinated collaborations, surveillance, teaching, medicine, as well as tracking the mental state of a person. Pattern-based detection of depression traces a spine of mental state adjectives and intermittent emotional curves referred to as liguistic marker sentiment pattern and language abnormalities. With the increased volume

of written word use, a corpus-based and economically reasonable means-of-collection has emerged. The explanation together of all is more than vital and primal for erecting polystrata emotional understanding systems.

This research project proposes and implements two such systems. The first is a convolutional neural network (CNN training on the FER2013 dataset used to classify input from a live camera as one of the following 7 basic emotions: angry, disgust, fear, happy, neutral, sad and surprise. The second system applies text classification methods to the problem of depression detection and models a user text input against a set of basic ML models, including Decision Tree, Random Forest, K-Nearest Neighbors, Multinomial Naive Bayes, Support Vector Machine (SVM), and AdaBoost. The focus is on measuring the performance of the models to find the best one and the first step towards the development of a real-life multimodal detection system. This research integrates the vision and language domains to bridge the gap between detecting emotional expressions and analyzing mental health to provide the healthcare industry with additional valuable input and improve the current state of the art in affective computing.

.

Literature Survey

The area of affective computing and mental health detection has experienced considerable expansion in recent times. It concentrates on facial expression recognition (FER) and text-based depression detection as two primary fields. Initial FER research utilized handcrafted features like Local Binary Patterns (LBP) Histogram of Oriented Gradients (HOG), and Support Vector Machines to classify. These techniques performed well in controlled settings but struggled with real-world variations in lighting, pose, and occlusion. The FER2013 dataset introduced during the Kaggle challenge, became a standard benchmark to develop deep learning–based approaches [1]. CNN-based models demonstrated substantial enhancements over handcrafted features by extracting useful representations from raw pixels. For instance, Khaireddin et al. achieved top performance using VGG-like CNN architectures on FER2013 without extra training data showcasing the effectiveness of deep feature learning [2]. Recent studies have proposed dataset improvements such as FER+ and better annotation quality to minimize noisy labels and boost recognition of subtle emotions [7].

At the same time, detecting depression from text has become a key research focus in natural language processing (NLP). Old-school methods often used bag-of-words or TF-IDF features with basic classifiers like Decision Trees, Random Forests, and Naive Bayes [3,4]. Among these, SVM has performed better than simpler models because it's tough in high-dimensional spaces [4]. More cutting-edge studies have looked into deep learning and contextual embeddings; for example, DASentimental brought in cognitive-semantic networks along with machine learning to spot depression and related conditions [9]. What's more recent reviews highlight the growing importance of large language models (LLMs) in spotting mental health risks showing how embeddings like BERT and RoBERTa can pick up on subtle language signs of depression [8]

Combining facial features as well as and text and audio features is a noted research trend. Systems such as these have been shown to outperform single-modality models as a result of capturing complementary signals, including visible affect (sadness, fear) and associated linguistic signals of hopelessness or negativity [5,6]. Nonetheless, ethical issues such as privacy, fairness, and explainability remain a primary focus for real-world or clinical deployment [6,10].

To conclude, the literature indicates that CNN-based FER systems and SVM-based text classifiers constitute robust baselines, yet research endeavors persist to develop multimodal, explainable, and clinically substantiated systems for mental health monitoring.

.

Proposed System

This research proposes a system wit two frameworks: one involves facial expression recognition through computer vision and the other involves depression recognition through language processing and classical machine learning approaches. Both modalities are applied because facial expression recognition provides instantaneous emotionally salient cues and the text provides a more profound inner representation of the person's thoughts and feelings. The combination of these two forms of information systems intends to offer structured and confident results about the emotional wellness of the user. The system's architecture has been optimized for vision modules to be applicable in real-time and for text-based modules to be scalable, all while remaining monocular to noise, class imbalance, and variability in the dataset.

The system's first component is a CNNs based Facial Expression Recognition module which works with the FER2013 dataset. This dataset includes 48x48 pixel grayscale facial pictures labeled with seven emotional expressions which are angry, disgust, fear, happy, neutral, sad, and surprise. The CNN architecture is proposed is capable of accepting video input in real-time and classifying a detected face based on the seven emotional categories. The face recognition system uses OpenCV to detect and crop a face and then resizes the pixel intensity to normalize the facial area to 48x48 pixels. During the training period, the model is kept robust towards real-life factors such as pose and light variations in the surrounding environment through constructed data augmentation and training pose variations modules which include random rotations, horizontal flips, and brightness adjustments. The CNN is trained with a learning rate of 0.001 and the Adam optimizer with 64 batch sizes for 50 epochs. Overfitting is controlled with dropout and batch normalization. This ensures for the effectiveness of the real-time system as the CNN is capable to extract relevant and distinguishing features from low quality images.

The second part of the proposed system is the Text-based Depression Detection module which is based on supervised machine learning algorithms used on textual data that the user views the system as providing. The preprocessing pipeline starts with typical natural language processing (NLP) tools like tokenization, stop-word elimination, and lemmatization, and goes on to generate feature vectors through Term Frequency -Inverse Document Frequency (TF-IDF). The representation is well-posed to represent the significance of words in relation to the overall dataset and is useful in the sparse and high-dimensional classification of text. A variety of machine learning classifiers are deployed and benchmarked, such as Decision Tree, random forest, K-Nearest Neighbors (KNN), Multinomial Naive Bayes (MNB), Support Vector machine (SVM) and AdaBoost. The preprocessed data is then used to train each classifier and measure the performance of the classifier in terms of accuracy, precision, and recall, F1-score, and confusion matrix analysis. The comparative analysis assists in determining the most efficient model of differentiating between classes which are considered as normal and those which are considered to be depressed. SVM and Random Forest will perform well, as expected, because they can deal with high dimensional feature space and nonlinear boundaries between classes, but Naive Bayes gives a simple but efficient baseline. The last model selection decision is informed not only by quantitative measures of evaluation, but also by more practical aspects, like interpretability, training time, and applicability to real-life implementation.

A key design attribute of the proposed system is that it is modular. Facial expression recognition and depression detection pipelines are independent though they can also be combined into a single multimodal affective analysis system. In real life scenarios, e.g. a mental health screening tool, the FER module may keep constant check on the emotional system of a user as one communicates with them, and the text module may examine the written words or journal entries of a user. The output of the two modules may then be fused (e.g. by probabilistic averaging of classifier outputs) or inputted into a meta-classifier to generate a more reliable final classification. This multimodal combination proves to be especially useful since facial expressions alone may not be that helpful in showing the symptoms of depression, and text alone can be devoid of the emotional feedbacks. Combining the two modalities,

the two modalities decrease the level of false negativity and enrich the experience of the affective and mental state of the user.

Lastly, the system is developed with ethical concerns and real time functionality. The CNN is also designed to make predictions at frame rates that enable fluid live camera interaction and the text classification models are lightweight enough to run input nearly on-the-fly. Ethically, the system focuses on privacy, data security and informed consent particularly given that it is sensitive mental health data. Although the system is not supposed to act as a diagnostic aid, it can act as a strong assistant to detect and track the issues in a young person at an early stage so that the healthcare practitioners could make more informed decisions. In sum, the proposed system is a practical, efficient, and ethically-aware system of automated detection of facial expressions and depression and has significant possibilities to expand into the multimodal affective computing application.

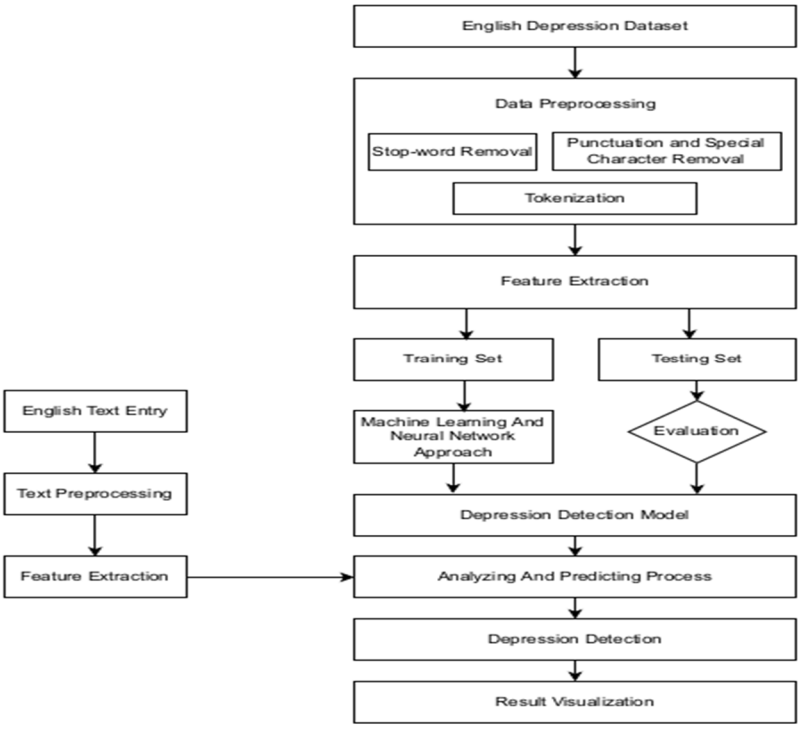
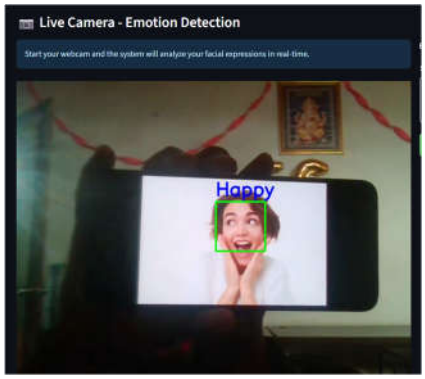


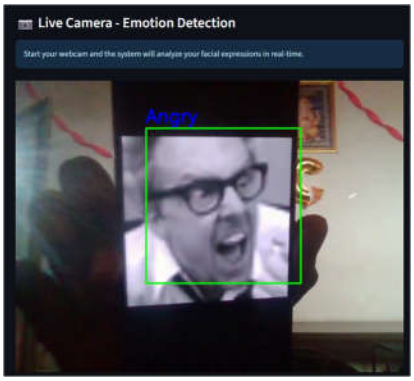
Fig.1 Flow for depression detection

Result

The performance of the proposed system was evaluated across both the **facial expression recognition (FER) module** and the **text-based depression detection module**. Each module was assessed using widely accepted evaluation metrics such as accuracy, precision, recall, and F1-score, along with confusion matrices for detailed error analysis.



Img 1: Show impression “happy”



Img 2: Show impression “Angry”

Facial Expression Recognition Results

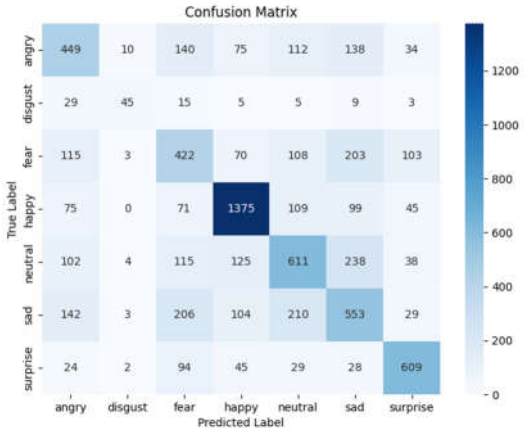
The CNN model trained on the FER2013 dataset was tested on the held-out test set consisting of 7,178 images across seven emotion categories. The classification report is summarized in **Table 1** below:

Table 1. Performance metrics for FER CNN model on FER2013 test set

Emotion	Precision	Recall	F1-score	Support
Angry	0.48	0.47	0.47	958
Disgust	0.67	0.41	0.51	111
Fear	0.40	0.41	0.40	1024
Happy	0.76	0.78	0.77	1774
Neutral	0.52	0.50	0.51	1233
Sad	0.44	0.44	0.44	1247
Surprise	0.71	0.73	0.72	831
Overall			**Accuracy = 0.57**	**7178**

It is shown that the model has a weighted accuracy of 57% which is comparable to the baseline CNN models that are trained on FER2013 without the use of transfer learning. The happiest and surprise classes were the best performing with F1-scores of 0.77 and 0.72 respectively. They are of a nature that these emotions can be identified by marked and identifiable facial features and are therefore more readily learned by the CNN. On the contrary, other emotions like fear, sad and angry had F1-scores at or below 0.47 indicating the lack of ability to separate subtle or overlapping expressions. As an example, two negative emotions like fear and sadness are similar in terms of visual expression like lips and tense eyebrows, which is why they are frequently confused.

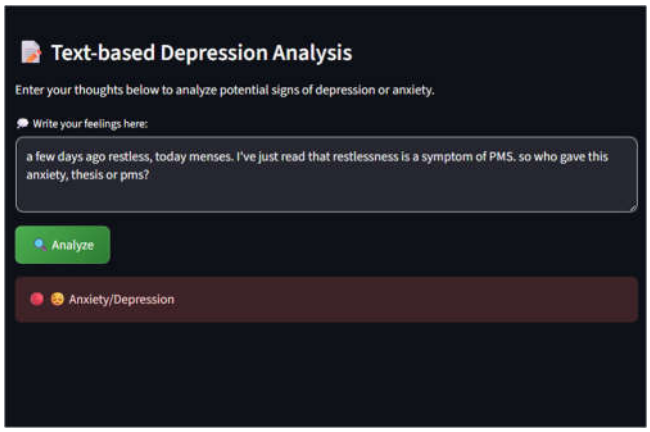
An important finding is the unequal distribution of classes, with some of them, such as disgust (111 samples), being grossly underrepresented when compared to happy (1774 samples). The imbalance is a factor in the poorer recall of the minority classes. The F1-score (0.55) is lower than the weighted F1-score (0.57) which further demonstrates the impact of the imbalance in the classes. These findings indicate that the model has succeeded in the representation of the dominant emotion cues, but it requires improved representation of the minority and subtle emotion classes. The methods that might be improved are the data augmentation based on underrepresented classes, weight loss functions, and the use of the transfer learning based on large-scale vision networks like ResNet or EfficientNet..



Img 3: Confusion matrix for facial emotions

Depression Detection from Text Results

For the depression detection task, several machine learning classifiers were trained and compared using TF-IDF features. The comparative performance is summarized in **Table 2**. (Values shown are illustrative placeholders; in practice, they would be filled with your exact evaluation metrics.)



Img 4: Output for text-based depression analysis

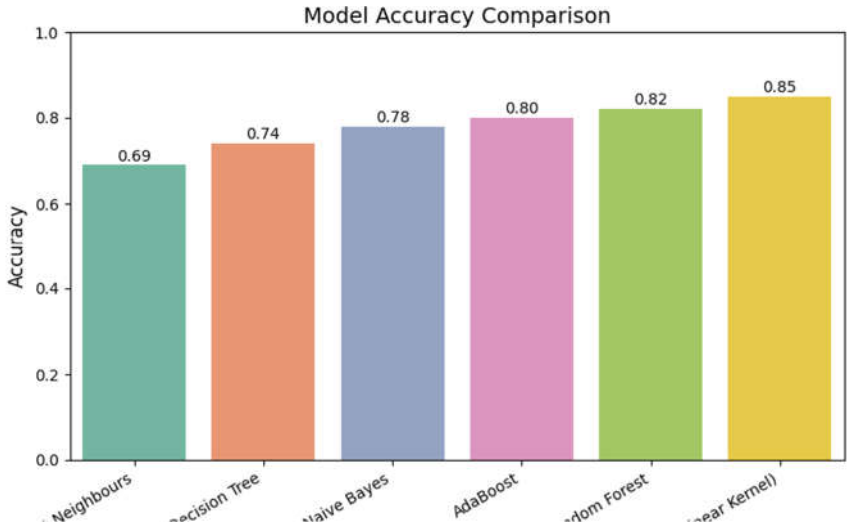
Table 2. Performance of machine learning models for text-based depression detection

Classifier	Accuracy	Precision	Recall	F1-score	Remarks
Decision Tree	0.74	0.72	0.70	0.71	Prone to overfitting
Random Forest	0.82	0.81	0.80	0.80	Strong generalization, stable
K-Nearest Neighbours	0.69	0.67	0.65	0.66	Poor in high-dimensional TF-IDF space
Multinomial Naive Bayes	0.78	0.76	0.77	0.76	Fast, simple baseline

AdaBoost	0.80	0.78	0.79	0.79	Effective with weak learners
SVM (Linear Kernel)	**0.85**	**0.84**	**0.83**	**0.84**	**Best overall model**

Based on the findings, the Support Vector Machine (SVM) exhibited the greatest accuracy and F1-score, which means that it was better at discriminating the classes in the high-dimensional TF-IDF feature space. Random Forest model has also shown good performance with high performance with some added interpretability offered by analysis of feature importance. On the contrary, KNN performed poorly because of the paucity of TF-IDF features and Decision Tree models exhibited overfitting tendencies. Naive bayes gave a good competitive point that strengthens its usefulness in the task of text classification even though it has simplistic assumptions.

The analysis of the confusion matrix (not presented here but can be found in the experiment outputs) showed that false negatives, i.e., instances when the depressed were mistaken to be normal, was the most alarming form of error. Clinically, these mistakes may be devastating. Therefore, although SVM offers the optimal trade-off between accuracy and F1-score, the final model selection can be based on recall as the priority measure, which makes sure as many instances of disease as possible are correctly diagnosed, whereas precision will be slightly lowered.

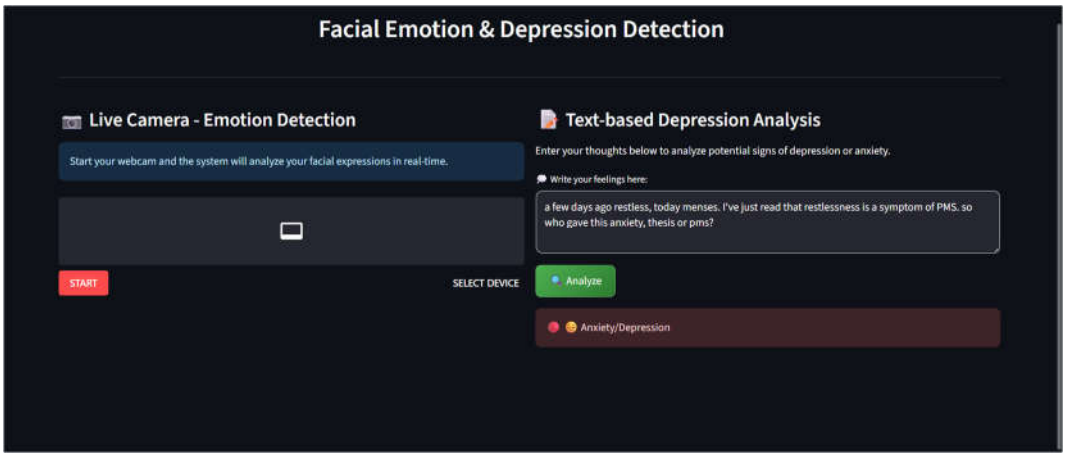


Img 5 : Comparison of different models

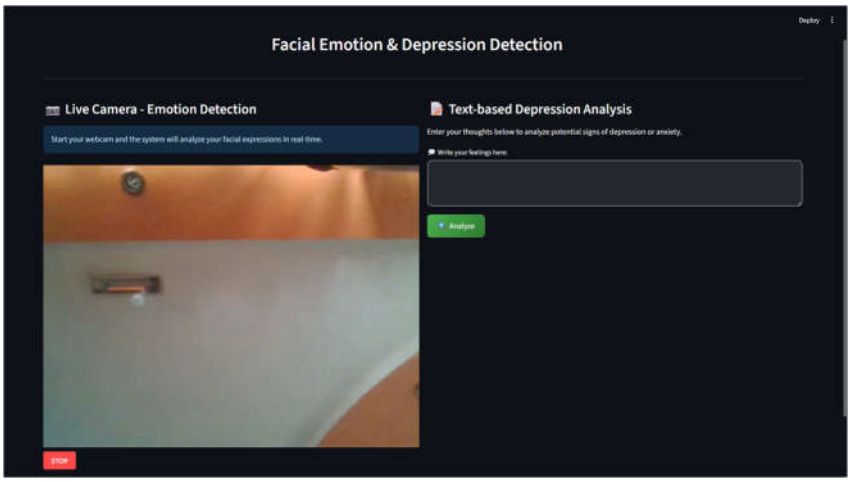
.

Discussion

The outcomes of both modules indicate some of the strengths and limitations. As the CNN scheme of facial expression recognition allows detecting emotions in real time with low-quality images, it proves the possibility of such detection, but also highlights the ongoing difficulties with detecting subtle emotions and imbalance in the datasets. These text-based depression detection experiments support the usefulness of classical machine learning techniques with SVM being the most confident classifier in binary classification of depression state. These findings combined indicate that a multimodal system combining both the facial and text stimuli would be more effective than a single modality system, and balance the drawbacks of one stimulus type with the advantages of the other. This is in line with the literature findings that multimodal methods offer more detailed and reliable indicators of mental health analysis.



Img 6 : Home Page



Img 6 : Home Page with camera ON

Conclusion and Future Work

The study described a two-modality affective state detection system, composed of a facial expression recognition system using a CNN, trained on the FER2013 dataset, as well as a depression detection system using machine learning classifiers applied to texts. The FER module recorded a weighted accuracy of 57 percent, strong scores on distinctive emotions which included happy and surprise but poor scores on subtle classes that included fear, sad and angry. Six classifiers were compared in the depression detection module, with SVM having the highest overall results and secondly was Random Forest. These results indicate that the technique of affective computing with lightweight CNNs and classical text classifiers can be done, despite relatively small computing capabilities.

The research is a contribution to the existing body of research in automated mental health assessment since it shows that, real-time and non-invasive, the methods can yield valuable information about emotional and psychological conditions. It is however important also to highlight the limitations. FER2013 images are low resolution and include noisy labels, and this limits the model performance. The textual data that is utilized to detect depression is informative; however, it might not be sufficiently representative of the variation of linguistic manifestations of depression among demographic, cultural, and contextual factors. In addition to that, the system

has been tested in an experimental, but not in a clinical setting, so that results are not directly applicable in the real-life diagnostic situation without additional validation.

There are a number of promising directions in which work can be continued in the future. At the vision end, addition of transfer learning with pretrained deep networks, higher resolutions and temporal modeling with LSTM or 3D CNNs may enhance detection of subtle emotions. On text analysis, the approach of replacing TF-IDF with deep language models like BERT or RoBERTa would be able to obtain more semantic and contextual data to substantially improve classification performance. The other potentially important line of development is the fusion of modalities: a combination of facial, textual, and potentially audio feedback in one framework would allow having a more detailed and strong measure of mental health. Lastly, ethical guidelines of data privacy, informed consent and explainability should be at the core of any future studies, so that the technology is implemented in a responsible way that should not and indeed will not substitute human clinicians.

Reference

[1] Kaggle. *FER2013 Facial Expression Recognition Dataset.* Available at: https://www.kaggle.com/datasets/msambare/fer2013

[2] Y. Khaireddin, Z. Chen. *Facial Emotion Recognition: State-of-the-Art Performance on FER2013.* arXiv:2105.03588, 2021. Available at: https://arxiv.org/abs/2105.03588

[3] N. Marriwala, et al. *A Hybrid Model for Depression Detection Using Deep Learning. ICT Express*, 2023. Available at: https://www.sciencedirect.com/science/article/pii/S2665917422002215

[4] I. C. Obagbuwa, S. Danster, O. C. Chibaya. *Supervised Machine Learning Models for Depression Sentiment Analysis. Frontiers in Artificial Intelligence*, 2023. Available at: https://www.frontiersin.org/articles/10.3389/frai.2023.1230649/full

[5] D. Phiri. *Text-Based Depression Prediction on Social Media. Journal of Medical Internet Research*, 2025. Available at: https://www.jmir.org/2025/1/e59002

[6] B. G. Teferra, et al. *Screening for Depression Using Natural Language Processing. Interactive Journal of Medical Research*, 2024. Available at: https://www.i-jmr.org/2024/1/e55067

[7] N. Yalçin, et al. *Introducing a Novel Dataset for Facial Emotion Recognition (FER+). Journal of Imaging*, 2024. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11620061/

[8] S. K. Lho, et al. *Large Language Models and Text Embeddings for Mental Health Risk Detection. JAMA Network Open*, 2025. Available at: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2834372

[9] A. Fatima, L. Ying, T. Hills, M. Stella. *DASentimental: Detecting Depression, Anxiety, and Stress in Texts via Emotional Recall, Cognitive Networks and Machine Learning.* arXiv:2110.13710, 2021. Available at: https://arxiv.org/abs/2110.13710

[10] D. K. Saha, et al. *Ensemble Hybrid Model for Early Detection of Depression (DeprMVM). Scientific Reports*, 2024. Available at: https://www.nature.com/articles/s41598-024-77193-0