# AI-Enhanced Emotional Assessment: Detecting Depression Levels through Visual and Vocal Expressions

Ms. Swati Suryawanshi[1], Mrs. Roomana Hasan[2], Mr. Kantilal Chandre[3]

*Abstract*—**In Automatic depression assessment supported visual and vocal cues may be a rapidly growing research domain. This exhaustive review of existing approaches as reported in over sixty publications during the last ten years focuses on image processing and machine learning algorithms. Visual indication of depression, many proceed used for data gathering, and existing datasets are reviewed. The article describes techniques and algorithms for visual feature extraction, dimensionality reduction, decision methods for classification, and regression approaches, also as different fusion strategies. A quantitative meta-analysis of reported results, counting on performance metrics robust to chance, is included, identifying general trends and key pending issues to be treated in future studies of automatic depression assessment utilizing visual and vocal cues alone or together with cues. The proposed work also administered to predict Depression levels consistent with the current input of videos using deep learning also as NLP.**

*Index Terms*—**Image Processing, Machine Learning, Classification Rule, Convolution Neural Networks, NLP etc**

## I. INTRODUCTION

In many situations, humans who are depressed are totally ignorant of their disturbed mental condition. They are incapable to recognize the cause of even unhappiness in them and ultimately such users/peoples fall into a state of mind where they start having suicidal tendencies. In some cases, peoples do know that they are suffering from depression, but they are hesitant to seek any kind of help from anyone mainly due to the wrongly conceived notion of 'humiliation' associated with depression. It is more useful to recognize the signs of depression at the initial steps of depression [1].

Depression is recognized in the primary grades, only a simple one-hour discussion with a counselor may be of tremendous help for the people. This may completely change the negative case of perception of that student into a positive one. Such a student can be provided excellent counseling on how to trade with mental stress and can be controlled to reflect the correct way to success. The most valuable form of non-verbal conversation is the facial expressions of a person. Many studies have been done for finding out the facial expressions that are related to depression [2].

The popular work is essentially offered to find out the appearance of depression in college students by studying their facial features. This system mainly uses different image processing techniques for face detection, NLP for speech identification, feature extraction, and classification of these features as depressed or non-depressed. The system will be trained with features of depression. Then videos of different students/peoples with a frontal face will be captured using a web camera. Then the facial features of those faces will be extracted for prediction of depression. Based on the level of depression features the student will be listed as depressed or non-depressed [3].

## II. RELATED WORK

- Asim Jan, Hongying Meng, Yona Falinie Binti A. Gaus, and Fan Zhang, in this article automatic depression assessment based on visual and vocal cues is a rapidly growing research domain. The present exhaustive review ofexisting approaches as reported in over sixty publications during the last ten years focuses on image processing and machine learning algorithms. Visual manifestations of depression, various procedures used for data collection, and existing datasets are summarized [4].
- Cynthia Solomon, Michel F. Valstar, Richard K. Morriss & john crowe, this work aims to not only determine audio features that differ between healthy and depressed people, but also to investigate how they change when people with depression try to conceal their true emotions.

Based on a set of optimized features, our goal was automatic depression recognition which will still be able to correctly classify a person as depressed even if they are trying to hide their depression. Healthy individuals who alter their behavior to appear depressed were not of interest in this study [5].

E. Jenkins and E. M. Goldner, the objective of this study is to synthesize extant literature on approaches currently being applied to understand and address this condition. It is hoped that the findings can be used to inform practitioners and guide future research. A scoping review of the scientific literature was conducted with findings categorized and charted by underlying research paradigm. Currently, the vast majority of research stems from a biological paradigm (81%). Research on treatment-resistant depression would benefit from a broadened field of study. Given that multiple etiological mechanisms likely contribute to treatment resistant depression and current efforts at prevention and treatment have substantial room for improvement, an expanded research agenda could more effectively address this significant public health issue [6].

Cynthia Solomon, Michel F. Valstar, Richard K. Morriss & john crowe, this article provides a systematic review and meta- analysis of the literature on automatic emotional facial expression in people with non-psychotic disorders compared to healthy comparison groups. Studies in the review used an emotionally salient visual induction method, and reported on automatic facial expression in response to congruent stimuli. In depression, decreases in facial expression are mainly evident for positive affect. In eating disorders, a meta-analysis showed decreased facial expressivity in response to positive and negative stimuli. Studies in autism partially support generally decreased facial expressivity in this group [7].

## III. PROPOSED SYSTEM

### A. System Architecture

Fig.1 is our system architecture diagram, in which user will register and the details get store in the local database, the user will login and provide a video, which is then separated in two parts a visual and audio [8]. In image processing CNN will be used for detecting face and to find emotions on face like happy, sad, angry, fear, neutral, disgust In this process images is taken from video and different frames are generated.

In audio extraction, when we upload a video then audio is separated from video and wave generation takes place. For speech recognition we use bag of words logic, we have some positive word dataset as well as some negative word dataset, with the help of this we can find the probability of how many words person speaks positive or negative, at last we apply Naive Bayes algorithm for final result [9].

After getting final result the user has a facility to see his report. In this project to address the problem of stress detection three modules have been mainly defined in order to measure the differences of stressed and non-stressed users on social media platforms: System Framework, Social Interactions and Attributes categorization [10].

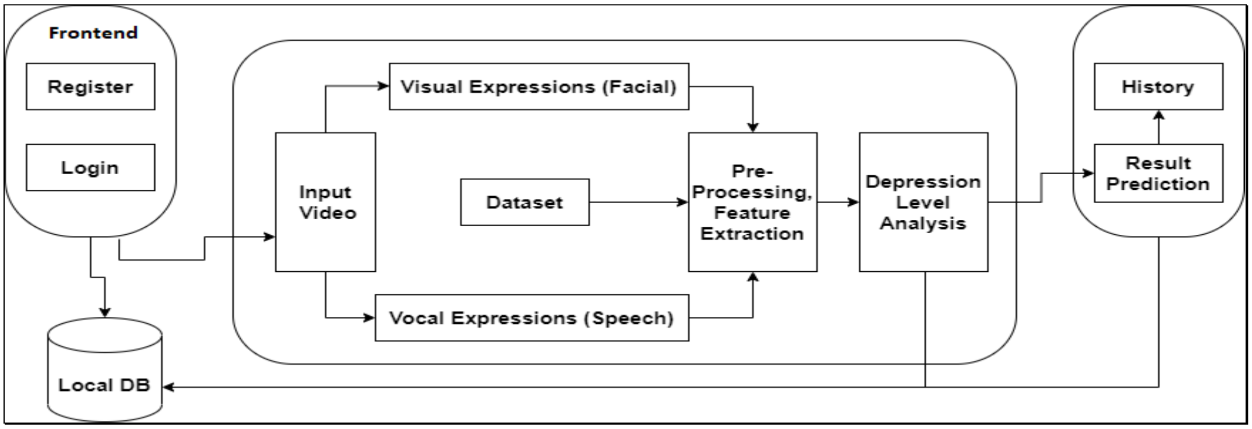
Frontend
Register
Login
Local DB
Input Video
Visual Expressions (Facial)
Dataset
Vocal Expressions (Speech)
Pre-Processing, Feature Extraction
Depression Level Analysis
History
Result Prediction

Figure 1. System Architecture

B. System Framework

In this framework we propose a novel hybrid model - a factor graph model combined with Convolution Neural Network to leverage contents and social interaction information for stress detection. Experimental results show that the proposed model can improve the detection performance by 6-9% in F1-score. By further analyzing the social interaction data, we also discover several intriguing phenomena, i.e. the number of social structures of sparse connections (i.e. with no delta connections) of stressed users is around 14% higher than that of non-stressed users, indicating that the social structure of stressed users' friends tends to be less connected and less complicated than that of non-stressed users [11].

*Social Interaction*

We analyze the correlation of users' stress states and their social interactions on the networks, and address the problem from the standpoints of: (1) social interaction content, by investigating the content differences between stressed and non- stressed users' social interactions; and (2) social interaction

structural diversity, social influence, and strong/weak tie. Our investigation unveils some intriguing social phenomena. For example, we find that the number of social structures of sparse connection (i.e. with no delta connections4) of stressed users is around 14% higher than that of non-stressed users, indicating that the social structure of stressed users‘friends tends to be less connected and complicated, compared to that of non- stressed users[12].

C. Attributes Categorization

We first define two sets of attributes to measu

re the differences of the stressed and non-stressed users on social media platforms: namely image and speech-level attributes from a user 's single user and user level attributes summarized from a user 's weekly activities.

## IV. SYSTEM MODELS

A. Emotion and Face Detection

Facial expression for emotion detection has always been an easy task for humans, but achieving the same task with a computer algorithm is quite challenging. The current approaches primarily focus on facial investigation keeping background intact and hence built up a lot of unnecessary and misleading features that confuse CNN training process. The current manuscript focuses on five essential facial expression classes reported, which is displeasure/anger, sad/unhappy, smiling/happy, feared, and surprised/astonished [13]. This is divided into 3 parts:

A. Facial Detection — Ability to detect the location of face in any input image or frame. The output is the bounding box coordinates of the detected faces.

B. Facial Recognition — Compare multiple faces together to identify which faces belong to the same person. This is done by comparing face embedding vectors.

C. Emotion Detection — Classifying the emotion on the face as happy, angry, sad, neutral, surprise, disgust or fear Humans are used to taking in nonverbal cues from facial emotions. Now computers are also getting better to reading emotions. So how do we detect emotions in an image? We have used an open source data set — Face Emotion Recognition (FER) from Kaggle and built a CNN to detect emotions. The emotions can be classified into 7 classes — happy, sad, fear, disgust, angry, neutral and surprise.

D. Speech to Text Conversion

Speech recognition is an important feature in several applications used such as home automation, artificial intelligence etc. In this process recorded audio was given to Google which creates converted wave file which is in the text format from that audio file.

## V. ALGORITHM DETAILS

A. CNN Algorithm:

CNN is one of the main categories to do image recognition, image classification. Object detection, face recognition, emotion recognition etc., are some of the areas where CNN are widely used. CNN image classification takes an input image, process it and classify it under certain categories (happy, sad, angry, fear, neutral, disgust). CNN is a neural network that has one or more convolutional layers [14].

- Step 1: Dataset containing images along with reference emotions is fed into the system. The name of dataset is Face Emotion Recognition (FER) which is an open – source data set that was made publicly available on a Kaggle.
- Step 2: Now import the required libraries and build the model
- Step 3: The convolution neural network is used which extracts image features f pixel by pixel

- Step 4: Matrix factorization is performed on the extracted pixels. The matrix is of m x n.
- Step 5: Max pooling is performed on this matrix where maximum value is selected and again fifixed into matrix.
- Step 6: Normalization is performed where every negative value is converted to zero.
- Step 7: To convert values to zero rectifified linear units are used where each value is filtered and negative value is set to zero.
- Step 8: The hidden layers take the input values from the visible layers and assign the weights after calculating maximum probability.

B. Speech to Text Conversion:

- Text Mining:

Text Mining is the process of deriving meaningful information from natural language text. A Text Mining refers to the process of deriving high quality information from the text. The overall goal is, essentially to turn text into data for analysis, via application of Natural Language Processing [15].

- Natural Language Processing (NLP):

Natural language processing (NLP) is a field of artificial intelligence in which computers analyze, understand and derive meaning from human language in a smart and useful way. By utilizing NLP, we can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation. NLP primarily acts as an important aspect called as speech reorganization in which system analyze primary source of audio data in the form of spoken words. In NLP, syntactic analysis is used to assess how the natural language aligns with the grammatical rules. Here are some syntax techniques that can be used:

1. Tokenization*:* Tokenization is an essential task in natural language processing used to break up a string of words into semantically useful units called tokens. Generally, word tokens are separated by blank spaces and sentence tokens by stops.

2.Part-of-Speech Tagging*:* It involves identifying the part of speech for every word. It signifies the word is noun, pronoun, adjective, verb, adverb, preposition or conjunction.

3. Bag of Words*:* It splits each string into words and listing it into vocabulary and converts every word of data into its root word.

Naive Baye's Classifier:

It is probability-based algorithm mostly used in text classification. Naive Baye‘s model is easy to build and particularly useful for very large data sets. Along with simplicity, Naïve Baye‘s is known to outperform even highly sophisticated classification methods. Naive Baye‘s algorithm observes each feature independently even if they are related.

## VI. CONCLUSION

We propose a unified deep learning-based framework for Depression Detection. In conclusion, we presented a novel approach to optimize word-embedding for classification tasks. We performed a comparative evaluation on some of the widely used deep learning models for depression detection from tweets on the user level. We performed our experiments on publicly available datasets. Our experiments showed that our CNN- based models perform better than CNN-based models. Models with optimized embeddings managed to maintain performance with the generalization ability. We presented our results of a study that looked into the automatic detection of depression using audio and video features in a human–computer interaction setting. In particular, we set out to discover how hard it would be to fool or cheat such an automated system. In our study maximum matched healthy and depressed participants, we found that depressed participants seemed to follow the predicted pattern of lower energy levels in speech. Many of the prosodic and spectral features that have before been used in emotion recognition were also found to be significant in depression recognition. However, not all features that were significant in differentiating depressed and healthy participants were the same as with those used in emotion recognition.

## REFERENCES

[1] Girard, Jeffrey M., Jeffrey F. Cohn, Mohammad H. Mahoor, Seyed mohammad Mavadati, and Dean P. Rosenwald ―Social risk and depression: Evidence from manual and automatic facial expression analysis‖ 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1- 8. IEEE, 2013.

[2] Deepak A Vidhate, Parag Kulkarni ―Performance comparison of multiagent cooperative reinforcement learning algorithms for dynamic decision making in retail shop application‖, International Journal of Computational Systems Engineering, Inderscience Publishers (IEL), Vol 5,Issue 3,pp 169-178, 2019.

[3] Alghowinem, Sharifa, Roland Goecke, Jeffrey F. Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. ‖Cross-cultural detection of depression from nonverbal behavior‖ 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1-8. IEEE, 2015.

[4] Deepak A Vidhate, Parag Kulkarni ―A Framework for Dynamic Decision Making by Multi-agent Cooperative Fault Pair Algorithm (MCFPA) in Retail Shop Application‖, Information and Communication Technology for Intelligent Systems, Springer, Singapore, pp 693-703, 2019.

[5] Pampouchidou, A., O. Simantiraki, C-M. Vazakopoulou,

[6] C. Chatzaki, M. Pediaditis, K. Marias et al. ―Facial geometry and speech analysis for depression detection‖ 39th Annual International Conference on Engineering in Medicine and Biology Society (EMBC), pp. 1433-1436. IEEE, 2017.

[7] Deepak A Vidhate, Parag Kulkarni ―A Novel Approach by Cooperative Multiagent Fault Pair Learning (CMFPL)‖, Communications in Computer and Information Science, Springer, Singapore, Volume 905, pp 352-361, 2018.

[8] Harati, Sahar, Andrea Crowell, Helen Mayberg, and Shamim Nemati. ―Discriminating clinical phases of recovery from major depressive disorder using the

[9] dynamics of facial expression‖ 38th Annual International Conference of Engineering in Medicine and Biology Society (EMBC), pp. 2254- 2257, IEEE, 2016.

[10] Deepak A Vidhate, Parag Kulkarni, ―Exploring Cooperative Multi-agent Reinforcement Learning Algorithm (CMRLA) for Intelligent Traffic Signal Control‖, Smart Trends in Information Technology and Computer Communications. SmartCom 2017, Volume 876, pp 71-81.

[11] Cohn, Jeffrey F., Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai ―Detecting depression from facial actions and vocal prosody‖ 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. ACII 2009., pp. 1-7. IEEE, 2009.

[12] Deepak A Vidhate, Parag Kulkarni, ―A Novel Approach for Dynamic Decision Making by Reinforcement Learning-Based Cooperation Methods (RLCM)", International Conference on Intelligent Computing and Applications, Springer, Singapore, pp 401-411, 2018.

[13] Tasnim, Mashrura, Rifat Shahriyar, Nowshin Nahar, and Hossain Mahmud. ―Intelligent depression detection and support system: Statistical analysis, psychological review and design implication‖ 18th International Conference on Health Networking, Applications and Services (Healthcom), pp.1-6 IEEE, 2016.

[14] Pampouchidou, Anastasia, Kostas Marias, Manolis Tsiknakis, P. Simos and Fabrice Meriaudeau ―Designing a framework for assisting depression severity assessment from facial image analysis‖ International Conference on on Signal and Image Processing Applications (ICSIPA), pp.578-583, IEEE, 2015.

[15] Deepak A Vidhate, Parag Kulkarni, ―Multiagent Cooperative Reinforcement Learning by Expert Agents (MCRLEA)‖, International Journal of Intelligent Information Systems, Science Publishing Group, volume 6, issue 6,pp72-84,2017.

[16] Meng, Hongying, Di Huang, Heng Wang, Hongyu Yang, Mohammed A I -Shuraifi, and Yunhong Wang.

[17] Depression recognition based on dynamic facial and vocal expression features using partial least square regression‖ 3rd ACM international workshop on Audio/visual emotion challenge, pp. 21-30, ACM, 2013.

[18] Deepak A Vidhate, Parag Kulkarni, ―A novel approach to association rule mining using multilevel relationship algorithm for cooperative learning‖ 4th International Conference on Advanced Computing & Communication Technologies, pp 230-236, 2014.