Exploring The Sentiment Landscape of Depression: A Multimodal Analysis Approach

Prof. D. R. Anekar[1], SujalDhamne[2], Rushikesh Waman[3], Vishal Divekar[4], Rushikesh Salunke[5]
[1, 2, 3, 4, 5] Dept of IT
[1, 2, 3, 4, 5] Savitribai Phule Pune University, Pune, India

Abstract- *Dysthymia is a common mental disorder that can be debilitating, causing people to feel lethargic and unmotivated. It is a major contributor to disability globally. Many depressed individuals are unable to receive the right care since there are few objective ways for assessing depression. Researchers may soon be able to distinguish subtypes of dysthymia based on speech patterns thanks to developments in acoustic characteristics and emotional sensing technologies. Our goal was to identify a group of linguistic traits that characteristics that could potentially reveal or predict dysthymia along with the correlation between this feature set and other symptoms such as suicidality. For examining this correlation. In order to construct a substantial speech feature set, we extracted with many characteristics as we could in accordance with prior studies. In the previous system, people who were found to be depressed were treated with medicines. However, in our system we are providing some techniques based on which persons' level of Dysthymia is being determined. Using Beck Depression Inventory (BDI) technique some questions are to be answered. Facial expressions detect the face, expressions ofpeople. In rare cases the nearby doctor/suggestions/notification are used for Dysthymia detection.*

I. INTRODUCTION

The human mind is affected by many factors, including affection and relationship issues. This can lead to depression, which is often treated using machine leaming. Machine leaming is a process that uses past experiences to provide the best solution when the same situation arises in the future.[2] It takes into account several factors, including user emotions. One of the main factors causing mental illness is depression. It significantly impairs everyday functioning and is a primary contributor to suicidal. thoughts. Machine learning can aid in identification and produce potential treatments for depression.

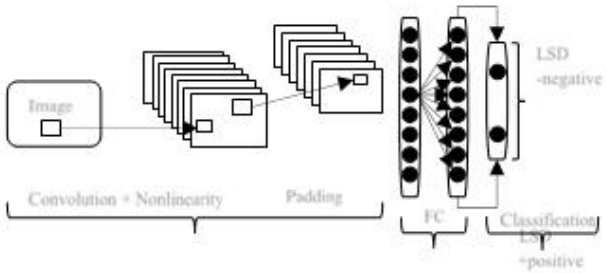


Fig 1: CNN Architecture

CNNs, also known as convolution neural networks, are deep learning networks that learn only from input. CNNs are very useful for seeing patterns in pictures of objects. The categorization of non-image data, such as speech, time - series data, and signal data, may also benefit tremendously from their utilisation.

Either the kernel, a filter, or feature detectors:

The kernel in a convolutional neural network is only a filter that is employed to extract information from the picture.

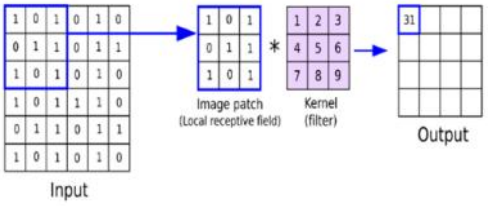


Fig 2: Filter or kernel

Stride:

The neural network's filter's stride parameter determines how much movement there is across the picture or video. We completed stride 1, now we'll go through them one at a time.

The following two pixels will be skipped if stride 2 is provided.

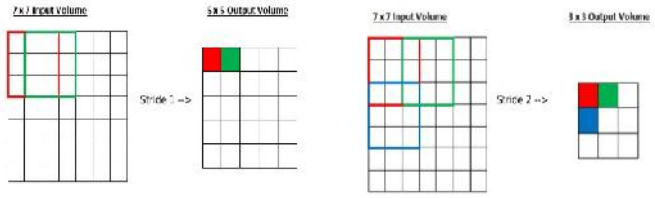


Fig 3: stride

Padding:

Since padding indicates the number of pixels that are added to an image during processing by the CNN kernel, convolutional neural networks (CNNs) can benefit from it. For instance, any extra pixels will have a value of 0 if the padding in a CNN is set to zero. When we scan a picture with a filter or kernel, its size will shrink. We must refrain from doing that in order to preserve the image's original dimensions and retrieve certain low-level information.

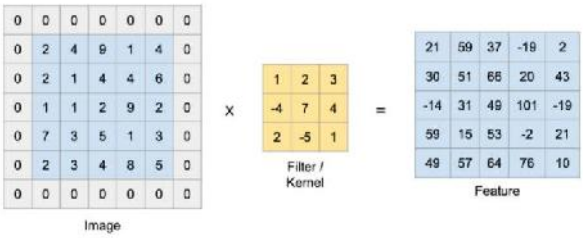


Fig 4: Padding of Images

Pooling

Convolutional neural networks use a method called pooling to generalise the characteristics that the convolutional filters have retrieved, enabling the network to detect features regardless of where they are in the picture.

FC (Fully Connected Layer)

A feedforward neural network is a basic layer that is fully linked. The last layer of the network is a layer that is entirely linked. The last pooling or convolutional layer's output, which has been flattened before being given to serves as the input nodes and is completely linked.

Non-Linearity Layers

A convolutional neural network's nonlinear layer is formed of an activation function that takes the feature map generated by the convolutional layer and outputs an activation map. An element-wise action on the input volume that results in the input and output having the same dimensions is known as an activation function.

1.Sigmoid

The logistic function, from which the sigmoid function is descended, is denoted by (x) or sign (x).

$$\sigma(x)=1/(1+\exp(-x))$$

2. Tanh

The range [-1, 1] is used to compress Tanh Tanh real-valued numbers. The activation saturates like sigmoid neurons, but unlike them, its output is zero-cantered.

3.ReLU

The Rectified Linear Unit (ReLU) has become quite popular in recent years. It computes the value of the function () =max (0,).

In other words, at 0 threshold, the activation just exists.

III. METHODOLOGY

The two components of our suggested structure are picture flow and audio flow. The visual stream has three CNN-based network models in total:one for body gesture and movement identification, one for identification of landmarks, and one for image recognition of faces. Grid search optimization weighs the scores of the audio and visual streams. For categorization, It uses the overall weighted score.

A. The model of facial gestures

The facial emotion model is made up of the facial regions model as well as the facial image model.While analysing videos, we recognise facial emotions using Frame Awareness Networks (FAN) [1]. Figure 1 is a schematic of the FAN framework. Using a variable number of face images as input, it creates a fixed-dimension feature representation for recognizing facial emotions from a facial video. The characteristic embedding component and the frame attention module are the two components that make up the whole network. The characteristic embedding module, a deep CNN, embeds each facial image into a feature vector. The frame attention module learns weights while the feature vectors are adaptively aggregated to produce a single discriminative video representation.

FAN first applies coarse weights to certain frame properties using an FC layer as well as a sigmoid function. Pre-processing and data cleaning mathematically calculate the weight of a i-th frame. We believe that learning weights from

both local and global variables is much more reliable. Using non-linear mapping and individual frame features, the rather coarse previous weights are

B. The audio Model

For the speech component, audio (speech) data is first converted to text data using the Google Cloud Speech to Text API. After data transformation, we use the Bag of Words technique for sentiment analysis. The arc is a simple and effective technique, but it cannot take word order into account. To overcome this problem, we use RNN to capture word order. In the BoW technique, we first preprocess the text to remove stop words. The vocabulary, i.e. the list of words, is created after the preprocessing. After creating the document-term matrix, we apply sentiment analysis algorithms to determine the sentiment of each document.

C. Fusion Strategy

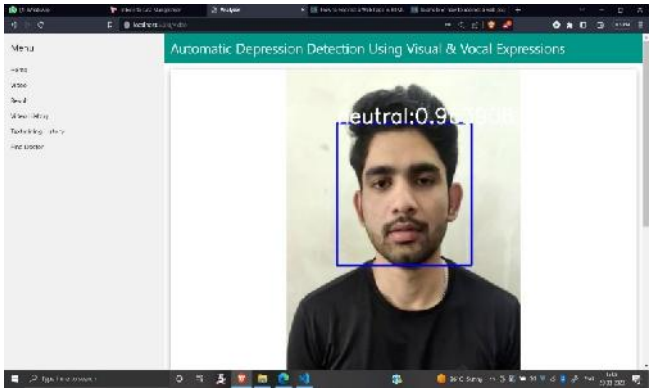
In earlier section, we discussed the several networks we employed in our framework; each one is crucial to classifying emotions and works in concert with the others. Therefore, we must employ a method to combine the output from each network. Each network in our architecture will provide an evaluation vector for every sample that indicates the likelihood that the sample corresponds to a certain emotion. As indicated in the calculation, we utilise weighted sum to combine each score in this paper to arrive at the final score.

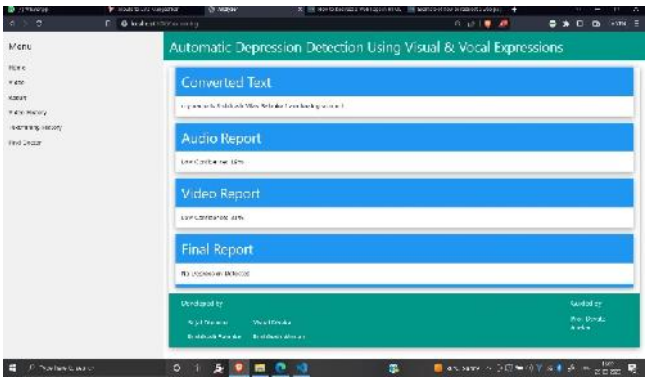
IV. EXPERIMENT

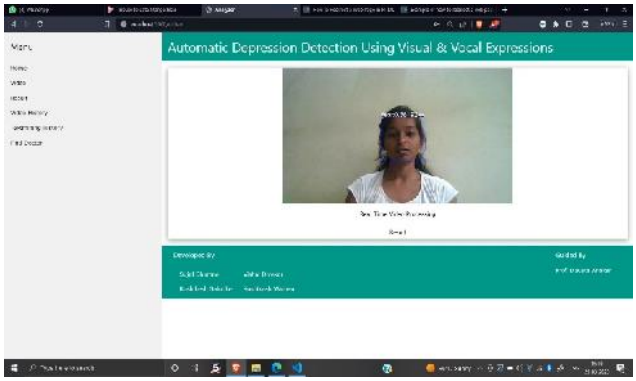
A. Pre-Processing the Emotion Recognition Challenge dataset is impacted by anomalous circumstances including light fluctuation, facial occlusion, and more. Pre-processing is thus required in our architecture. Our first step is to align, normalise, and resize the to 224x224 px, the face.Next, we select frames from video every 2sec interval to use as inputs for the facial image model. Similar to this, after employing 98 facial landmarks initially, we entered 3 frames into the facial landmarks model. To prepare for training, we utilise Cloud Speech-to-Text api to convert the audio into text format to extract features from the audio model. B. Conclusion and Discussion We use several neural networks to extract complementing features for the Multimodal (Audio and Facial) based emotion identification challenge in order to get better and more reliable performance. We have primarily used two types of fusion approaches for our research: fusion by segments and fusion by video. Additionally, each study will include a variety of fusion techniques, such as voting on the highest value and the score produced by each model. Experiments reveal that the latter has superior accuracy. The outcomes of the various networks that we presented are shown the Facial Landmarks Model achieved an accuracy of 70.00%. while the Facial Expression Model was 70.71 percent accurate, that of the Audio Model is 49.29%, It's vital to note that although while the audio-fusion model performs less accurately than other networks, it still has a significant impact because of the complementing data it provides. Finally, the top fusion framework scored 76.43% on the database of validation. In the past, accuracy was determined using segment-based data. Each of the five segments of a video may have a different prediction outcome due to this. As a result, we suggest a novel approach to determining accuracy rate: computation by video unit. Each video's five clips are counted, and the forecast with the greatest number of occurrences is chosen as the final outcome.

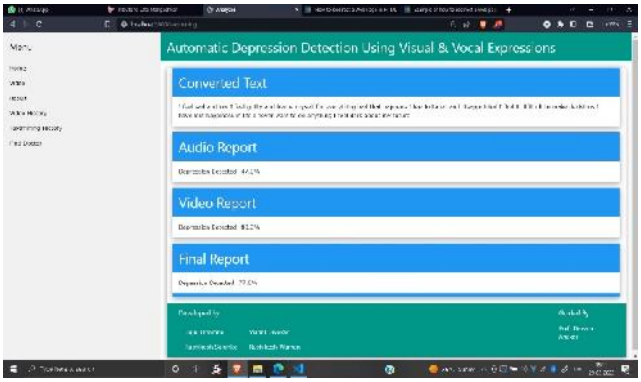
V. RESULT

Model	**Accuracy**
Facial Landmarks Model	70.00%
Facial Expression Model	70.71%
Audio Model	49.29%









VI. CONCLUSION

In this study, we provide a multi-feature framework for the depression detection problem that recognises sentiment. Audio and video are the two complementing aspects of the extracted information. We employ a convolutional neural network to extract sentiment characteristics from video data, and for audio we use Bag of Word technique. This approach is straightforward and effective. The challenge's experiment results shown that our suggested framework is superior for the task of sentiment analysis for dysthymia.

REFERENCES

[1] Q. Wu, Y. Liu, Q. Li, S. Jin and F. Li, "The application of deep learning in computer vision," 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 6522-6527.

[2] D. Ballard, Generalizing the Hough transform to detect arbitrary shapes, Pattern Recognition, vol. 13, no. 2, pp. 111122, 1981.

[3] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, vol. 13, no. 2, pp. 111122, 1981.

[4] V. Bhavana, G. M. Surya Mouli and G. V. Lakshmi Lokesh, "Hand

[5] Gesture Recognition Using Otsu's Method," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, 2017, pp. 1-4.

[6] Y. Liu, J. Zhang, and J. Tian, an image localization system based on gradient Hough transform, MIPPR 2015: Remote Sensing Image Processing, Geographic Information Systems, and Other Applications, 2015.

[7] M. J. Cheok, Z. Omar, and M. H. Jaward, ''A review of hand gesture and sign language recognition techniques,'' Int. J. Mach. Learn. Cybern., vol. 10, no. 1, pp. 131–153, Jan. 2019.

[8] Q. Xiao, Y. Zhao, and W. Huan, ''Multi-sensor data fusion for sign language recognition based on dynamic Bayesian network and convolutional neural network,'' Multimedia Tools Appl., vol. 78, no. 11, pp. 15335–15352, Jun. 2019.

[9] E. K. Kumar, P. V. V. Kishore, M. T. K. Kumar, and D. A. Kumar, ''3D sign language recognition with joint distance and angular coded colour topographical descriptor on a 2–stream CNN,'' Neurocomputing, vol. 372, pp. 40–54, Jan. 2020.

[10] J. Wu and R. Jafari, ''Wearable computers for sign language recognition,'' in Handbook of Large-Scale Distributed Computing in Smart Healthcare.
Scott J. Social network analysis. Thousand Oaks: Sage; 2017.

[11] Serrat O. Social network analysis. In: Knowledge solutions. Singapore: Springer; 2017. p. 39– 43.

[12] Mikal J, Hurst S, Conway M. Investigating patient attitudes towards the use of social media data to augment depression diagnosis and treatment: a qualitative study. In: Proceedings of the fourth workshop on computational linguistics and clinical psychology—from linguistic signal to clinical reality. 2017.

[13] Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol. 2016;9:77–82.

[14] Ofek N, et al. Sentiment analysis in transcribed utterances. In: Pacific-Asia conference on knowledge discovery and data mining. 2015. Cham: Springer.

[15] Yang Y, et al. User interest and social influence based emotion prediction for individuals. In: Proceedings of the 21st ACM international conference on Multimedia. 2013. New York: ACM.

[16] Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol. 2010;29(1):24–54.

[17] Pennebaker JW, Francis ME, Booth RJ. Linguistic inquiry and word count: LIWC 2001, vol.71. Mahway: Lawrence Erlbaum Associates; 2001. p. 2001.

[18] Holleran SE. The early detection of depression from social networking sites. Tucson: The University of Arizona; 2010.
[19] Greenberg LS. Emotion-focused therapy of depression. Per Centered Exp Psychother. 2017;16(1):106–17.
[20] Haberler G. Prosperity and depression: a theoretical analysis of cyclical movements. London: Routledge; 2017.
[21] Guntuku SC, et al. Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci. 2017;18:43–9.
[22] De Choudhury M, et al. Predicting depression via social Media. In: ICWSM, vol. 13. 2013. p. 1–10.
[23] De Choudhury M, Counts S, Horvitz E. Predicting postpartum changes in emotion and behavior via social media. In: Proceedings of the SIGCHI conference on human factors in computing systems. New York: ACM; 2013.
[24] Zhang L, et al. Using linguistic features to estimate suicide probability of Chinese microblog users. In: International conference on human centered computing. Berlin: Springer; 2014.
[25] Aldarwish MM, Ahmad HF. Predicting depression levels using social media posts. In: 2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS). 2017.
[26] Zhou J, et al. Measuring emotion bifurcation points for individuals in social media. In: 2016 49th Hawaii international conference on system sciences (HICSS). 2016. Koloa: IEEE.
[27] Wang X, et al. A depression detection model based on sentiment analysis in micro-blog social network. In: Trends and applications in knowledge discovery and data mining (PAKDD). 2013.
[28] Nguyen T, et al. Affective and content analysis of online depression communities. IEEE Trans Affect Comput. 2014;5(3):217–26.
[29] Park M, McDonald DW, Cha M. Perception differences between the depressed and non- depressed users in Twitter. In: ICWSM, vol. 9. 2013. p. 217–226.
[30] Wee J, et al. The influence of depression and personality on social networking. Comput Hum Behav. 2017;74:45–52.