Grenze International Journal of Engineering and Technology, June Issue

Automatic Depression Level Detection

Prof.Rupali Umbare[1], Vedant Bhamre[2], Danish Tamboli[3], Trushant Jadhav[4] and Pranav Kurle[5]
[1-5]Department of Information Technology, JSPM'S Rajarshi Shahu College of Engineering, Pune, MH-India
Email: vedantbhamre17@gmail.com

***Abstract*— According to physiological research, there are a variety of variances in both speech and face movements. These facial and vocal expressions are shared by healthy and depressed people. On the basis of this information, we offer the Multimodal Attention Feature Fusion and a novel Spatio-Temporal Attention (STA) network technique that are utilized to get the multimodal representation of depression signals to be able to predict the amount of personal depression. Correctly, we first separate segmenting the speech amplitude spectrum and video into predetermined lengths and submitting them to the STA network, which focuses on the audio and video frames used to detect depression in addition to integrating the attentional processing of spatial and temporal information mechanisms. The output of the STA network's final full connection layer is where the audio and video segment-level functionality is acquired. In order to collect the changes in every aspect of the audio and segment-level features for videos and summarize them as an audio and video feature level, this study also provides the eigen evolution pooling approach. The MAFF is then used to create a multimodal representation composed of modal complementary data, which is then inputted into a support vector regression predictor to determine the severity of the depression. The utility of our strategy is illustrated by experimental findings on the depression databases for AVEC2013 and AVEC2014.**

***Index Terms*— Convolutional Neural Network, Multimodal Audio/Video Segment-Level Depression Detection Feature, Machine Learning.**

I. INTRODUCTION

Depression is a condition that causes people to have extremely low moods and the inability to engage in typical social interactions. More gravely, we can observe that depression can also cause behaviors that contribute to self-harm and suicide. As a result, depression will overtake heart disease as the second biggest cause of death by 2030. Fortunately, we can state that early diagnosis and therapy can assist people in quickly getting out of problems. However, the diagnostic process is typically challenging and heavily dependent on the doctors, which can prevent some patients from receiving timely, effective therapy. Finding a system for automatically diagnosing depression is therefore vital to help clinicians work more effectively. The model of automatic depression identification has new opportunities thanks to new algorithms, and this could lead to model improvement by increasing model accuracy and accurately forecasting depressed clients. According to physiological research, depressive patients' speech and facial movements differ slightly from those of healthy people.

II. REVIEW OF LITERATURE

[1] Yanfei Wang et. Al focused on multiple instances learning as well as Sampling, Slicing, Long Short-Term

Grenze ID: 01.GIJET.9.2.34

Memory, and Multiple Instance Learning are methods for feature manipulation. Depression is detected by binary classification, uses the Support Vector Machine algorithm. It covered the capability to detect various small video clips of depression symptoms. Maximum accuracy of 81.06% is achieved by using this learning technique.

[2] Sri Harsha Dumpala et. al offer a multi-task learning framework for enhancing depression performance in terms of severity prediction based on the acoustic aspects of brief speech audio recordings, as well as the usage sentiment and feeling embeddings that are.. The suggested training for many tasks using regression and classification improves the assessment the level of depression, according to experimental findings. Additionally, we demonstrated that, when compared to two separate networks, a CNN that can perform multiple tasks achieves higher classification of sentiment and emotion performance. When paired with acoustic characteristics, sentiment-emotion embedded patterns in this multi-task CNN considerably enhanced the accuracy of estimating depression severity. These enhancements imply that the suggested methods may be applied to creating clinical applications.

[3] Anastasia Pampouchidou et. al Although there are many different approaches to related algorithms documented in the literature, automatic depression evaluation is still in need of considerable development compared to present practices ability to differentiate between various types of depression and how MDD differs from other mood disorders is just one clinical research concern that has to be addressed methodically. Further research is necessary to understand individual variation brought on by concomitant personality disorders or traits, as well as the impact of culture and ethnicity. It's intriguing that, despite the fact that such data can be useful in understanding ongoing emotional responses, physiological activity measured through EMG, BVP, skin conductance, and respiration has not been included in the reviewed multimodal studies, with the exception of those who recorded heart rate using a non-contact, facial video-based system.

[4] Yuan Gong et. al proposed that, as a common mental condition, major depressive disorder must be accurately diagnosed in order to provide targeted intervention and care. Participants in this challenge are asked to use audio, video, and text from an interview that lasts between 7 and 33 minutes to build a model that forecasts the severity of depression. It is difficult to find, collect, and keep crucial temporal details for such lengthy interviews because doing so will result in the loss of the majority of temporal facts. As a result, we suggest an unique topic modeling-based method for doing context-aware analysis. Our tests demonstrate that the suggested strategy outperforms the challenge baseline and the context-unaware method by a wide margin across all criteria. The ability of our approach to find a variety of temporal features that have an underlying relationship with depression and further to build models on them was also discovered by analyzing the features chosen by the machine learning algorithm, which is a task that is challenging for humans to complete.

[5] Asif Sa et. al Despite the fact that depression and social anxiety disorder are relatively widespread, many people who are depressed or anxious choose not to seek counselling. The majority of current tests for these diseases are based on client self-report and clinician judgement, making them cumbersome to perform, vulnerable to subjective bias, and unavailable to patients who have difficulty accessing therapy. The development of methods for identification, assessment, prevention, and therapy might benefit from objective indicators of depression and social anxiety. For the purpose of identifying symptomatic persons and state affect from extensive spoken audio data, we provide a weakly supervised learning system. We specifically present NN2Vec, a novel feature modelling approach that makes use of the innate relationship between voice states and symptoms/affective states. Additionally, we offer a novel MIL modification of the BLSTM classifier, known as BLSTM-MIL, in order to comprehend the temporal dynamics of vocal states in weakly labelled data. We tested our framework using spontaneous audio speech recordings from 105 participants, including speakers who were very socially uncomfortable.

[6] A technique for speech depression detection using deep convolutional neural networks was developed by Karol Chlasta et al. After analyzing five network typologies, ResNet-34 and ResNet-50 were determined to be delivering the best categorization results. The results suggest a workable new method that uses audio spectrograms and quick voice samples for initial screening of depressive individuals. The potential for the spectrograms to generate CNN learnable features was found. This held true despite the challenge of utilizing voice as a predictor of depressed symptoms. We think that the solution's use of 15-second sample intervals helped to reduce the effect of noise. Our system can be used independently or as a part of a more complex, hybrid, or multi-modal strategy.

[7] Richard Caruana et. al proven approach for training artificial neural networks with many outputs, we present multitask learning (MTL) in connection-ism. In actuality, MTL in connection-ism can be seen in the traditional NETtalk application and earlier work on giving suggestions to neural networks. We present an empirical example that demonstrates how MTL can still enhance generalization performance even when the similarity between numerous tasks is challenging to learn and recognize. We go on to show that this progress cannot be

attributed to anything other than a shared inductive bias resulting from the similarity of the tasks. We provide a brief explanation of how to create multitask decision trees from the top down in order to demonstrate the generality of the MTL methodology. Decision trees are not typically used to learn several tasks; therefore, this is noteworthy. By doing this, a system is created that generalizes particular conceptual clustering techniques, enhancing their applicability in fields where the separation between features (information that will become available in the future) and classes (objects we wish to forecast) must be maintained.
[8] Robert J. McAulay et. al worked-on analysis/synthesis method was used to analyze speech that was both clear and interfered with in various ways. In every instance, natural-sounding, high-quality synthetic speech was produced. The technique may also be applied to the parametric representation of non-speech sounds, such as music and particular marine biological noises. Finally, it is important to keep in mind that tools used to change the width of are essential for high-quality speech reconstruction in addition to updating the average pitch. It's vital to remember that, despite using the frequency analysis window, there are no voicing choices made during the analysis and synthesis process.

III. METHODOLOGY

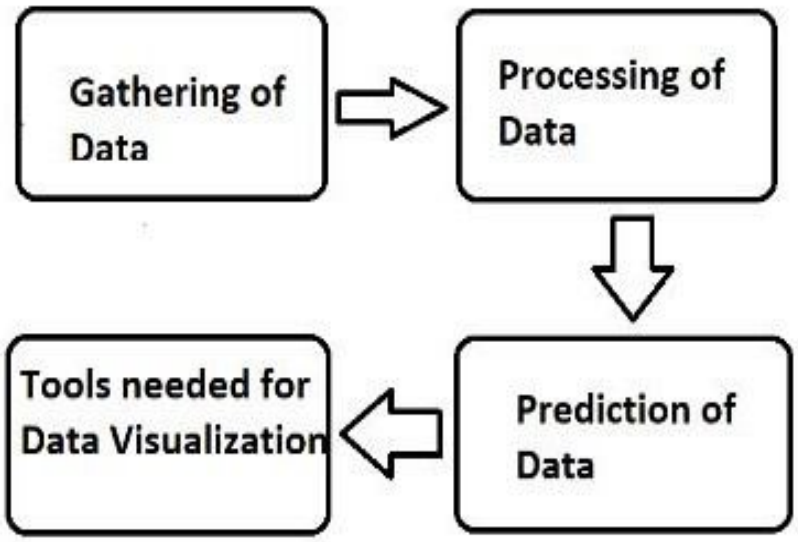

Figure 1: Methodology for gathering of data

1. Gathering of Data: Gathering of data is required at the first stage as we need to create a data-set which can used to analyze and generate a better working model. The data will be collected in the form of audio and video and it will be collected sufficiently in order to create a better working model.
2. Pre-Processing Of Data: Steps involved in pre-processing of data are: Data Cleaning, Feature Selection & Data Transformation.
Data Cleaning is the process which involves removing and fixing the missing or incorrect data which is stored in the database as whenever we create a data set it is bound to have some errors and those errors should be cleaned which will give better and accurate results.
Feature Selection is the process of picking up appropriate and approximate features from the data-set and then accordingly we can direct a way in which our model can be influenced.
Data Transformation is a process where the needs and behavior of algorithm is taken into account and the data is changed accordingly such as the structure or the format of the data.
3. Projection/Prediction of Data: Is a step where we refer to the output after the model is trained on the data set which was provided for training where we can predict the face gestures from the video data set and the tone/audio notes of the voice collected from the audio data set to make predictions based on the previous data.
4. Tools needed for Data Visualization: Data Visualization tools provide a very easy way to create visuals on a data set which are huge. Statistics of the data which are generally not visible are clearly stated and the underlying patterns which are there can be easily uncovered.
The flow chart is shown in the following fig

IV. System Architecture

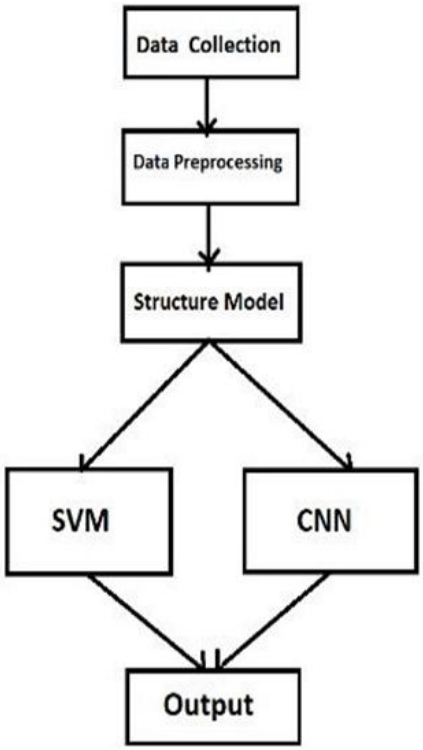

Figure 2: System architecture for our model

Algorithms Used:

A. SVM

The acronym SVM stands for "Support Vector Machine," a machine learning under supervision algorithm that the ability to create regression and classification models that perform well on both linearly and non-linearly distinguishable data. The SVM algorithm performs classification with the help of margin. The objective of the algorithm is to find the borderline which can most accurately distinguish the data points in n-dimension space as of which the boundary line is called the hyperplane.

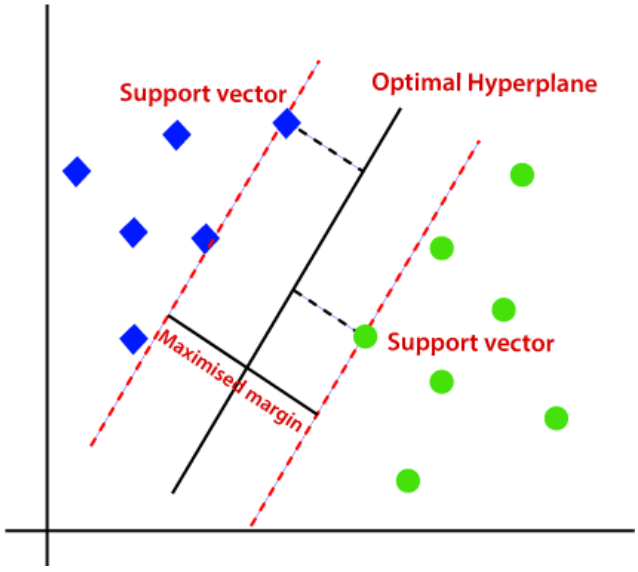

Figure 3: Support Vector Machine Representation

B. CNN

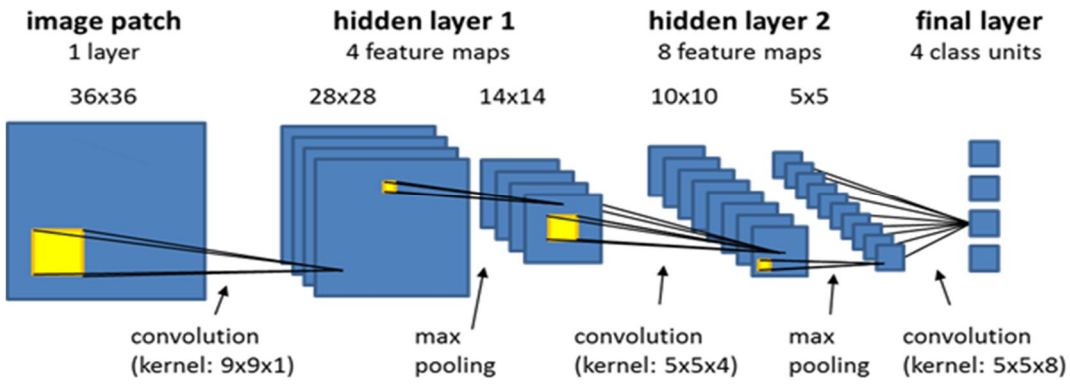

Figure 4: Convolutional Neural Network Representation

CNN: stands for a particular kind of (Convolutional Neural Network) DNN (Deep Neural Networks) and they are applied in visual memory which can be analyzed. This algorithm uses a technique called convolution which is an operation in mathematics in which the operation is performed on double number of functions which then creates 3rd function which shows how the size or form of one is modified by the other. The main aim or objective of the CNN is that to reduce the complexity of the images so that those images can be processed easily and additionally it will also not loose any of its features which will then give us better predictions.

V. Conclusion

According to physiological research, both facial and verbal activity differ little between depressed and very healthy people. In light of this reality, we develop a multimodal spatiotemporal representation paradigm for automatic depression level identification. The suggested STA network focuses on frames relating to depression detection in addition to integrating secular information. In addition, by removing the information between processes, the suggested MAFF method enhances the multimodal representation's quality. Experimental AVEC2013 and AVEC2014 results show that our approach has a decent performance in terms of detection.
Human speech is a sophisticated combination of words and feelings. Every word might mean something different depending on the context in which it is used. Every user will have a different mental state, making it challenging to understand their input. Their feelings can help us grasp what they're going through even better. Making a schedule and choosing the therapies are also aided by this. If the data is accessible, we also think about applying this approach to identify more diseases. To increase the detection accuracy, we will partition various tasks and train separate models in the future.

References

[1] "Automatic Depression Detection Via Facial Expression Using Multiple Instance Learning," IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, by Yanfei Wang, Jie Ma, Bibo Hao, Pengwei Hu, Xiaoqian Wang, Jing Mei, and Shaochun Li
[2] Estimating severity of Depression From Acoustic Features and Embedding of Natural Speech | IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 2021 IEEE, Sri Harsha Dumpala, Sheri Rempel, Katerina Dikaios, Mehri Sajjadian, Rudolf Uher, Sageev Oore.
[3] Automatic Assessment of Depression Based on Visual Cues: A Systematic Review, by Anastasia Pampouchidou, Panagiotis Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis, IEEE Transactions on Affective Computing, 2017.
[4] "Topic Modeling Based Multi-modal Depression Detection," Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Yuan Gong and Christian Poellabauer, 2017, pp. 69–76.
[5] A Weakly Supervised Learning Framework for Detecting Social Anxiety and Depression, Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, vol. 2, no. 2, pp. 1–25, 2018. A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman, and J. J. Stankovic.
[6] "Automated speech- based screening of depression using deep convolutional neural networks," Procedia Computer Science, vol. 164, pp. 618–628 (2019). K. Chlasta, K. Wok, and I. Krejtz.
[7] Multitask learning: A knowledge-based source of inductive bias, R. Caruana, Proceedings of the ICML, 1993.
[8] Speech Analysis/Synthesis Based on a Sinusoidal Representation by Robert J. McAulay, VOL. ASSP-34, NO. 4, AUGUST 1986.