[1]Ankit Kumar

[2] Snehal Godse

[3]Sagar Kolekar

[4] Dilip Kumar Jang Bahadur Saini

[5] Deepak Pandita

[6] Pulkit

Decoding Stress with Computer Vision-Based Approach Using Audio Signals for Psychological Event Identification during COVID-19

Abstract: - Interpreting psychological events can be costly and quite complex. It is simple to translate such experiences into a person's spoken and nonverbal cues. The suggested model investigates a computer vision-based method for using an individual's audio signal to identify stressful psychological events. Different people's input speech signals are recorded and compared to the common questionnaire. A series of inquiries pertaining to the second stage of COVID-19 events are included in the questionnaire set. Through additional processing, these speech signals are converted into frequency components by means of the Fast Fourier transformation (FFT) method. A long short-term memory module processes each frequency component and produces temporal information from each frequency band. The features of speech signals are extracted into the temporal frames by this module. The VGG 16 algorithm is used to further classify each temporal frame into stress and un-stress classes. A classifier with 16 layers of architecture is called VGG 16. A feed-forward convolutional neural network called VGG 16 is used to divide the vast array of speech signal features into classes: stressed and unstressed. The proposed model attempts to recognize speech signals as stress indicators. A standard set of questionnaires with a series of interrogation-style questions has been used to develop the stress symptoms in an individual's mind. The audio signals generated by each person's responses are recorded and subsequently analyzed for stress and un-stress classes. The proposed model was able to identify stress in speech signals with 98% accuracy. The time and cost implications of the suggested model are relevant. Medical research is typically costly and time-consuming.

Keywords: LSTM; VGG 16; CNN model; data preprocessing; speech signal.

I. INTRODUCTION

Artificial intelligence (AI) has brought improvement to the healthcare system. Health monitoring based on AI has great significance. The AI brings convenient techniques to outline emotional states and draw them out into statistical specifications. It makes the recognition of health issues more convenient and automatic [1]. Medical tools and examinations take expensive effort for the identification of health issues. Complex health issues like psychological disorders are expensive and time-consuming its recognize [2]. Among the various emotional states, the proposed system focuses on the stress status of an individual based upon his/her speech signals that are fetched against situations given in the form of interrogation using a standard questionnaire set. Bio signals like ECG are used in earlier techniques for the detection of emotional status. The suggested method makes use of speech signals because they may be easily obtained from microphones without interacting with a person's body [3]. This functionality is useful for consumers as well as for creating a sizable database for a stress-detection system. Nonetheless, bio-signal-based algorithms often have better accuracy than speech-based stress-detection systems. Despite this performance obstacle, improving neural network-based techniques by gathering a lot of data makes speech-based stress-detection systems more attractive [4].

Detecting stress in individuals is important for early intervention and prevention of long-term health problems. One method for stress detection is analyzing speech signals, as speech patterns can reveal changes in emotional and psychological states. Stress detection from speech signals involves analyzing various features of speech

[1]Dr. APJ Abdul Kalam Technical University, Lucknow, 226031, India, 7667ankit@gmail.com

[2]JSPM's Rajarshi Shahu College of Engineering, Pune, Snehal_mca@jspmrscoe.edu.in

[3]SIT- Symbiosis Institute of Technology, Symbiosis International University, Pune, India , Sagar.kolekar@sitpune.edu.in

[4]*Corresponding author: Pimpri Chinchwad University, Pune, 412106, India,dilipsaini@gmail.com

[5]Pimpri Chinchwad University, Pune, 412106, India, Pandita.deep@gmail.com

[6] Jindal Global Business School, O.P. Jindal Global University, Sonipat, Haryana, India ,pulkitcdacnoida@gmail.com

signals such as spectral characteristics, timing, and statistical properties to identify patterns indicative of stress [5]. Machine learning (ML) algorithms can also be trained on a dataset of labeled stress speech to automatically classify new instances as stressed or not stressed. While stress detection from speech signals has the potential to be a non-invasive and effective tool for early detection and prevention of stress-related health problems. Millions of individuals worldwide suffer from stress, a prevalent mental health problem [6]. Anxiety, despair, and weariness are just a few of the many physical and emotional symptoms it can bring on. Early detection of stress is critical to prevent it from escalating and causing further health complications.

Recently, there has been a increasing interest in using speech analysis as a tool for stress detection [7]. Speech is a powerful tool for stress detection since it is a direct reflection of the speaker's mental state. Studies have shown that there are various acoustic and prosodic features of speech that can indicate stress. These features include changes in pitch, loudness, speaking rate, and spectral characteristics of speech signals. Stress detection from speech signal involves analyzing these features and developing algorithms that can accurately classify speech signals as stressed or not stressed. Machine learning (ML) techniques have been employed for stress detection from speech signals. Stress detection from speech signal has many potential applications, including in clinical settings for mental health screening and monitoring, workplace stress management, and public speaking training [8]. However, there are still many challenges to overcome in this field, including improving the accuracy and reliability of stress detection algorithms and addressing issues related to privacy and data protection. Stress detection from speech signal can be achieved by analyzing various features of speech signals. Some of the techniques that can be employed are:

- Frequency domain analysis: Frequency domain analysis involves analyzing the spectral characteristics of the speech signal. This can help in identifying changes in the fundamental frequency and formant frequencies, which are indicative of stress.
- Time domain analysis: Time domain analysis includes investigation of changes in amplitude, duration, and timing of speech signals. A stressed person may have changes in these parameters, such as increased amplitude, longer duration, and altered timing.
- Prosodic analysis: As mentioned earlier, prosodic features such as changes in pitch, loudness, and speaking rate can be indicative of stress.
- Statistical analysis: Various statistical features can be extracted from speech signals to identify stress patterns. Examples include mean and standard deviation of the frequency or amplitude of the speech signal.

The rest of the paper is divided into subsequent parts. Section2 discusses the literature survey. Section 3 illustrates the proposed methodology. Section 4 contains feature extraction. Section 5 describes the experimental results. A last section draws our conclusion and outlines the future work.

II Related Works

The translation of speech signal into stress has many advantages over the medical inspection. Speech signals are translated using computer vision techniques. Microphones make it simple to capture signals from speech, which are then transformed into electronic data. At first, vocal patterns are analogue in nature. Then, feature vectors are generated by translating the transformed data into bits. The traits make it difficult for a medical tool to investigate connected emotions. These characteristics are important for constructing a sizable dataset for a computer vision-based method to investigate a person's emotional traits [9]. The psychology of a person contains emotional activities, which manifest themselves in a variety of ways, including through body language, posture, gestures, eye contact, and expressions on the face, among others. Speaking is the most effective approach to describe emotional activity [10]. Different speech signals are produced depending on psychological processes. Artificial neural systems improve the accuracy and reliability of audio processing. According to a study by the American Psychological Association, changes in psychological processes lead to the unpleasant emotion of stress. It is simple to detect some quality of audio abnormalities in a stressed person. The number of hormones secreted by the body, such as cortisol, has an effect on the vocal cords. There have been many studies exploring stress detection from speech signal using various techniques and algorithms.

In [11], the authors used a combination of spectral and prosodic features of speech signals and was trained and tested on a publicly available dataset of stress speech. The results showed that the proposed model achieved an accuracy of 92.4% in detecting stress from speech signals, outperforming other state-of-the-art models. The study [12] highlights the potential of using deep learning models such as CNN and BiLSTM for stress detection from speech signals and provides valuable insights for future research in this area. In [13], the authors proposed a model that combines deep learning and transfer learning techniques to detect stress from speech signals. The

model was trained on a dataset of stress speech and achieved an accuracy of 91.4%. In [14], the authors proposed a model that combines Convolutional Neural Networks (CNN) and attention mechanism to detect stress from speech signals. The model was trained on a dataset of stress speech and achieved an accuracy of 87.6%.

In [15], the authors proposed a deep neural network in order to train stress representation from video input data. Some salient features from the data have been obtained to measure the intensity level of stress. In [16], the authors proposed a technique based on computer vision in which speech signal has been transformed into histogram representation in which variation of signal frequencies are observed. The methodology uses linear regression model to implement the prediction of the speech signal as stress and non-stress. The algorithm has been applied over AVEC 2014 stress dataset. The detection of depression, stress and anxiety depends of the variation of intensity of vocal cords [17]. The stress has least variation and the depression has the most variation that can be seen from frequency representation of vocal cords. The classification model such as SVM, CNN, RNN, etc has been tested over speech signal in the past decade [18]. Mostly, stress recognition is observed through the facial activities in the past work. There are some demerits in the detection of depression and stress from facial expression. The facial expression can be easily controlled and may be confused with other emotional activities [19]. The audio analysis of an individual is found to be very robust for the identification of emotional characteristics. Stress recognition may be easily confused with sad emotions by facial dataset. The working on facial dataset for the stress detection is found to be very obsolete and prone to achieve various kinds of errors.

The speech signals are more reliable and authenticated that contains the exact psychological activities from an individual's mind. Frequency components are divided from a frame of speech signal and hence observed for the variation [20]. The correlation of such variations has been seen with set of questionnaires or the activities that is provided to the individual during the interrogation. Computer vision makes the recognition of stress more efficient and cost efficient as compare to the medical investigation. The medical exploration is time consuming and quite expensive. In [21], the authors implemented an automatic detection of emotional states using facial cues. The approach uses the movement of eyes, mouth and head to justify the state of emotions in an individual. This algorithm discovers the short-term stress state. Hence, this method may fail to discover depression which is a long-term mental disorder. In [22], the authors extracted middle level of facial features containing action units that describe pose and movements. This work was conducted using video dataset from which the facial feature units are extracted. The method was applied on AVEC 2016 video dataset in which facial action units are tracked and extracted out. Recent studies have shown the relationship between the changes in the personality with the mental health. The correlation between these can give the understanding of disorder in an individual. Body may release multiple biological hormones that bring changes in the action units of a subject. These action units are tracked out frame to frame by techniques such as Kalman filter, viola jones, etc. for the investigation of level of mental disorder. In [23], the authors found the effect of non-melancholic depression event on the personality disorder state. The study shows how the personality is associated with depression and stress. They also show the influence of depression on person's temperament and personality. The temperament inventory comes out in the form of facial action units that may reflect the level of stress. The algorithm uses standard anxiety and depression scale to measure the level of stress in an individual. This scale depends on the personality features like reward dependence, persistence, self-directedness, cooperativeness. These characteristics are translated into score point generated by scale that shows the level of stress and depression. The score shows that if the neuroticism traits are low in an individual then there could be less depression. The scale also measures the extra version traits that come out through the regular exercises. The research shows that the person has less stress factor if his/her extra version states are strong.

In [24], the authors implemented the stress recognition techniques on 632 subjects for 1 year and founded that a person has much effective with stress that has low self-directedness, low cooperativeness and high harm avoidance capability. These features are the reflected by personality of an individual. In [25], the authors also investigated the influence of stress state on a person's personality. The study uses personality traits such as conscientiousness, neuroticism, extraversion etc. for the estimation of stress level. More the personality features are visible the more accuracy decision will be made by the algorithm. If personality shows less information, then it prediction becomes inaccurate. Hence, the detection of stress from personality features is quite obsolete and less effective. The measurement of stress level in various patients can be done using some standard scaling parameters. In [13], the authors reveal the source of the mental stress of using Holmes-Rahe stress scale that made an analogy of events of life with the scaling factors. But this scale does not found to be very effective in the prediction of stress accurately.

In the proposed methodology, speech signals are recorded from standard microphone [9] against questionnaire dataset. The questionnaire dataset has been taken from Microdata library [9] in which some standard questionnaire has been prepared. The questionnaire has been taken in the form of text data. These questionnaires have been asked to 300 individuals and these responses are recorded in the form of speech signal. The questionnaire dataset contains standard questions related to the events of COVID 2nd phase of India. These questions are general and contribute in provoking people's attitude to gather the information their speech. The proposed model applied fast Fourier transform for the translation of speech signal into frequency coefficients. These frequency coefficients have analyzes for any variations. LSTM model has been used in this experiment that used to store frequency components and extracts hidden contextual information from it. The LSTM model stores long-term temporal features of speech signal. LSTM is a special type of recurrent neural network that shows long-term dependencies on various feature parameters. The basic unit of the LSTM model is memory block that stores the information of features. The model contains non-linear gating units that maintain every state of speech signal over the frames. The model is used to regulate the temporal information from one memory cell to other. The LSTM has three important gates i.e. input gate, forget gate and the output gate. Each gate is used to update its information. The algorithm is applied to extract the scalogram information. The proposed model also applied VGG 16 algorithm that is made of convolutional and pooling layers. The VGG-16 architecture is used to analysis the features of speech signal and classify them into stress and non-stress classes. The convolutional layer of the model break down the features into set of frames and train the model with them. The model analyses the variation in the scalogram and maps it with the cortisol hormones. The scalogram features are mapped with the frequency coefficients of the speech signals. The scalogram representation shows the variation of hidden features such as pitch, noise etc.

III MATERIALS AND METHODS

The flow chart of the suggested methodology is shown in Fig. 1, which comprises the modules for the classification of stress and non-stress features. The flow chart expresses the working modules of the models. According to the flow chart, the model first read the input speech signals of participants. These speech signals are taken from the responses gathered against the questionnaire dataset. Speech signals are transformed into frequency components by Fast Fourier Transform (FFT) technique. Then, the Long Short-Term Memory (LSTM) method has been applied for the feature extraction. Then, classification task has been done by using VGG-16 algorithm. We discussed the components of the proposed models as follows.

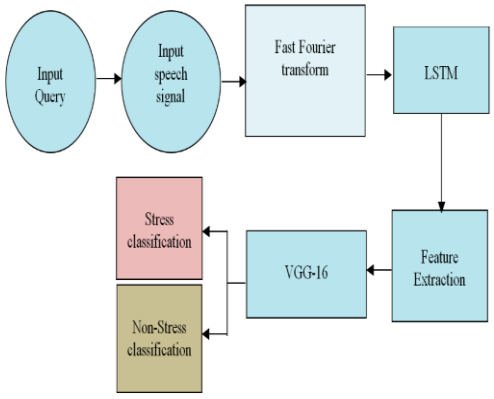

Figure 1: Proposed methodology

Dataset

Identify applicable sponsor/s here. If no sponsors, delete this text box (sponsors).

We have shown the questionnaire dataset in Tab. 1. Therein, the sample questions from the standard dataset are displayed. We have taken the sample of the speech of one of the participants against the one of the questions from questionnaire dataset. The amplitude variations of the host audio are observed against the time samples. The responses are recorded in the form of audio signals for which frequency coefficients are calculated.

Table 1: Dataset of the questionnaire

Source of Stress	Never(0)	Almost Never (1)	Some-times (2)	Fairly Often (3)	Very Often (4)
Have your personal plans been changed or affected by the COVID in the last 7 days?	58	62	81	64	35
Did you do any work for pay during COVID at a job or business in the last 1 month?	65	86	51	48	50
Think about your life for past 30 days, how do you think; was it a really tough time?	78	58	63	71	30
Have you currently covered by any of the types of medical insurance or any health plan?	59	68	75	62	36
Have you had any friend or member who is very close to you die from COVID-19 or recovered from 2nd wave of COVID.	71	86	64	30	49
How did you communicate with your family, friends and relatives during the last 1 month? Did you use any phone, text, email, or Internet something to connect with you close one?	65	37	79	66	53
After 1 march 2020, what did you think about the change in your life Did you talk with any of your neighbors?	89	58	67	48	38
After the second wave of COVID, did you voluntarily appear in any organization or association as a part of helping hands?	67	86	65	54	28
In the past 1 month, have you ever felt nervous, anxious, sad, lonely etc.	65	75	60	54	46
Do you feel any slackening in health sector in the past 1 month?	67	58	37	84	54
Did lockdown seriously affect your job or financial source?	69	54	67	82	28
Have you realized any major changes in your life after 1 march 2020?	37	48	82	64	69

The FFT Method

The FFT (Fast Fourier Transform) performs the translation of the signal data into frequency representation where the data variation can easily be observed and maintained. The proposed methodology has applied FFT technique

to the speech signal taken from participants to generate the frequency coefficients. These frequency coefficients are non-overlapping components in which each frequency band contains the unique set of information. Then magnitude of the information has been represented in the form of frequency bands. FFT is the modified version of the discrete Fourier transform (DFT). In DFT technique, frequency components are also estimated in the same way as happened in FFT. But FFT technique is much faster and efficient for the generation of non-collapsing energy components. The higher and lower frequency bands are important as it contain maximum information of the speech signal. These energy bands are utilized for the analysis further towards feature extraction as follows,

$$F[a]= \sum_{p=0}^{n-1} F[p]e^{-j2\pi ap/n} \quad (1)$$

where $e^{-j2\pi ap/n} = -j\sin\left(\frac{2\pi ap}{n}\right) + \cos(2\pi ap/n)$,F[a] represents the input matrix 'a' in the frequency domain. Although it is comparable to DFT, its computation time is faster than DFT at O(Nlog2N). The intensity of a signal matrix's pixel values is represented by its frequency coefficients. The information of the signal is included in the frequency bands, which take the shape of real and imaginary components. We can calculate the energy of each frequency band as follows,

$$E = \sum_I f(I)^2 \quad (2)$$

where$f(I)^2$is the square function of the energy band's intensity value, and E stands for the energy band's magnitude.

The LSTM Method

Recurrent neural networks (RNNs) with LSTM (Long Short-Term Memory) architecture are frequently employed in tasks involving sequence-to-sequence prediction and natural language processing. The basic LSTM cell consists of three main components named, input gate, forget gate, and output gate. The Input gate determines which values from the input sequence must be used to update the cell state. The forget gate describes which information from the previous cell state should be forgotten, while the output gate describes which values from the updated state of cell should be output to the next layer or to the final prediction. An LSTM cell also includes a cell state, which serves as a kind of memory. The input and forget gates are used to notify the cell state, which is then transmitted through the output gate to generate the desired output. LSTM networks can be stacked to create deeper models, and can also be bidirectional such that they can process input sequences both forward and backward in time. Therefore, LSTM are especially helpful for jobs like translating languages, in which a sentence's meaning may vary depending on the context of the whole section.

D. The VVG Architecture

The VGG architecture has several variants including VGG-16 and VGG-19, which have 16 and 19 layers, respectively. The max-pooling layer and the ReLU activation function come after each convolutional layer in the VGG architecture. The convolutional layers are designed with a small receptive field of 3x3 pixels and a stride of 1 pixel. The feature maps' depth is maintained while their spatial dimensions are shrunk thanks to the max-pooling layers. Fig.1 shows the architecture of VGG 16, in which convolutional layers are represented with "Convl" followed by their numbers. It represents audio signals is to use the Mel spectrogram, which is an emblem of the frequency content (FC) of a signal against thetime. The Mel spectrogram can be used as the input to the modified VGG 16 network with the number of Mel frequency bands representing the input channels. VGG 16 has been used to extract features from the audio data, which can then be used for classification. The VGG 16 architecture can be adapted for audio classification by modifying the input layer to accept audio signals instead of images. This technique utilizes 16 layers of CNN model for the better recognition of features.

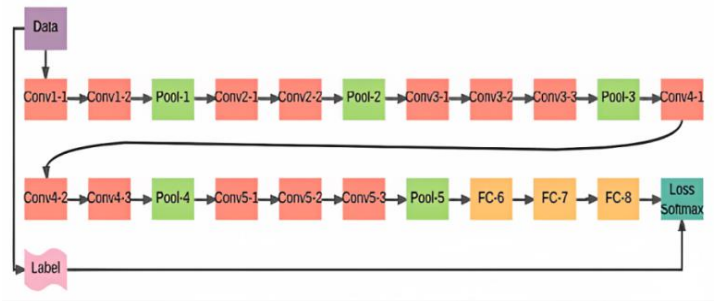

Figure 2: Architecture of VGG-16

IV RESULTS AND DISCUSSION

We have used the metric of Receiver Operating Characteristic (ROC)curve for the training and validation proof of our dataset. In Fig. 3, the graph has been plotted between loss and the epoch. The training and validation speech signal has been iterated multiple times in order to measure the efficiency of the proposed model. We can see that the validation loss remain lower than that of the train loss, which shows the validity of the proposed model.

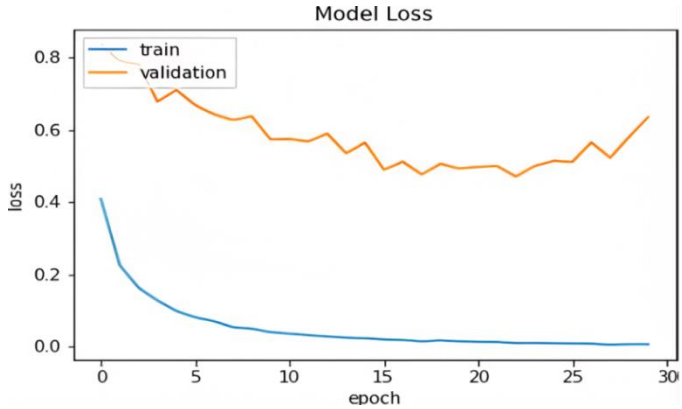

Figure 3:ROC curve for both validation and training

The FFT plot is shown in Fig. 3, which contains the representation of normal speech signal into frequency components. The FFT component translates the speech signal into frequency coefficients that shows the variation against the questionnaire. On the other hand, Fig. 4 shows the FFT plot for the stress sample speech signal that shows the inconsistent variation of frequency of speech signal. This is because of the fact that normal that normal speech signal has constant variation and consistency. However, the stressful speech signal has abrupt variations against unusual questionnaire. Fig. 5 shows the two-dimensional (2D) representation of stress audio signal. Inconsistent variation can be seen in 2D representation of stress signal.

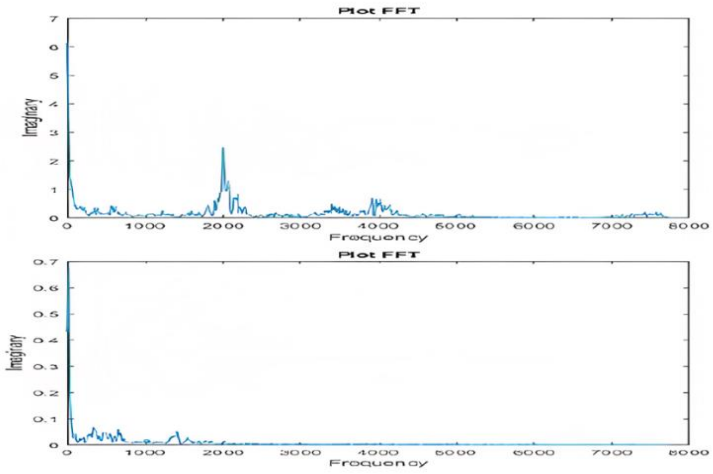

Figure 4: FFT plot of normal audio

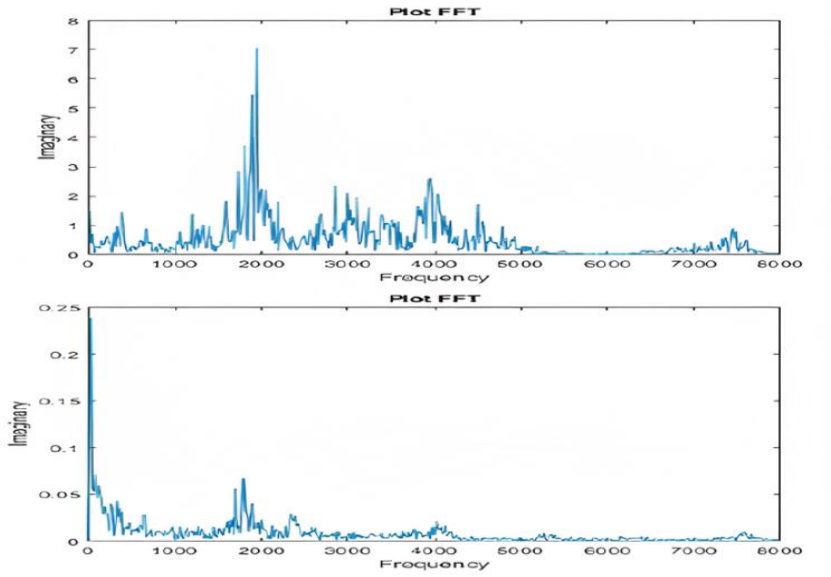

Figure 5:FFT plot of stress audio

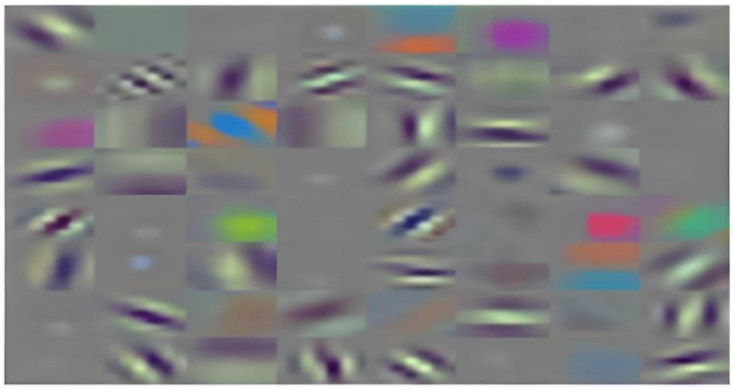

Figure 6:Representation of stress audio in 2D form

Figure 6 shows the scalogram representation of stress speech signal. This scalogram representation shows the variation of hidden features such as pitch, noise etc. The scalogram of stress speech signal varies from one frame to another frame. All the hidden features are visible in first scalogram. However, in the other scalogram the hidden features are less. This shows the inconsistency of the person's speech against the various questionnaires. As the scalogram contains the variation of hidden features so if the variations become inconsistent then it shows the symptom of stress. Fig. 7 shows the ROC curve for the accuracy and the error. The upper part of the figure shows the true recognition of speech into the stress class. The loss curve has been mentioned at the below of the figure that shows the failure achieved by model against the misclassification of the speech signal. The accuracy of the suggested work is found as 98%. The false acceptance rate and the false rejection rate are very less. The proposed model classified the features of speech signal into stress and non-stress classes. The model is robust as it also analyses the scalogram of the speech signal. VGG-16 architecture performs deep analysis of the variation of speech signal according with respect to questionnaire. The model is able to predict cortisol level. The cortisol is a type of hormone which is able to response against stress or irregular blood flow. The proposed model establishes the correlation of speech variation with the cortisol hormone.

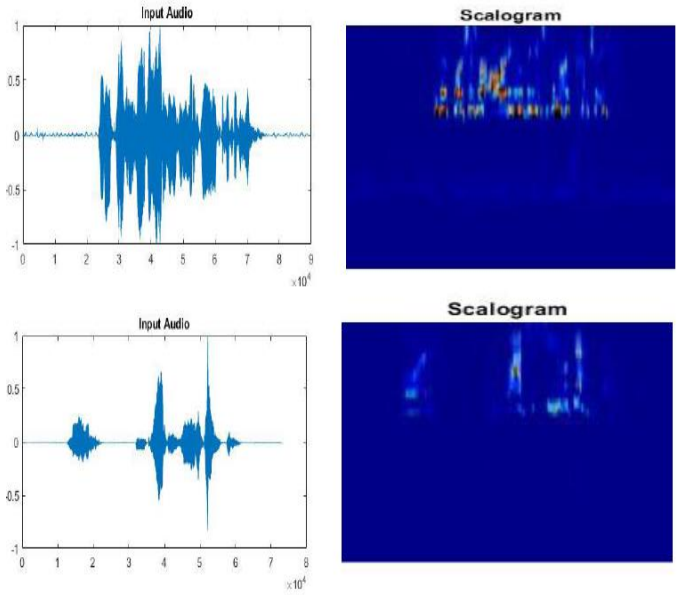

Figure 7:Scalogram of stress Audio

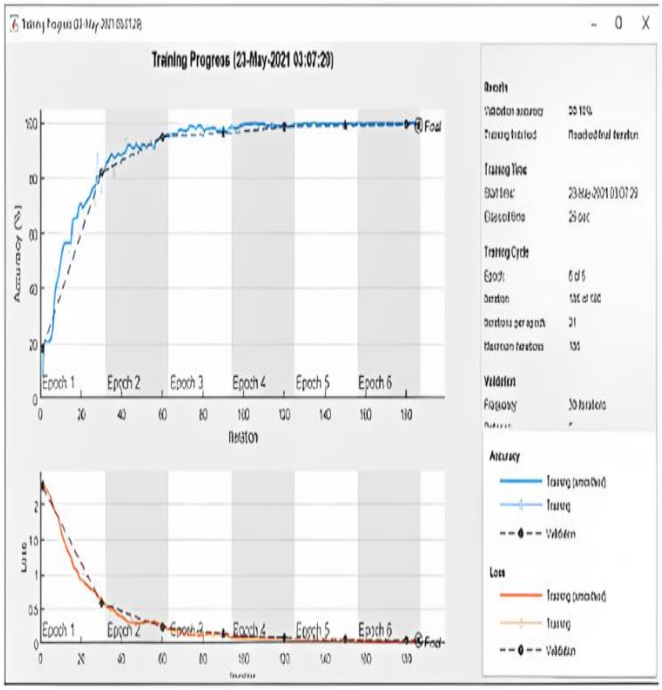

Figure 8:ROC curve for an input speech signal

The level of cortisol hormone becomes high when a person is in stress. In the proposed system, the questionnaire set provokes the person and so the variation in the speech can be recorded and analyzed. Fig. 8 shows the plot of stress and non-stress features of the speech signal. The blue color dot in the figure signifies the stress individuals and the red color dot characterizes the non-stress individuals. We can that that the model is able to clearly distinguish stress and non-stress features, which proves the validity of the proposed model.

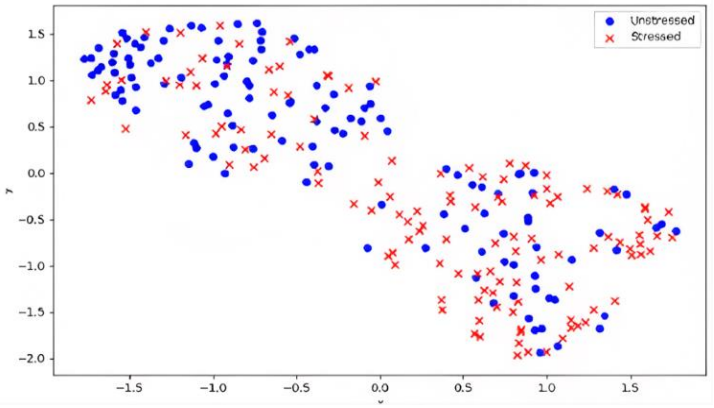

Figure 9: Virtualization of features for stress and non-stress classes

Fig. 9 demonstrates the graphical user interface (GUI) and working of the proposed model. It shows the input speech signal, VGG 16 architecture, 2 D representation of input speech signal and the scalogram of the signal. The GUI shows the accuracy rate of 90% for the sample input signal. The GUI contains some controls also on the left side that can help to effectively use this model. In Tab. 2, we preset an evaluation compression of our proposed models with the related models.

Table 2: Comparative evaluation of the suggested model

Authors	Dataset	Findings	Algorithm	Accuracy
Gupta et. al [24]	Self-made speech dataset	Stress and non-stress	LSTM-RNN	66.4%
P. Vitasari et al. [25]	AVEC dataset	Identification of Depression	ResNet-50 model	93.67%
Kejriwal et al. [26]	AVEC 2013, AVEC 2014	Depression and non-depression	ResNet-101 algorithm	92.23%
Liu et al. [27]	DAIC-WOZ dataset	Identification of Anxiety	LSTM Algorithm	91.54%
Kuchibhota et al. [28]	DAIC-WOZ dataset	Identification of depression disorders	C-CNN-AVL	93.84%
Xefteriset al. [29]	AVEC 2014	Identification of depression and anxiety disorders	Bi-LSTM Algorithm	89.64%
Hyewon H. et al. [30]	Oz (DAIC-WOZ) database	Analysis of stress disorder	Ensemble-50 1D CNN	72%
Proposed Scheme	Micro-Library[9]	Stress	LSTM-VGG16	98%

V. CONCLUSION

The suggested approach successfully distinguishes between characteristics in the stress and non-stress groups. The suggested approach successfully used pre-processing methods to improve voice signal feature characteristics. These methods and filters enhance feature quality of audio signals. The frequency coefficients are extracted from speech using FFT algorithm. Then, feature extraction has been performed by using LSTM algorithm. Each feature components are stored into temporal frames. The features are further classified using VGG 16 feed-forward convolutional neural network architecture that uses pooling layer to further disintegrate the features. The correlation of these features are calculated and mapped with the class. Hence, the suggested model provides the robust classification of psychological events into stress and non-stress classes. The proposed model is able to obtain 98% of stress detection accuracy. The VGG-16 model is thus determined to be extremely effective in the provided feature space and also resistant against any inaccuracy, according to the recommended method. The study can be expanded in the future to incorporate the recognition of a variety of different emotions utilizing the suggested algorithms to carry out greater accuracy rates. The suggested approach can be expanded in the future to assess other emotional states that are related to speech signals, such as happiness, rage, excitement, etc. The model's robustness can be increased by using further resilient neural network algorithms

VI ACKNOWLEDGEMENT

The author thanks Natural Sciences and Engineering Research Council of Canada (NSERC) and New Brunswick Innovation Foundation (NBIF) for the financial support of the global project. These granting agencies did not contribute in the design of the study and collection, analysis, and interpretation of data.

Funding Statement: The author(s) received no specific funding for this study".

Availability of Data and Materials: The data underlying this article will be shared (after the patent of the underlying research is filled) upon reasonable request to the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] J.F.Cohnet al.,"Detecting depression from facial actions and vocal prosody,"inProc.IEEE ACII, Amsterdam , Netherlands, 2009, pp. 1–7.

[2] S. Alghowinemet al., "Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behav- iors,"IEEE Trans. Affect. Comput. vol. 9, no. 4, pp. 478–490, 2018.

[3] T. S. Wingenbach, C. Ashwin and M. Brosnan, "Correction: Validation of the amsterdam dynamic facial expression set–bath intensity variations (ADFES-BIV): A set of videos expressing low, intermediate, and high intensity emotions,"PloS One, vol. 11, no. 12, p. e0168891, 2016.

[4] E. B. SÖNMEZ, "An automatic multilevel facial expression recognition system,"J. Appl. Nat. Sci. , vol. 22, no. 1, pp. 160-165, 2018.

[5] Afzali, A. Delavar, A. Borjali and M. MIRZAMANI, "Psychometric properties of DASS-42 as assessed in a sample of Kermanshah High School students," Int. J. Behav. Sci.,vol. 5, no. 2,pp. 81-92, Aug. 2007.

[6] N. Bayram and N. Bilgel, "The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students," Soc.Psy.Epidy., vol. 43, no. 8, pp. 667–672, 2008.

[7] R. Li, andZ. Liu, "Stress detection using deep neural networks," BMC Medical Inform. Decis. Mak., 20, 1-10, 2020.

[8] Gaballah, A. Tiwari, S. Narayananand T. H. Falk, "Context-aware speech stress detection in hospital workers using Bi-LSTM classifiers,"in Proc. ICASSP,Toronto, Canada, 2021, pp. 8348-8352.

[9] G. Shanmugasundaram, S. Yazhini, E. Hemapratha, and S. Nithya, "A comprehensive review on stress detection techniques,"in Proc.IEEE ICSCAN, Amsterdam, Netherland, 2019, pp. 1-6.

[10] S. Cohen, "Perceived stress in a probability sample of the United States,"inThe social psychology of health , 1st ed., NV, USA: Sage Publications, Inc. 1998, pp. 31–67.

[11] Y. Jiaoet al."Feasibility study for detection of mental stress and depression using pulse rate variability metrics via various durations,"Biomed. Signal Process. Control., vol. 79, 104145, 2023.

[12] M. S. Hilmyet al.,"Stress Classification based on Speech Analysis of MFCC Feature via Machine Learning,"in Proc.IEEE ICCCE, Kuala Lumpur, Malaysia, 2021, pp. 339-343

[13] X. Zhanget al.,. "Perceived academic stress and depression: The mediation role of mobile phone addiction and sleep quality,"Front. Publ. Hlth., vol. 10, p. 66. 2022

[14] Zhanget al.,"Relationship between Depressive Mood and Parenting Style of Junior High School Students and Educational Countermeasures,"Int. J. Educ. Hum., vol. 6, no. 3, 61-65, 2023.

[15] G. M. Kalatzantonakis-Jullien, N. Stefanakisand G. Giannakakis, "Investigation and ordinal modelling of vocal features for stress detection in speech," in Proc. IEEE ACII, Cambridge, United Kingdom, 2021,pp. 1-8.

[16] J. D. Henry and J. R. Crawford, "The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample," Br. J. Clin. Psychol., vol. 44, no. 2, pp. 227–239, 2005.

[17] T. H. Holmes and R. H. Rahe, "The social readjustment rating scale," J. Psychosom. Res., vol. 11, no. 2, pp. 213–218, 1967.

[18] N. P. Dhole, S. N. Kale, "Stress Detection in Speech Signal Using Machine Learning and AI," in Proc. ICMLIP,Hyderabad, India, 2020, pp. 11-26.

[19] Kuppuswamy, "Manual of socio-economic status scale," Delhi: Manasayan Publication, 1st ed. 1962.

[20] L. Manea, S. Gilbody, and D. McMillan, "Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis," Can. Med. Assoc., vol. 184, no. 3, pp. E191–E196, 2012.

[21] H. Liu, L. Chen, M. Zhang, and H. Zhang, "A machine learning-based system for detecting stress and depression using speech features,"J. Ambient Intell. Humaniz. Comput., vol. 12, no. 5, 4449-4460, 2021.

[22] M. Aldosari, S. Islam, H. Almuhareb, R. Alharthi, and A. Alshehri, "Stress and Depression Detection from Speech Using Deep Learning Techniques,"J. Med. Syst., vol. 45, no. 9, p. 97

[23] M. Shah, S. Hasan, S. Malik, and C. T. Sreeramareddy, "Perceived stress, sources and severity of stress among medical undergraduates in a Pakistani medical school," BMC Med. Educ., vol. 10, no. 1, p. 2, 2010.

[24] M. Guptaand S. Vaikole, “Audio Signal Based Stress Recognition System using AI and Machine Learning,”J. Algebr. Stat., vol.13, no. 2, pp.1731-1740, 2022.
[25] P. Vitasari, M. N. A. Wahab, A. Othman, T. Herawan and S. K. Sinnadurai, “The relationship between study anxiety and academic performance among engineering students,” Procedia Soc. Behav. Sci., vol. 8, pp. 490–497, 2010.
[26] J. Kejriwal, S. Beňuš, and M. Trnka, “Stress detection using non-semantic speech representation,”in Proc.IEEEICR, Kosice, Slovakia,2022, pp. 1-5.
[27] W. Liuet al.“A longitudinal study of perinatal depression and the risk role of cognitive fusion and perceived stress on postpartum depression,”J. Clin. Nurs., vol. 32, n. 5-6, pp. 799-811, 2023.
[28] S. Kuchibhotla, S. S. Dogga, N. G, Thota, G. Puli, M.S.R. Niranjan and H.D. Vankayalapati, “Depression Detection from Speech Emotions using MFCC based Recurrent Neural Network,”in Proc. IEEE ViTECoN, Vellore, India, 2023, pp. 1-5.
[29] V. R. Xefteriset al.,“Stress detection based on physiological sensor and audio signals, and a late fusion framework: An experimental study and public dataset,”Res. Sq., 2023.
[30] H. Hyewon B. Kyunggeun and K. Hong-Goo, “A deep learning-based stress detection algorithm with speech signal,”in Proc. of ACM AVSU, New York, NY, USA, 2018, 11–15.