International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8178
***Article Received:** 24 June 2023 **Revised:** 28 August 2023 **Accepted:** 07 September 2023*

Automated Video and Audio based Stress Detection using Deep Learning Techniques

Dr. Deepali Godse[1], Nilofar Mulla[2], Dr. Rohini Jadhav[3], Milind Gayakwad[4], Rahul Joshi[5], Kalyani Kadam[6], Jayashree Jadhav[7]

[1]Bharati Vidyapeeth's College of Engineering for Women,
Pune, Maharashtra, India
deepali.godse@bharatividyapeeth.edu
[2]Bharati Vidyapeeth's College of Engineering for Women,
Pune, Maharashtra, India
nilofar.mulla2006@gmail.com
[3]Associate Professor, Bharati Vidyapeeth (Deemed to be University),
College of Engineering, Pune, India
rbjadhav@bvucoep.edu.in
[4]Assistant Professor, Bharati Vidyapeeth (Deemed to be University),
College of Engineering, Pune, India
mdgayakwad@bvucoep.edu.in
[5]Symbiosis Institute of Technology, Pune,
Symbiosis International (Deemed University), Pune, India.
rahulj@sitpune.edu.in
[6]Symbiosis Institute of Technology, Pune,
Symbiosis International (Deemed University), Pune, India.
kalyanik@sitpune.edu.in
[7]Bharati Vidyapeeth's College of Engineering for Women,
Pune, Maharashtra, India
jayashree.jadhav@bharatividyapeeth.edu

Abstract— In today's world, stress has become an undoubtedly severe problem that affects people's health. Stress can modify a person's behavior, ideas, and feelings in addition to having an impact on mental health. Unchecked stress can contribute to chronic illnesses including high blood pressure, diabetes, and obesity. Early stress detection promotes a healthy lifestyle in society. This work demonstrates a deep learning-based method for identifying stress from facial expressions and speech signals.An image dataset formed by collecting images from the web is used to construct and train the model Convolution Neural Network (CNN), which then divides the images into two categories: stressed and normal. Recurrent Neural Network (RNN), which is used to categorize speech signals into stressed and normal categories based on the features extracted by the MFCC (Mel Frequency Cepstral Coefficient), is thought to perform better on sequential data since it maintains the past results to determine the final output.

Keywords- Convolutional neural network; Deep learning; Image Processing; Recurrent neural network; Stress detection.

I. INTRODUCTION

The emotional states of a person, such as stress and worry, have considerable effects on his or her standard of living. Since stress is one of the main factors contributing to serious chronic health issues, stress management is crucial in the modern world to maintain a low stress level and lower health risks. Mind-set, feeling, character and other down to earth data about the condition of the speaker are available in each verbally expressed expression. Currently, interest in exploration is growing in this area as the number of potential applications grows and verbal emotions have also generally been concentrated in absence. A discourse signal contains references to the speaker in about 25% of its data [1].

Routinely, mental and physiological subject matter experts conclude pressure state of a singular utilizing poll-based pressure investigation. This approach conveys parcel of vulnerability also, is inconsistent as it relies altogether upon the people's reactions and individuals will be faint to answer the poll [2].

The acknowledged term for discourse signals conveying data on the speaker's physiological pressure is "focused discourse". Both internal (emotion, weariness, etc.) and external (noise, vibration, lack of sleep, etc.) elements can contribute to stress. The physiological effects of stress include, but are not limited to, changes in the heart rate, respiratory rate (such as faster, more erratic breathing), changes in the musculature (such as greater muscle tension), etc. [1].

Many examinations were embraced to inspire and distinguish pressure, in light of an individual's physiological boundaries. A circumstance can be unpleasant when something mental, for example, tireless stress over losing an employment, moving toward work cutoff time, and so on., occurs. Such a stressful situation might release a cascade of stress hormones that cause major physiological changes like a racing heart, revived breath, tense muscles, the formation of sweat globules, and so on. These physiological changes lead to the body's physical (or "instinctive") response. Affected individuals emit corresponding biosignals during these physiological changes.

These biosignals help to identify pressure by measuring the physiological proportions of a person. Different physical sensors were utilized for this motivation behind programmed pressure Identification [2].

Stress is more pervasive and intense than ever before in today's society. Unmanaged stress could be a factor in a number of health issues, endangering people's moods, thoughts, behaviours, and general well-being. The ability to recognize stress can enable people to actively manage it before negative effects occur.

II. LITERATURE SURVEY

Although several researchers have developed methods for detecting stress, but they were primarily concerned with doing so through the use of speech or facial expressions. Deep learning, a branch of Artificial Intelligence (AI), is used to distinguish pressure using a collection of informative metrics including electrodermal action, skin temperature, and pulse estimations, unease, and stress [3].

It is recommended to use a cutting-edge feeling recognition model built on recurrent neural networks (RNNs) that takes into account the conversation's context and the emotional states of each party. For the purpose of extracting features from conversational sound data, a bag of words (BoW) -based methodology is provided [4].

In order to demonstrate a two-leveled stress detection network (TSDNet), the facial articulations and activity motions of the clients in the video were examined. TSDNet first individually learns face and action-level descriptions before integrating the findings with a stream-weighted integrator with local and global considerations for recognizing stress [5].

Using multimodal datasets collected from wearable physiological and movement sensors, various AI and deep learning techniques have been proposed for stress discovery on individuals. These techniques can keep a person informed about various pressure-related medical conditions [6].

The model that forecasts which subject is related to the textual data is put forward. These models allow for the online emotion detection of users. These feelings are also examined in order to understand stress or sadness. In summary, the Bidirectional Encoder Representations from Transformers (BERT) model and Machine Learning (ML) models have very high detection rates. The benefits of this research can be shown in terms of mental wellness. The outcomes, which are assessed using different metrics at the macro and micro levels, show that the trained algorithm can identify the emotional condition based on social interactions [7].

A study talks about a way to find expressions of tension and relaxation in tweeter datasets, such as feeling investigation to find feelings or sentiments about day-to-day life.

Examining feelings involves the automated extraction of information about feelings from text. Here, the author used the TensiStrength framework to distinguish between feeling strength and informal English text on person-to-person communication destinations. TensiStrength is a way to measure how well social media instant messages show stress and ease. TensiStrength uses a lexical method and a number of other principles to find direct and indirect expressions of tension or relaxation to give the treatment while reducing strain. [8].

The high-level sign handling, considering the Galvanic skin reaction, blood volume, understudy development, and skin temperature, is essential for structure work on pressure identifiable proof. The alternative approach to this problem also uses a few physiological indicators and visual cues (eye end, head advancement) to identify a person's level of stress while they are working [2].

III. MATERIALS AND METHODS

A. Proposed Methodology

Stress is identified in the proposed system using a small collection of supervised data.

For feature extraction and classification, a convolutional neural network is suggested, while an RNN is employed to identify speech stress. Real-time video of the user is captured by the system, and features are derived from the data. The system determines whether or not the person is stressed based on features extracted by the deep learning models as shown in Fig. 1.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8178
Article Received: 24 June 2023 Revised: 28 August 2023 Accepted: 07 September 2023

Dataset Gathering

Image dataset and audio dataset are required to train the CNN and RNN model respectively. Images on web are gathehered to form a dataset to train the CNN model. For model training 220 samples of images are used. While the model is tested using 30 samples.

The real time audio data is gathered to train the RNN model for speech recognition. Here, out off 205 samples 185 samples are used for training while 20 samples are used for testing the model.

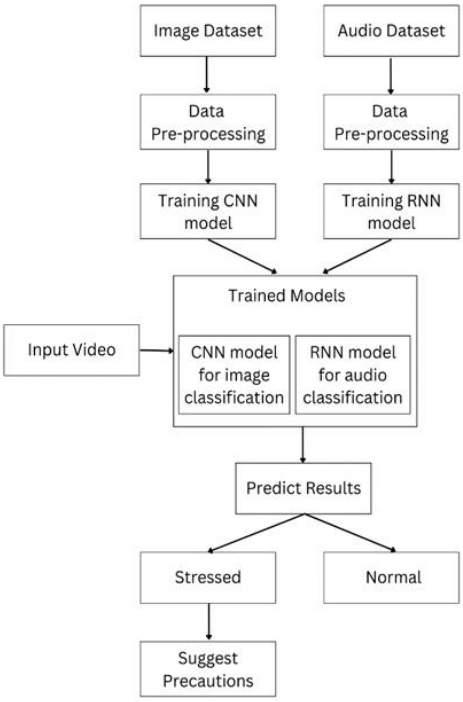

Figure 1. System Architecture

Data Pre-processing

The dataset gathered is first processed to eliminate the nonessential data such as background noise in audio data. Also the image dataset contains images of varying sizes which are resized into a size of 224 x 224.

Convolutional Neural Network (CNN)

The proposed method employs CNN to extract and classify image features in order to identify stress from facial expressions. The gathered data illustrating numerous facial expressions, such as sadness, anger, happiness, confusion, and surprise, is used to train the CNN model. Convolution layer, max pooling layer, fully connected layer, and output layer are the four layers that make up the CNN model [9], [10]. The fully linked layer is used for classification, the first two layers are utilized for feature extraction. The CNN model receives input images that have first been resized to 156 by 156 pixels. The photographs are sent to CNN in a batch of size 16. The image is converted into vector form by the Convolution layer using a 3x3 filter before being sent to the Max Pooling layer [11]. The filter is applied to the image within the specified size, and the dot product of the image's pixel value and the filter values which takes the form of a vector is calculated. The introduction of linearity, which is used for picture classification, is then made using the ReLU activation function [10], [11]. A kernel of size 2x2 slides over the vector picture in the pooling layer to extract the most noticeable feature from the area of the image covered by the kernel. This layer's primary function is to flatten the image's dimensions [10]. Edges, corners, forms, texture, and other features are a few that these layers' extract [11]. The flatten layer reduces the multidimensional vector produced by the convolution and pooling layers to a one-dimensional vector. To identify the photos, the fully connected layer uses the 1D vector as its input and correlates features to a certain category. The group of photos that fit into a specific category make up the CNN layer's final output. Fig. 2 includes the layered architecture of CNN.

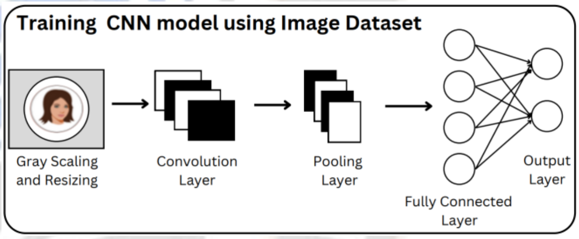
Training CNN model using Image Dataset
Gray Scaling and Resizing
Convolution Layer
Pooling Layer
Fully Connected Layer
Output Layer

Figure 2. CNN Architecture

In Fig. 3 the detailed flow of how features are extracted from the Convolution layer and the Max-Pooling Layer is shown. Later these features are used for image classification.

The image dataset containing 250 samples of stressed and normal images is used as input to the CNN model.

Training data(), Testing Data() Activation function(), Padding(), Loss(), Optimizer(), Epochs() are few measures that are used to evaluate the performance of the model.

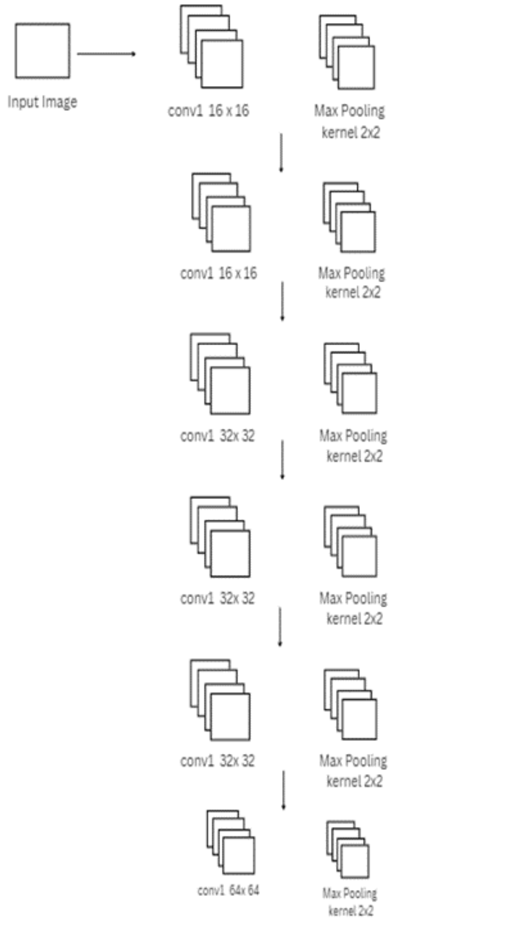

Figure 3. Feature Extraction

Following is the step wise mathematical representation of CNN model:

Feature extraction by convolutional layer():

$$x_{i,j}^{l} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \omega_{ab} y_{(i+a)(j+b)}^{l-1} \quad (1)$$

Max-pooling layer():

Size of input = N*N

Size of kernel = k*k

Output Block size = N/k*N/k (2)

Activation function():

$$f(x)=max(0,x) \quad (3)$$

Dense layer(): (4)

output=activation(dot(input,kernel)+bias)

Epochs():

Number of Steps per Epoch (5)
=(Total Number of Training Samples)/(Batch Size)

Recurrent Neural Network (RNN)

- The Recurrent Neural Network is considered best for the sequential data and hence it is used for extracting features from the audio signals [12]. RNN stores the output of each window in its memory which is used for calculating the final output or the overall output from the model [13]. Initially the audio signals are pre-processed to remove background noise by setting a threshold value. The frequency in the audio below the threshold value is considered as the background noise that is removed. Then the Mel Frequency Cepstral Coefficient (MFCC) is used for extracting features from the audio signals [14]. It has its filters that are used to identify the features and extract them. The features extracted by the MFCC are mainly Fourier transform (FT), Discrete Fourier transform (DFT), inverse discrete Fourier transform (IDFT), etc. These features are passed to the RNN model [14]. The RNN has layers such as Hidden Layer, Forget Gate and the Output Layer [15]. The hidden layer consists of neurons, the input features are multiplied with the weights and neurons to produce of an output. The forget gate erases the unnecessary data that is stored in the memory of the RNN which is not useful in calculating the final output. The final Output of the RNN is in the form of two categories as shown in Fig.4, the features that associate with the particular category are used for classification of emotions from the audio signals.

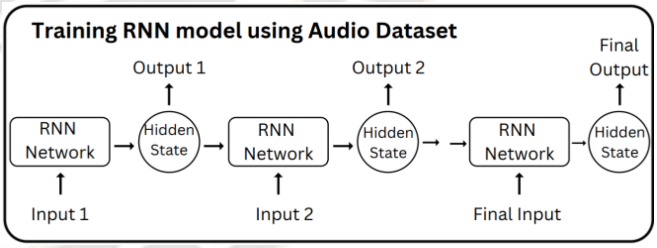

Figure 4. RNN Architecture

The primary cell introduces S0 with zeros or some irregular number in light of the fact that no past state is seen. U be one more grid of aspect dxn where d is the quantity of neurons in the main RNN cell and n is the information jargon size. W is another lattice whose aspect is dxd. b is inclination whose aspect is d*1. For finding the result from the main cell, another framework V is taken whose aspect is kxd where c is inclination with aspect kx1.

$$S_1=UI_1+WS_0+b \quad (6)$$

$$O_1=VS_1+c \quad (7)$$

B. Results and Discussion

- In the experimental setup, the total 205 speeches of real-time data are used. These speeches go through RNN framework by following feature extraction using MFCC module. The developed system gives the accuracy of 87.31% at 300 epochs for RNN model and 94.45% at 100 epochs for the CNN model.
- Table I shows the distribution of data used for training and testing the models. Out of the total 205 samples collected, 185 samples are used for training while 20 samples are used for testing the RNN model. Similarly for training the CNN model, 220 image samples are used while the model is tested using 30 samples that gives accuracy of 94.45%.

TABLE I. DATA DISTRIBUTION

Data Type	**Training**	**Testing**	**Accuracy**
Speech	185	20	87.31%
Images	220	30	94.45%

The graphs in Fig. 5 and 6 show the accuracy and loss plot of the system. Fig. 5 shows the number of epochs from 0 to 100 on X-axis and accuracy value on Y-axis. Similarly,

Fig. 6 shows the number of epochs on X-axis and loss value on Y-axis. From the graphs it is clearly visible that the system accuracy increases with the increase in number of epochs

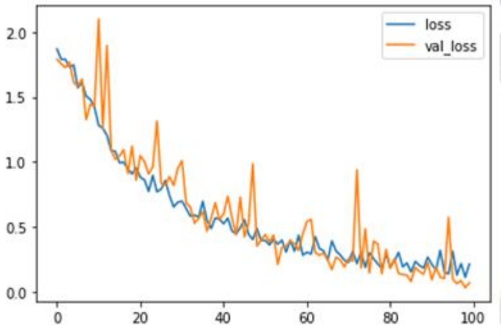

Figure 5. Accuracy plot

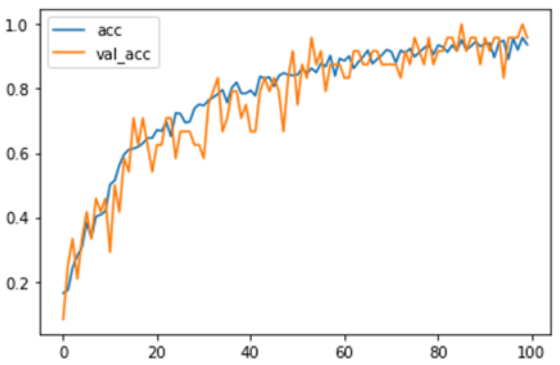

Figure 6. Loss plot

C. Conclusion

This paper contains a systematic review of previous research restricted for detecting stress either through facial expressions or audio signals which was conducted by the researchers worldwide. By analyzing both audio data and facial expressions, stress is identified with greater accuracy by the suggested approach.

In this project, stress is detected from video and audio using deep learning model which calculates the features and classify the stress to the respective category. In this paper neural networks are used to get the accuracy of 94.45% on 100 epochs and 87.31% at 300 epochs on CNN and RNN models, respectively. The model's performance can be enhanced in the future.

REFERENCES

[1] G. W. Evans and J. M. McCoy, ‘‘When buildings don’t work: The role of architecture in human health,’’ J. Environ. Psychol., vol. 18, no. 1, pp. 85–94, 1998. [10.1006/jevp.1998.0089]

[2] A. Liebl, J. Haller, B. Jödicke, H. Baumgartner, S. Schlittmeier, and J. Hellbrück, ‘‘Combined effects of acoustic and visual distraction on cognitive performance and well-being,’’ Appl. Ergonom., vol. 43, no. 2, pp. 424–434, 2012. [10.1016/j.apergo.2011.06.017]

[3] Sara Aritizabal, Kunjoon Byun and Nadia Wood, “The Feasibility of Wearable and Self-Report Stress Detection Measures in a Semi-Controlled Lab Environment,” IEEE Access, vol. 9, pp. 102053-102068, 2021. [10.1109/ACCESS.2021.3097038]

[4] SadilChamishka, Ishara Madhavi and RashmikaNawaratne, “A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling,” Multimedia Tools and Applications, vol. 81, pp. 35173–35194, 2022. [https://doi.org/10.1007/s11042-022-13363-4]

[5] Huijun Zhang and Ling Feng, “Video-Based Stress Detection through Deep Learning,” Sensors, vol. 20, 2020. [https://doi.org/10.3390/s20195552]

[6] Ciabattoni L., Ferracuti F., Longhi S., Pepa L., Romeo L. and Verdini F., “Real-time mental stress detection based on smartwatch,” in Proc. 2017 IEEE International Conference on Consumer Electronics (ICCE), pp. 110–111, 2017. [10.1109/ICCE.2017.7889247]

[7] Mike Thelwall, “TensiStrength : Stress and relaxation magnitude detection for social media texts,” Elsevier Information Processing and Management, vol. 53, pp. 106-121, 2016. [10.1016/j.ipm.2016.06.009]

[8] Russell Li and Zhandong Liu, “Stress Detection using Deep Neural Networks,” in Proc. The International Conference on Intelligent Biology and Medicine (ICIBM), vol. 20, 2020. [10.1186/s12911-020-01299-4]

[9] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do and Kaori Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into Imaging, vol. 9, pp. 611–629, 2018. https://insightsimaging.springeropen.com/articles/10.1007/s13244-018-0639-9]

[10] Nadia Jmour, SehlaZayen and AfefAbdelkrim, "Convolutional neural networks for image classification," in Proc. International Conference on Advanced Systems and Electric Technologies (IC_ASET), pp. 397-402, 2018. [10.1109/ASET.2018.8379889]

[11] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6645-6649, 2013. [10.1109/ICASSP.2013.6638947]

[12] Aditya Amberkar, Parikshit Awasarmol, Gaurav Deshmukh and Piyush Dave, "Speech Recognition using Recurrent Neural Networks," in Proc. IEEE International Conference on Current Trends toward Converging Technologies, 2018 [10.1109/ICCTCT.2018.8551185]

[13] Harshawardhan S. Kumbhar and Sheetal U. Bhandari, "Speech Emotion Recognition using MFCC features and LSTM network," in Proc. 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), 2020. [10.3390/app11219897]

[14] Sudharsan.R, Hands-On Deep Learning Algorithms with Python, 1st ed. ,Packt Publishing, 2019. [https://www.packtpub.com/product/hands-on-deep-learning-algorithms-with-python/9781789344158]

[15] Introduction to Supervised Deep Learning Algorithms, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/05/introduction-to-supervised-deep-learning-algorithms/

[16] People feeling more stressed in India, 2022. [Online]. Available: https://www.statista.com/statistics/1320246/india-people-feeling-more-stressed-by-age/