Stress Detection in Women Using Speech Analysis

A. Sharada, R. Mamatha(✉), K. Meghana, and A. Monika

Department of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad, Telangana, India
{sharada,mamatha.kovuru}@gnits.ac.in, Kskmeghana1040@gmail.com, Allamonikajan03@gmail.com

Abstract. "One of the common issues that everyone deals with is stress. One of the most prevalent emotional states in people is stress. Positive stress can motivate women to make important achievements. Yet, stress may also be damaging and destructive, negatively affecting many facets of one's life. Stress makes it more challenging to mature as it becomes extreme or chronic. Women are more prone to exhibit a wide range of emotions. Heart rates, blood pressure, and skin temperatures may now be monitored via smartwatches. Still, measuring stress-related biomarkers requires an individual to provide a sample of their blood or other bodily fluid. Voice Stress Analysis (VSA) is a field of study examining how the brain responds to human stress by listening to women's voices while under stress. There is a lot of equipment available for detecting stress by keeping an eye on skin temperature, blood pressure, and heart rates, but for the most part, determining stress-related biomarkers still requires some degree of invasiveness. Given that physiological measurements have several limits, speech analysis makes stress assessment more alluring, especially given that it is currently both inexpensive and non-intrusive. With an Android app, stress can now be measured via speech signals that use software and artificial intelligence to analyze the pitch, jitter, energy, rate, frequency, length, and number of pauses among other speech and speech acoustic characteristics. The on-call dialogues are the main data source sent to the program, revealing whether the woman is under stress or not. Thus, utilizing speech using speech to measure stress has potential, but important validation, privacy, and ethical concerns must be suggested system can assist women by automatically identifying their stress from ordinary speech without any additional aid.

Keywords: Stress · Voice analyzer · Artificial Intelligence · Speech acoustics

1 Introduction

Women are stereotypically more emotional than men, both in terms of outward displays of good emotions like happiness and internalizations of negative emotions like despair, dread, anxiety, humiliation, and guilt. Stress may push women to take action and get things done, but it can also negatively impact many aspects of their life. Yet, despite

B. Raj et al. (Eds.): ICETE 2023, AER 223, pp. 797–808, 2023.
https://doi.org/10.2991/978-94-6463-252-1_80

technological advances that make it feasible to measure physiological reactions to stress like heart rate, blood pressure, and skin temperature, identifying stress-related indicators is still mostly intrusive, and women, being inherently unselfish, typically don't take the time to do so. Using a speech analyzer to evaluate stress is simpler than utilizing physiological measurements. Vocal Stress Analysis (VSA) examines women's mental states and brain responses to stress via their voices. Since speech analyzers are non-invasive and inexpensive, this is especially true. This study seeks to understand how females cope with stress. An Android app can analyze a person's vocal signal to determine their stress level. These applications analyze pitch, jitter, energy, tempo, frequency, and pause length and quantity using computer algorithms and AI. The technique relies on phone conversations to determine whether the females are under stress or not. Measuring stress using speech has untapped potential, but validation, privacy, and ethical issues must be addressed first. The recommended method may assess women's stress levels simply by listening to their conversations. According to the findings, it was found that the CNN model achieved 85% accuracy, which is much greater than that of decision trees, RCNN, and other machine-learning algorithms.

2 Literature Survey

Hindra Kurnaiwan1, Alexander V, and others. Due to greater awareness of chronic stress's health risks and the development of non-intrusive stress-monitoring equipment, stress management research is growing. GSR and verbal signs may indicate experimenter stress. This article automatically detects acute stress using GSR and speech data. Classifiers based on GSR characteristics had substantially poorer accuracy, reaching just 70% in identifying light from the high workload [1].

Kevin Tomba1, Joel Dumoulin1 and others. This research uses speech analysis to assess HR candidates' stress during initial interviews. Machine learning can classify mean energy, intensity, and mfcc to identify speech stress. We train and test our classification algorithms using the Berlin Emotional Database (EmoDB), Keio University Japanese Emotional Speech Database (keioESD), and Ryerson Audio Database of Emotional Speech and Song (RAVDESS) Neural networks improved stress detection accuracy to 97.98% (EmoDB), 95.83 (keioESD), and 89.16% (RAVDESS) [2].

Dr. S vaikole S. Mulaikar and others. Stress affects biochemistry, physiology, and behavior. We created multi-step stress-detection models employing deep learning frameworks, CNN structures, and audio-visual data to identify stressed states in voice signals. Emotion-labeled tagged classification will measure stress (stressed vs. un-stressed). A stronger conceptual multimodal method may improve detection. Professional cortisol testing of each raw audio stage would improve experimental outcomes. All these factors will be considered to improve stress detection [4].

Anakha P S, Aishwarya Devi, and others. Managing stress increased pressure. This study presents a continuous stress disclosure framework with a face ID module that analyzes facial features and a conversation signal module that deciphers speech signals, allowing the client to participate in video conferences with easily available specialists. Recognizing pressure's impacts on happiness starts with the suggested strategy [5].

A. Baum Explains stress and its causes. Stress is an unpleasant emotional experience with predictable biochemical, physiological, and behavioral changes to adjust to the

stressor or its effects. Chronic stress exceeds long-term stress. Before a stressor subsides, reactions may habituate and persist. According to Three Mile Island and Vietnam veteran statistics, stress and biobehavioral changes continue. Images or thoughts of the stressor cause stress [6].

A.kene and S.Thakare. Stress now. The WHO believes stress harms. Everyone's down. Everyone's stressed. Stress creates mental disease. Stress impacts thoughts, feelings, and actions. This article explored machine learning-based stress detection tests. PhysioBank stress levels. Gradient boost properly measured stress utilizing statistical feature selection and extraction. The model has an accuracy (83.33%), specificity (75%), sensitivity (75%), positive and negative predictive values (90%), error rate (16.66%), F1 Score (83.33%), and recall (75%). Gradient boost defeated KNN, Random Forest, and SVM. The model predicted stress [8].

J. Lee and L. Tashey. RNN(prop.)-ELM is the learning algorithm's system. RNN(prop.) solely classifies using high-level RNN features without the utterance-level classifier. RNN(prop.) performs similarly without the high-level features utterance-level classifier that considers temporal dynamics in the traditional system because RNN explicitly considers temporal dynamics. RNN(prop.)-ELM improves UA and WA measures by 12% and 5%, respectively [9].

Muray LR, Baber others. Summarize the 1995 ESCA-NATO Workshop on Speech Under Stress (Lisbon, Portugal). The authors define stress and describe different stress models that may be approved by the speech community based on Workshop discussions. Due to its widespread usage, stress has defied categorization. Stress models are unsophisticated and poorly understood. The authors recommend correlating stresses and strains. This field should also be studied (PsycINFO, 2018 APA) [13].

3 Proposed Approach

There are many stages involved as shown in Fig. 1, utilizing a CNN model to identify stressed speech. The audio is read and processed to produce a data frame for data analysis first. A waveform and spectrogram are shown to better understand the audio data. Any unnecessary pauses in the recording are subsequently cut out. The audio's frequency distribution is then determined by plotting and analyzing a Mel Power Spectrogram. Each audio file is given a name, and the data is checked for parity by plotting the distribution of emotions. Then, the librosa library is used to extract features from the audio recordings, and the data is divided into training and testing sets. After that, the data is processed and ready to feed into the CNN model. The audio data is transformed into Mel Spectrograms, which are then fed into the CNN model. Afterward, the CNN model is developed, coded, and educated using the data from the training set. To compare the model's results on the two datasets, a Train Valid Loss Graph is created. After the model is trained, it is written into a JSON file. The stress on the test data is then predicted using the trained model. The performance of the model may then be assessed by comparing the observed and anticipated values. To further assess the effectiveness of the model, a confusion matrix is utilized to display the total number of positive, negative, false positive, and false negative predictions. The model's output is "stressed" or "unstressed" depending

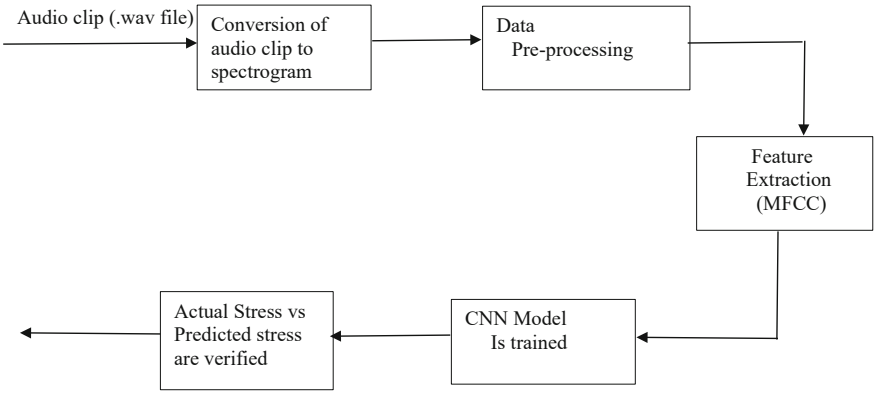

Fig. 1. Architecture of Stress Detection Model

on the projected value. To correctly identify stress in speech using a CNN model, one must first prepare data, extract features from that data, create a model, and train and evaluate that model.

3.1 Dataset Description

We used an audio-visual (voice and video) database called RAVDESS to study the effects of stress. A total of 24 actors' voices are included in the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset that can be found on Kaggle. All of the performers are in their twenties or thirties and come from different cultural backgrounds. There are a total of 7356 files in the dataset, with 24 performers each delivering two distinct vocal performances (neutral and expressive) and eight distinct emotional performances (happy, sad, furious, terrified, startled, disgusted, neutral, and calm). A professional sound booth was used to capture these high-quality recordings, which include a sampling rate of 48 kHz and a resolution of 16 bits. Each file is given its distinct name that describes the actor who recorded it, the mood conveyed, and the sort of speech or music that was used. Each file in the collection is accompanied by a spreadsheet with information such as the actor's age, gender, emotional category, and intensity. Speech analysis, emotion identification, and other machine learning applications like voice synthesis and cloning all make use of the RAVDESS dataset.

4 Experimental Analysis

4.1 Plotting Waveform

Waveforms make it easier to interpret sound signals. The vertical axis of a waveform depicts the amplitude of an audio source with time (the horizontal axis). A waveform can't be drawn without first being read from audio data by a software library or application. When being loaded into Python's Matplotlib, audio data may be shown as a waveform. The amplitude of the audio signal as a function of time is shown in the waveform plot.

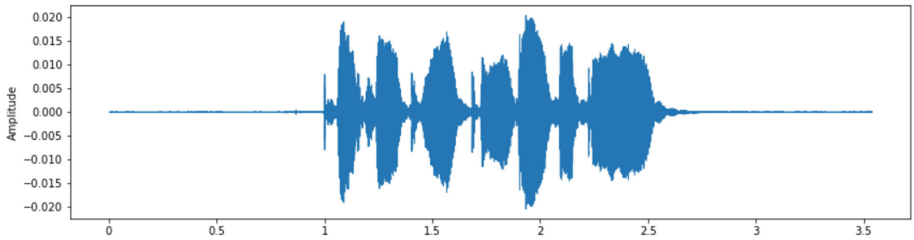

Fig. 2. Waveform of audio clip

The y-axis displays signal amplitude in volts or decibels, while the x-axis displays time in seconds or milliseconds. The audio signal may be seen in the waveform plot so that its frequency and duration can be understood. Signal patterns and sounds' beginning and ending points may be deduced from waveform graphs. The frequency content of a signal is not visible in a waveform plot. Audio stream frequency content may also be analyzed using spectral analysis. Finally, a waveform may be analyzed by plotting it on a graph to look for anomalies, discover the signal's characteristics, and spot any recurring patterns or trends. It displays a representation of the signal in the time domain but not its frequency content (Fig. 2).

4.2 Spectrogram

Audio spectrograms show frequency over time. A two-dimensional graphic shows frequency component intensity throughout time and place. Slicing audio into tiny, overlapping, time-bound bits produces a spectrogram. Fourier processing extracts frequency components from each window. Time and frequency are used to evaluate each frequency component's power. A two-dimensional spectrogram shows color intensity, frequency, and time. Higher-frequency colors have stronger colors. A spectrogram may show trends, patterns, and characteristics by tracing an audio stream's frequency content. They excel at understanding multi-frequency sounds like music and speech. Spectrograms can evaluate the frequency content of spoken words and detect formants, which help identify words and sounds. Spectrograms analyze, classify, and segment audio. Music analysis, audio compression, and noise reduction employ them. Finally, spectrograms show audio stream frequency over time. Audio analysis, processing, and categorization may use 2D frequency component strength or amplitude (Fig. 3).

4.3 Mel Power Spectrogram

Acoustic signal processing often uses a Mel power spectrogram. An audio stream's Mel-scale-normalized frequency content is shown over time. The human ear understands mel scale tones. Its non-linear frequency spacing is smaller at lower frequencies and coarser at higher ones. Mel power spectrograms improve sound perception by scaling the frequency axis to the Mel scale. Mel power spectrograms are created by segmenting the audio stream into overlapping constant-time frames. Fourier transforms recover window frequency components. Triangular overlapping filters convert frequency component intensity or

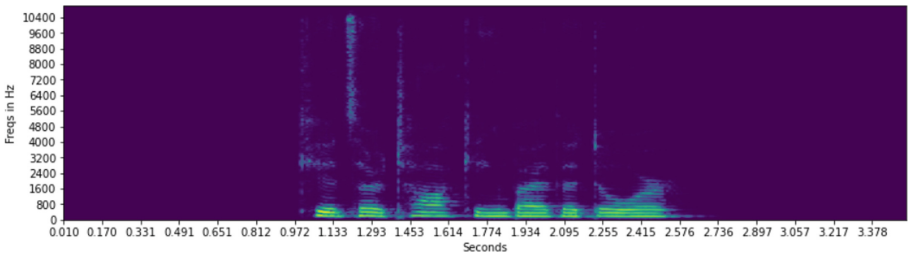

Fig. 3. Spectrogram

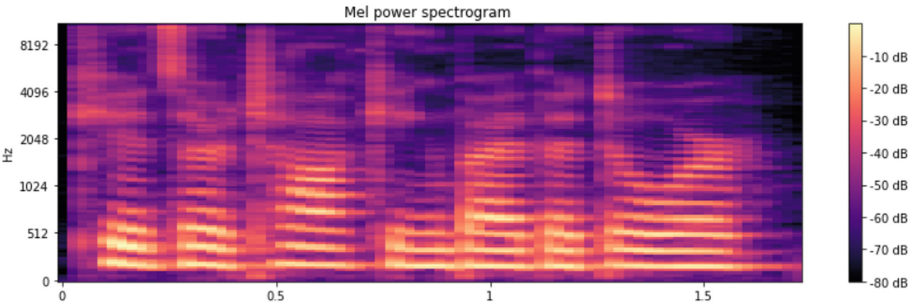

Fig. 4. Mel power spectrogram

magnitude to Mel scale. In the Mel power spectrogram, color intensity denotes frequency component strength while time and frequency are duration measures. Speech, music, and sound categorization employ mel power spectrograms. They may help distinguish speech and music. Mel power spectrograms may deconstruct words' tonal structure and extract formants by frequency in voice recognition. Lastly, a Mel power spectrogram, which scales frequencies to the Mel scale, may show an audio signal's frequency content with time. For audio analysis, processing, and classification, it better reflects human auditory perception (Fig. 4).

4.4 Mel Frequency Cepstral Coefficients

The Mel-Frequency Cepstral Coefficient (MFCC) is a popular representation of features used in audio and speech processing. They are a special kind of cepstral coefficient that accurately represents the spectral profile of an audio stream. The MFCCs are produced using a Mel power spectrogram, which is a spectrogram that employs a filter bank based on the Mel scale, to analyze an audio signal. The Mel power spectrum is logarithmized, and then discrete cosine transformed (DCT) to provide a collection of cepstral coefficients. The resultant MFCCs find widespread usage in a variety of audio-related applications, including voice recognition, speaker identification, and music analysis. In addition to the regular MFCC, the delta MFCC, the double delta MFCC, and the delta-delta MFCC are all variants that may give extra-temporal information about the audio signal. Generalizing, MFCCs' strength as a tool for audio signal processing and analysis stems from their capacity to record the signal's spectral profile (Fig. 5).

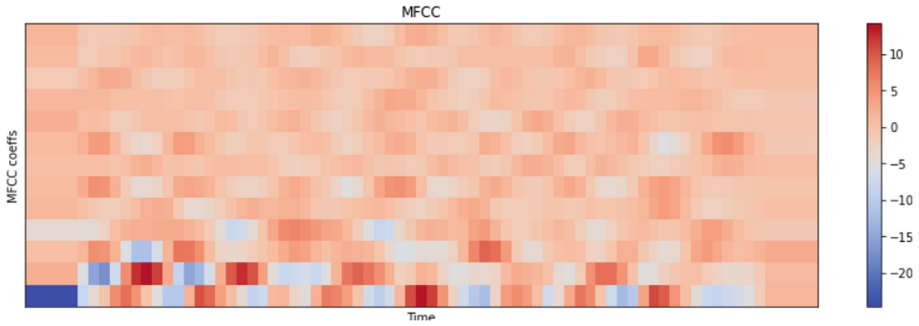

Fig. 5. Mel Frequency Cepstral Coefficients

4.5 CNN Model

A CNN model using speech analysis may detect women's stress. These models are trained on massive stress-labeled women's voice datasets. The CNN model can binary classify speech stress using audio data. Test data may be used to evaluate the CNN model after training it on the dataset. A CNN model may identify stress in women by examining speech stream features. Deep learning may improve stress detection. Applications may alter CNN layer count and size. Several models add convolutional or pooling layers after a few fully linked layers. Layer size and description affect model accuracy. The model's size and training duration may depend on each convolutional layer's filters and fully connected layer's units. CNN model training epochs (the number of times the dataset is processed through the model). Too many epochs overfit. The model's epoch count matters. Lastly, CNN speech analysis algorithms identify female stress effectively. CNN architecture layers detect stress. Fewer epochs may improve model performance.

Eight sequential structures. Start with a one-dimensional convolutional layer with 256 filters and a third-order kernel of 3. This layer outputs (x, y, z) (None, 252, 256). The output tensor activates the second layer. 1D convolutional layer 4 has 256 filters and a kernel size of 3. The third batch normalization layer standardizes the second layer's output. Fourth-layer normalized output. The fifth layer randomly rejects previous layer outputs to minimize overfitting. Max pooling reduces output-tensor spatial dimensions to (layer 6) (None, 31, 256). After activating the output tensor, a 1D convolutional layer with 128 filters and a kernel size of 3 follows. After three 1D convolutional layers with 128 filters and 3 kernel sizes, the eighth layer activates the output tensor. Batch normalization ensures consistency after integrating third, fourth, and fifth convolutional layer output tensors. Standardization gives activation and dropout layers the batch normalization layer's output tensor. Nine max pooling layers reduce output tensor spatial dimensions to (None, 3, 128). A 1D convolutional layer with 64 filters and a kernel size of 3 activates the output tensor in the tenth and eleventh layers. Flattening the second 1D convolutional layer output creates a 1D tensor (None, 192). After the six-unit dense twelfth layer, the activation layer activates the output tensor. Adjust 1,283,078 settings.

The 8 layers are:

- Input Layer: The model is fed a one-dimensional (1D) time series of shape (n timesteps, n features) as its input.
- Conv1D Layer1: This layer uses 256 3×3 filters with the padding value of "valid" and the strides value of "1" to generate an output of form (n timesteps-2, 256).
- Activation layer 1: the ReLU activation function is applied to the output of the preceding layer.
- Conv1D layer 2: This layer uses 256 3×3 filters with padding = 'valid' and strides = 1, and its output has a shape that is (n timesteps-4, 256).
- Batch Normalization layer: In order to achieve quicker convergence and improved generalization, this layer adds batch normalization to the output of the preceding layer.
- Activation layer 2: To the output of the prior layer, the ReLU activation function is applied here.
- Dropout Layer: Overfitting may be avoided by having this layer remove neurons at random. The layer's output has the structure (n timesteps/8), and the layer executes max pooling with a pool size of 8, and strides of 8.
- Conv1D Layer 3: This includes 128 3×3 filters, and its output has the form (n timesteps/8–2), 128, thanks to the padding = 'valid' and strides = 1 settings.
- Activation layer 3: This layer takes the layer 2 output and activates it using the ReLU function.
- Conv1D layer 4: This includes 128 3×3 filters with padding = 'valid' and strides = 1, for a total of 128 outputs in the form (n timesteps/8–4, 128). The output of the third activation layer is sent into the fourth, which uses the ReLU activation function.
- Conv1D layer 5: To generate an output of form (n timesteps/8–6, 128), layer 5 of the Conv1D model uses 128 filters of size 3×3, with padding = 'valid' and strides = 1.
- Batch normalization layer 2: This layer uses batch normalizing on the output of the first layer.
- Activation layer 5: This applies the ReLU activation function to the output of the previous layer.
- Dropout layer 2: This layer randomly drops out some of the neurons to prevent overfitting.
- MAXPooling1D layer 2: The second layer is a MaxPooling1D implementation, which, given a pool size of 4 and strides of 4, generates an output of shape (n timesteps/32) 128, where is the number of output samples.
- MAxPooling1D layer 2: The second layer is a MaxPooling1D, which uses a 4×4 pool and 4×4 strides to provide an output of shape (n timesteps/32, 128).
- Conv1D layer 6: Using 64 3×3 filters, padding = 'valid,' and strides = 1, layer 6 of the Conv1D network yields an output of size (n timesteps/32–2, 64).
- Conv1D layer 6: Using 64 3×3 filters, padding = 'valid,' and strides = 1, layer 6 of the Conv1D network yields an output of size (n timesteps/32–2, 64).
- Activation layer 6: Layer 6's activation function takes the output of layer 5 and applies the ReLU activation to it.
- Conv1D layer 7: This has an output of form (n timesteps/32–4, 64) and uses 64 filters of size 3×3, with padding = 'valid' and strides = 1.

- Activation layer 7: The seventh activation layer takes the output of the sixth layer and activates it using the ReLU function.
- Flatten layer: The output of the preceding layer is transformed into a 1D vector and flattened by this layer.
- Dense layer: This layer applies a fully connected layer to the preceding layer's flattened output, and it contains 6 output units.
- Activation layer 8: this softmax activation function is used on the layer 7 output.

5 Results and Discussions

The CNN model was trained and evaluated using male and female speaker audio samples and achieved % 85accuracy. The algorithm accurately determines if a speaker's voice was strained using an audio clip's trajectory. The CNN model may be used in contact center monitoring, mental health diagnostics, and stress treatment to detect speech stress. Nevertheless, further research is needed to test the model on larger and more diverse datasets and optimize its hyperparameters for accuracy and efficiency. Stats show audio clip stress. CNN predicts speaker stress, but values show it. The program correctly assessed most audio samples' stress. Known model. Computers predicted speaker stress. The CNN model predicted speakers' stress levels in audio samples better than others (Fig. 6).

6 Conclusion

Healthy stress may inspire greatness in women. Stress may also harm. Modern life is more stressful. Work, family, demanding employment and the desire for constant change worsen the problem. Machine learning algorithms can produce continuous, non-intrusive stress detection systems to enhance the quality of life. Speech analysis is one of the best stress detection tools. This study recognized stress using just voice cues. CNN architectures in deep learning systems do this. CNN outperformed the other classifiers with 85% accuracy using just voice features. An emotional labeling exam will measure stress (stressed vs. unstressed). This study proposes analyzing voice and sound data to detect stress-related speech patterns. CNN-style analysis and categorization may improve mental health and well-being. Future accuracy-improvement methods may use new models.

	actualvalues	predictedvalues
1	female_none	female_Stressed
2	female_none	female_Stressed
3	female_none	female_Stressed
4	female_Unstressed	female_Stressed
5	female_Unstressed	female_Stressed
6	female_Unstressed	female_Unstressed
7	female_Unstressed	female_Unstressed
8	female_Unstressed	female_Unstressed
9	female_Unstressed	female_Unstressed
10	female_Unstressed	female_Unstressed
11	female_Unstressed	female_Unstressed
12	female_Stressed	female_Stressed
13	female_Stressed	female_Stressed
14	female_Stressed	female_Stressed
15	female_Stressed	female_Stressed
16	female_Stressed	female_Stressed
17	female_Stressed	female_Stressed
18	female_Stressed	female_Stressed
19	female_Stressed	female_Stressed
20	female_Stressed	female_Stressed

Fig. 6. Results of the CNN model

References

1. Hindra Kurniawan1, Alexandr V. Maslov1,2, Mykola Pechenizkiy1 1Department of Computer Science, TU Eindhoven, the Netherlands hindra.kurniawan@gmail.com, m.pechenizkiy@tue.nl 2Department of MIT, University of Jyvaskyl ¨ a, Finland ¨
2. Kevin Tomba1 , Joel Dumoulin1 , Elena Mugellini1 , Omar Abou Khaled1 and Salah Hawila2 1HumanTech Institute, HES-SO Fribourg, Fribourg, Switzerland 2AIR @ En-Japan, Tokyo, Japan
3. Russell Li & Zhandong Liu Stress detection using deep neural networks. BMC Med Inform Decis Mak 20 (Suppl 11), 285 (2020). https://doi.org/10.1186/s12911-020-01299-4
4. Dr. S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, S. Dhas Associate Professor, Student, Student, Student, Student Department of Computer Engineering, Datta Meghe College of Engineering, Navi-Mumbai, India
5. Anakha P.S1 ,Aiswarya Devi2 ,Anjana S Nair3 3Student, Aishwarya Suresh ,Neema George, India
6. Recognition Of Human Mental Stress Using Machine Learning Paradigms - Mrs. Megha V Gupta, Research Scholar, Computer Engineering, Datta Meghe College of Engineering,
7. A. Baum. Stress, intrusive imagery, and chronic distress. Health psychology, 9(6): 653, 1990.
8. A. Kene and S. Thakare, "Mental Stress Level Prediction and Classification based on Machine Learning," 2021 Smart Technologies, Communication and Robotics (STCR), Sathyamangalam, India, 2021, pp. 1-7, https://doi.org/10.1109/STCR51658.2021.9588803.
9. J. Lee and I. Tashev. High-level feature representation using recurrent neural networks for speech emotion recognition. 2015.
10. Front. Bioeng. Biotechnol., 02 September 2020 Sec. Biomaterials
11. Yamaguchi M. Aragaki T. Eto K. Uchihashi K. Takai, N. and Y. Nishikawa. Effect of psychological stress on the salivary cortisol and amylase levels in healthy young adults. Archives of oral biology, 49(12):963–968, 2004.
12. L. V. D. Maaten and G. Hinton. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605, 2008
13. Muray I.R, Baber C, South A. Towards a Definition and Working Model of Stress and its Effects on Speech Communication, Speech Communication Journal, Volume 20, Issue 1-2, Nov 1996, pp 3-12.
14. Moriyama, T., Mori, S., and Ozawa, S. (2009). A synthesis method of emotional speech using subspace constraints in prosody. Journal of Information Processing Society of Japan, 50(3):1181–1191.
15. Seehapoch, T. and Wongthanavasu, S. (2013). Speech emotion recognition using support vector machines. In Knowledge and Smart Technology (KST), 2013 5th International Conference on, pages 86–91. IEEE.