PRAGUE UNIVERSITY
OF ECONOMICS
AND BUSINESS

Acta Informatica Pragensia
2023, Volume 12, Issue 1, pp. 104–122
Special Issue: Deep Learning Blockchain-enabled Technology
for Improved Healthcare Industrial Systems
https://doi.org/10.18267/j.aip.208

Article Open Access

Longitudinal Investigation of Work Stressors Using Human Voice Features

Indhumathi Natarajan [1], Maheswaran Shanmugam [1], Samiappan Dhanalakshmi [2], Santhosh Easwaramoorthy [1], Sethuraja Kuppusamy [1], Saravanan Balu [1]

[1] Department of Electronics and Communication Engineering, Kongu Engineering College, Erode-638060, India
[2] SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, India

Corresponding author: Maheswaran Shanmugam (mmaheswaraneie@gmail.com)

Abstract

Stress is a part of everyone’s life. Any event or thought that makes you upset, furious or anxious can set it off. It will affect the human health mentally and physically and produce a negative impact on nervous and immune systems in our body. The human voice carries a lot of information about the person speaking. It also aids in determining a person's current state. In this proposed method, stress was detected using a deep learning model. Automatic stress detection is becoming an intriguing study topic as the necessity for communication between humans and intelligent systems rises. The hormone called cortisol can also be used to determine the body’s stress state. For most people, however, it is not a viable option. Speech features are particularly affected by stress, which is combined with the aim that voice data would serve as an easy-to-capture measure of everyday human stress levels and hence as an early warning signal of stress-related health problems. The proposed technique extracts Mel filter bank spectral coefficients from pre-processed voice input and the spectrum coefficients are extracted. The features of Mel frequency cepstral coefficients are applied to feed-forward networks and long short-term memory to predict the status of stress output using a binary decision, i.e., unstressed or stressed. The Mel spectrum and spectrogram output shows the variation in stressed and unstressed voice features. The results of the proposed method indicate better performance compared to an existing model. The model was developed as a web application to be used by workers to test their state of stress at any time.

Keywords

Stress; MFCC; Mel filter bank; FFT; Mel scale; Spectrogram; LSTM.

Citation: Natarajan, I., Shanmugam, M., Dhanalakshmi, S., Easwaramoorthy, S., Kuppusamy, S., & Balu, S. (2023). Longitudinal Investigation of Work Stressors Using Human Voice Features. *Acta Informatica Pragensia*, 12(1), 104–122. https://doi.org/10.18267/j.aip.208

Special Issue Editors: Mazin Abed Mohammed, University of Anbar, Iraq
Seifedine Kadry, Noroff University College, Norway
Oana Geman, Ștefan cel Mare University of Suceava, Romania

Academic Editor: Zdenek Smutny, Prague University of Economics and Business, Czech Republic

1 Introduction

Our bodies' response to a challenge or demand is stress. Stress may be beneficial in small doses, e.g., when it helps escape in a danger situation or make a deadline. However, prolonged stress can be harmful to the health (Akcay et al., 2020). According to a World Health Organization (WHO, 2020) report, the physical and social consequences of stress are growing at an alarming rate and having an influence not just on young people's and children's lives but even on adults' as well. There are two types of stress: chronic and acute.

Acute stress is a temporary stress that passes rapidly. When slamming on the brakes, fighting with a partner or skiing down a steep slope, everyone can feel it. It aids in the management of perilous circumstances. It might also happen when a person is trying something new or intriguing. Everyone experiences intense stress at some point in their lives (Archana & Devaraju, 2020).

Chronic stress is a type of tension that lasts a long time. If a person has money troubles, an unsatisfactory marriage or employment problems, he or she may experience chronic stress. Chronic stress is defined as any sort of stress that lasts for days to weeks. Stress can cause health issues if one becomes unable to control it. Stress can affect the human body in different ways. It causes the release of hormones that stimulate the brain, tighten the muscles and raise the heart rate. These effects are beneficial in the short term since they can assist in dealing with stressful circumstances. This is how the human body defends itself (Bandela & Kumar, 2017). An extended period of stress can have a serious negative impact on one's health.

Stress is an inevitable condition. Among the many things that might stress out a person, social evaluation scenarios such as interviews and public speaking activities have a particularly negative impact on a person's confidence and well-being. Therefore, it is critical to create a solid plan to reduce and manage human stress in order to aid people in regaining their mental and physical equilibrium. Even when there is no threat, the body remains attentive while under chronic stress. This puts a person at risk for a variety of health issues, including

- obesity,
- high blood pressure,
- diabetes,
- menstrual problems,
- heart disease,
- skin problems, such as acne or eczema, and
- depression or anxiety.

A stressor is intrinsically neutral: it might result in either harmful stress (distress) or healthful, stimulating stress (eustress). Individual variations and reactions are what cause either eustress or distress. Any circumstance, encounter or external stimulation that raises someone's stress levels is referred to as a stressor. These occurrences or encounters, which may be physical or psychological, are viewed as dangers or difficulties for the person. Stressors can increase a person's vulnerability to both medical and psychological issues, according to researchers (Neil et al., 2005). When stressors are "chronic, extremely disruptive, or viewed as uncontrolled," they are more likely to have a negative impact on a person's health. Researchers in psychology typically divide the various stresses into four groups (Neil et al., 2005):

- catastrophes,
- major life events,
- micro stressors, and
- ambient stressors.

1.1 Catastrophes

Catastrophes are unforeseeable and unpredictable; as a result, they are entirely beyond the control of any one person. Even though it does not happen often, this kind of stressor usually adds a lot of tension to a person's life. According to Stanford University research, those who were affected by natural catastrophes significantly increased their stress levels thereafter (Saeed & Gargano, 2022). One of the most common acute and chronic problems is combat stress.

1.2 Major life events

Major life events include things such as getting married, starting college, losing a loved one, having a child, getting divorced, changing residences, etc. These occurrences, whether favourable or unfavourable, can produce apprehension and anxiety, which in turn breeds stress. For instance, studies have shown that stress levels rise as students move from high school to college, with first-year college students having a roughly two-fold higher risk of stress than those in their last year. The amount of time that has passed since the incident and whether it was positive or bad both affect whether it generates tension or not.

1.3 Micro stressors

Daily irritations and little inconveniences fall under this group. Making judgments, managing deadlines at work or school, navigating traffic, dealing with irritable people, etc., are a few examples. Conflicts with other individuals are a common component of this sort of stressor. However, daily stresses vary from person to person as not everyone views a particular situation as stressful.

1.4 Ambient stressors

They are worldwide low-grade stressors that are a normal component of the environment, as their name suggests. They are characterized as stresses that are persistent, negatively regarded, unimportant, physically observable and resistant to personal efforts to alter them.

In order to assess stress based on its biological and biochemical consequences, a number of strategies have been developed. For instance, the amount of stress has been measured using chemicals such as cortisol and adrenalin (Stephanie, 2022). Additionally, physiological impacts including heart rate, blood pressure and skin temperature are also utilised (Stephanie, 2022).

The majority of research on sentence-level stress detection has employed written content obtained from social media sites such as microblogs (Bartusiak & Delp, 2001). In identifying stress types, the authors employed a framework that included linguistic, visual and social qualities. Several researchers have used stressors and a stress subject lexicon to search tweets for stressful situations. Thousands of written Weibo tweets were gathered and manually sorted into ten types. On Russian speech transcriptions, a basic bidirectional recurrent neural network (RNN) may obtain decent results on word-level stress detection (Bou et al., 2000).

The main objective is to identify the stress level and stress is monitored periodically to reduce health issues of workers in any kind of firm. The stress levels are identified with the help of voice features and no sensors are required.

2 Literature Survey

A method consisting of six attributes, namely foot and hand galvanic skin response, electrocardiogram, heart rate, respiration and electromyogram, has been considered for stress detection (Archana & Devaraju, 2020). The researchers evaluated these features for threshold values using machine learning methods such as decision trees, naïve-Bayes, and k-nearest neighbour. Features were extracted using algorithms. They were decision tree with root nodes and sub nodes, and the dataset was further divided into sub-nodes as attributes; the attributes were divided into trained data and compared with a threshold value that predicts whether the person is under stress or not. The method used a variety of factors to distinguish between depressed and regular speech; nevertheless, the mel frequency cepstral coefficients

(MFCC)-based factor is more appropriate than the others since depressive speech or audio signals might include more information in increased energy bands than normal speech. The voice signal was divided into frames that are generally 10 to 30 milliseconds long. The quickest approach to compute the discrete Fourier transform (DFT) is to employ fast Fourier transform (FFT), a technique that can speed up DFT computations by a factor of a hundred. It produces the best results in both depressed and healthy persons. The authors proposed a method for extracting speech features from a male or female speaker's recorded voice and making comparisons to database templates. Linear predictive codes (LPC), MFCC, perceptual linear prediction (PLP), PLP-RASTA (PLP relative spectra) and other parameters can be used to parameterize speech. When extracting features, PLP and MFCC consider the nature of speech, whereas LPC predicts future features based on previous features (Bandela & Kumar, 2017).

For classification and recognition, techniques such as support vector machine (SVM), dynamic time warping (DTW), hidden Markov model (HMM) and vector quantization (VQ) can be utilised (Prabhu et al., 2021). A rapid and accurate automatic voice recognition method was developed that uses digital processing of speech signals and voice recognition algorithms. To represent the voice signal, digital signal processing techniques such as feature extraction and feature matching have been used. Several approaches have been tested, including liner predictive coding (LPC), HMM, artificial neural network (ANN) and others, in order to find a simple and effective way for speech signal. Following the signal pre-processing or filtering, the extraction and matching procedure takes place. The technology called pseudoscientific technique for inferring deceit from voice stress measurements is designed to distinguish between stressful and non-stressful responses using a deep learning-based psychological stress detection algorithm. Eight convolutional neural network (CNN) layers and completely linked layers make up this module. At each time sequence, the neural network layers gather the temporal information of the extracted features and calculate the frame-level output (Bartusiak & Delp, 2021; Bou et al., 2000).

Voice features have been extracted from MFCC and analysed using LPC as known from the literature, but frequency analysis of stress has not been done using any method. Longitudinal analysis provides improved health results. Thus, the proposed method concentrates on longitudinal analysis with the help of stress analysis applied periodically.

3 Existing Method

3.1 Architecture

In an existing system, Figure 1 shows that the statistical dataset has six attributes, namely foot and hand galvanic skin response, electrocardiogram, heart rate, respiration and electromyogram (Dhole & Kale, 2020). Based on these algorithms, data are pre-processed and then divided into a training dataset and a testing dataset.

The dataset is compared with the threshold values and that value is compared and predicts whether the person is in a stressed or unstressed condition (Fernandes & Ullah, 2021; Firoz et al., 2009). Some of the pre-processing steps are importing the dataset, importing the libraries and cleaning. The data are divided into 80% for training the model and 20% for testing the model (Burrowes et al., 2022; Dymecka et al., 2022; Robinson et al., 2023; AlShorman et al., 2022).

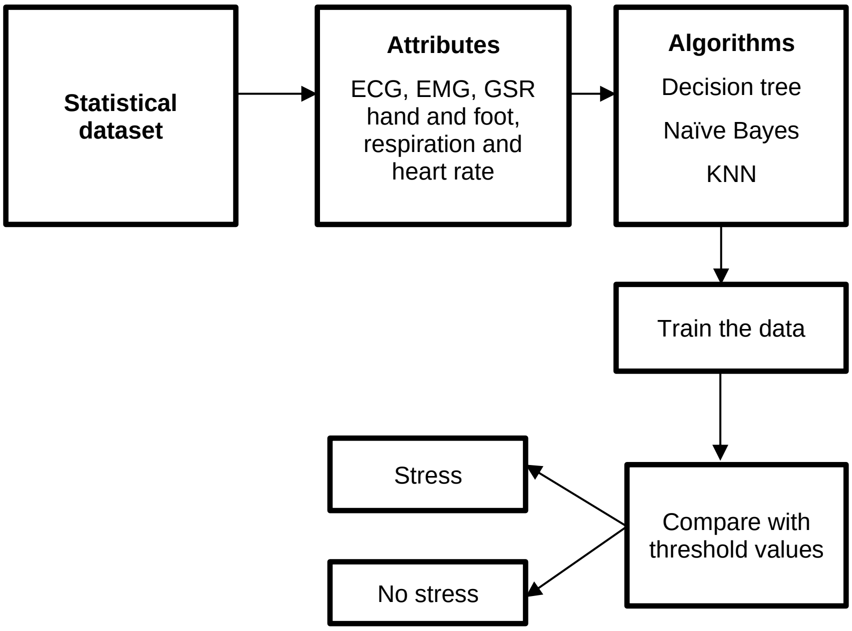


Figure 1. *System overview of stress identification process.*

3.2 Stress detection using machine learning

By considering the six attributes represented in Figure 1. Data are collected for the use of statistical datasets (Gupta & Shubhangi, 2022). Data pre-processing, involving data cleaning and also checking for any missing data. Feature extraction, where the original data are encoded into an unreadable format to avoid hacking.

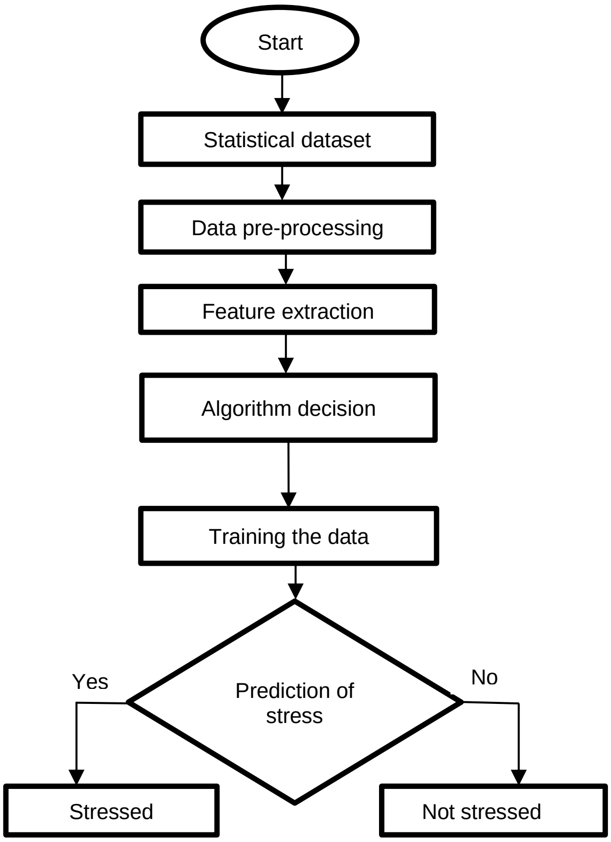


Figure 2. *Existing method flowchart.*

Features are extracted using algorithms such as decision tree with root nodes and sub nodes. The root node is the dataset, further divided into sub-nodes as attributes; the attributes are further divided into trained data and compared with a threshold value that predicts whether the person is under stress or not. Naïve Bayes is used for probability classification, calculating the problems based on assumptions and probability, and k-nearest neighbour is used for classification and predicts the nearest neighbour in the data. The dataset is split into two parts, namely training and testing. The training dataset is always larger than the testing dataset.

Based on the trained value, the result is tested as 0 for not stressed and 1 for stressed. The trained value is compared to the threshold value; based on these values, it is possible to predict and detect whether the person is under stress or not. The steps to predict the stress level are shown in Figure 2 (Hansen, 1996).

3.3 Inference of existing method

Stress detection involves the data set collection from the website which collects the statistical data and also considers the attributes for the dataset (He et al., 2011). Based on the statistical data, the dataset is trained and compared with the threshold values (Hilmy et al., 2021). The predicted values for stress detection are taken with the help of classification machine learning algorithms and values are obtained from the confusion matrix (Kalatzantonakis et al., 2021). A graph of predicted values from heart rate and hand galvanic skin response (GSR) of the physiological dataset for stress detection is shown in Figure 3.

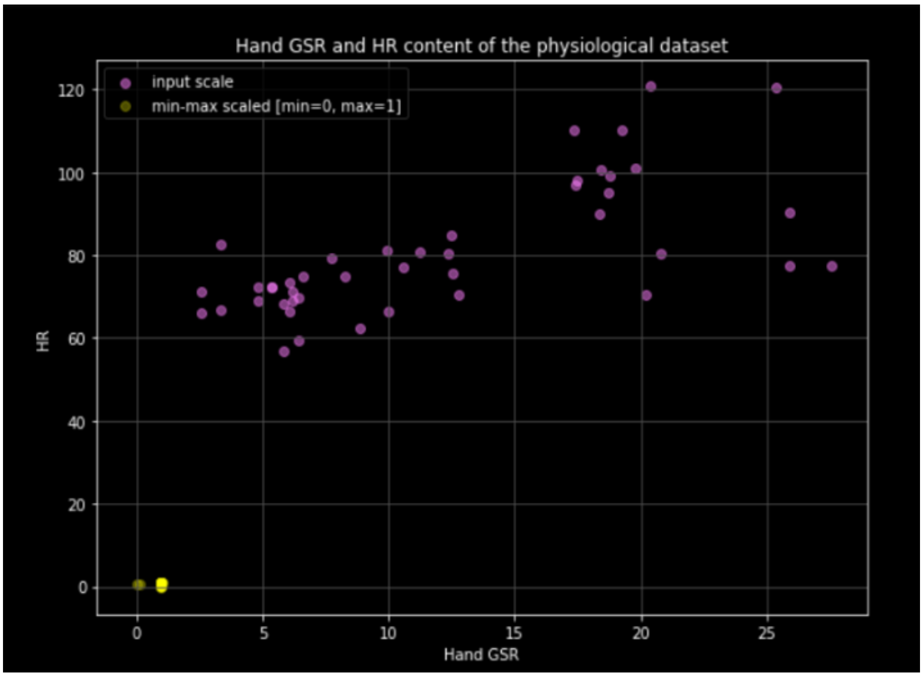
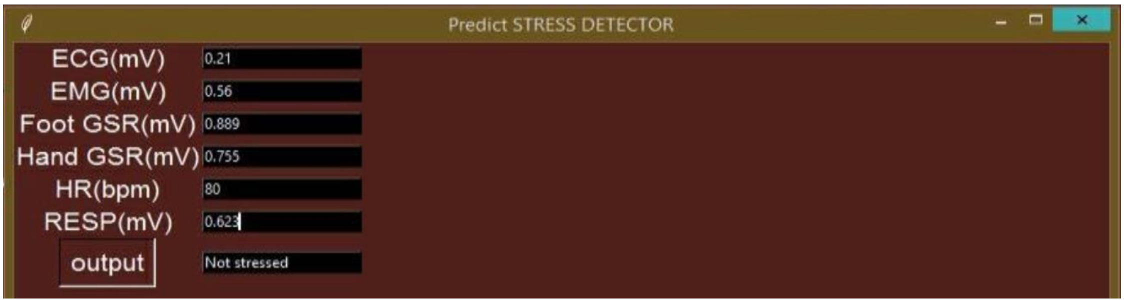
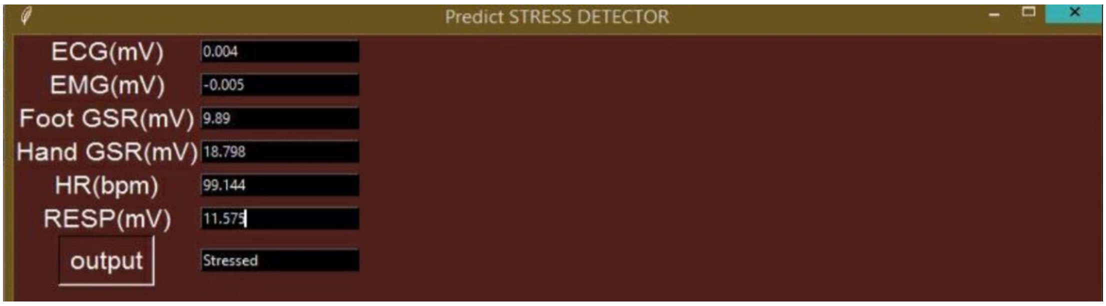


Figure 3. *Graph of predicted values. Source: (Archana & Devaraju, 2020).*



ECG(mV) 0.21
EMG(mV) 0.56
Foot GSR(mV) 0.889
Hand GSR(mV) 0.755
HR(bpm) 80
RESP(mV) 0.623
Not stressed

Figure 4. *Output for a person (not stressed). Source: (Archana & Devaraju, 2020).*



Predict STRESS DETECTOR

ECG(mV) 0.004

EMG(mV) -0.005

Foot GSR(mV) 9.89

Hand GSR(mV) 18.798

HR(bpm) 99.144

RESP(mV) 11.575

output Stressed

Figure 5. *Output for a person (stressed). Source: (Archana & Devaraju, 2020).*

The values are predicted from the physiological dataset on heart rate and hand galvanic skin response; the minimum value is 0 and the maximum value is 1. After entering the input values in the website, the output of the model whether the person is in a stressed or unstressed condition is shown in Figures 4 and 5.

4 Proposed method

4.1 Proposed architecture

The proposed architecture shown in Figure 6. The audio file is input to the feature extraction block, the MFCC is extracted from the input and pre-processing takes place. The pre-processed output is given to the LSTM architecture. In this way, status of stress is analysed. The LSTM architecture variables are as follows: Ct is the cell state vector, Ct-1 is activation of the memory cell C at the time step t-1, ht is the hidden state vector; ht-1 is the hidden state vector at the time step t-1; σ is the sigmoidal function, and Xt is the input.

By using the audio recordings of workers, we detect whether they are in a stressed condition or not. For this, we use an audio feature called the MFCC (Kejriwal et al., 2022). These MFCC features are then given as the input for an LSTM-based RNN model to predict whether they are in a stressed condition or not, as shown in Figure 6.

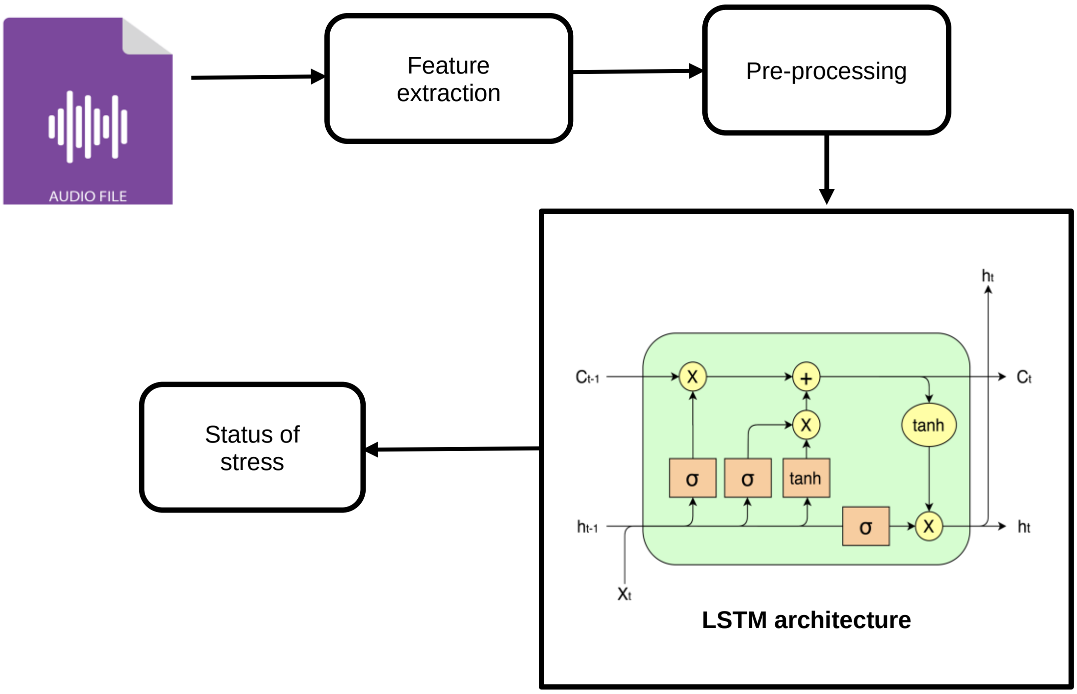


Figure 6. *Stress detection module.*

4.1.1 Data set description

We used the Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D, 2019). It consists of 7442 audio files from 91 people. Within those, there are 43 female actors and 48 male actors. The age range is between 20 and 74. The actors spoke at 4 different levels and in 6 different emotions.

4.2 MFCC

MFCC are one type of voice features which are widely used in automatic speech recognition and speaker recognition (Wang et al., 2015). The most important thing to understand about speech is that the vocal tract, which includes the tongue and teeth, filters the sounds that a human produces. This shape determines the sound that emerges (Kurniawan et al., 2013). We should be able to appropriately represent the phoneme being formed if we can exactly determine the shape. The form of the vocal tract is reflected in the short-time power spectrum envelope, and the role of MFCC is to appropriately represent this envelope (Langari et al., 2020).

4.3 Steps involved in calculating MFCC

Figure 6 denotes the following steps in extracting MFCC features from an audio recording.

1. Speech audio should be taken in WAV format, which is a time domain format (Li et al., 2007). The wave plot is shown in Figure 7.

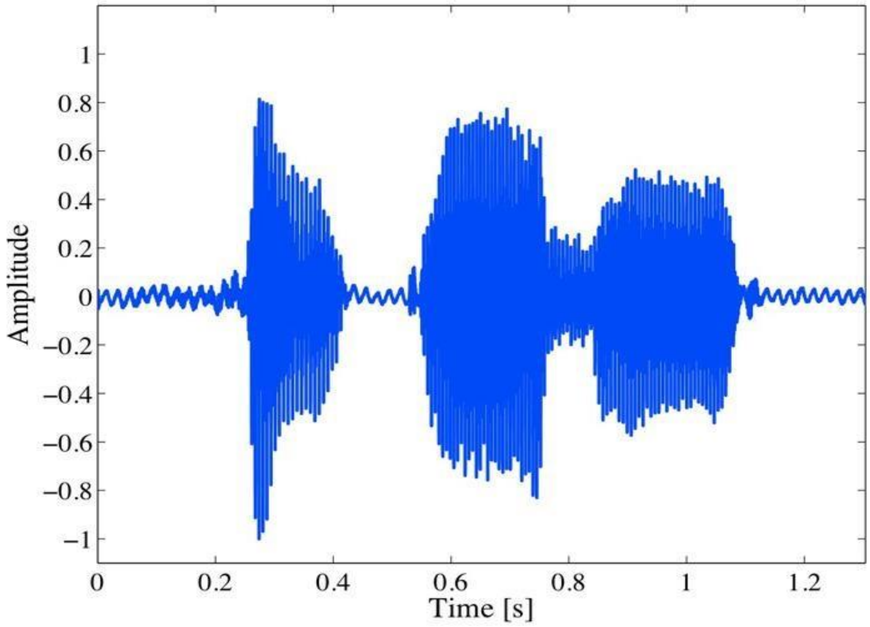


Figure 7. *Wave plot of speech audio file in WAV format.*

2. Here, the FFT is used to convert the audio signals into a periodogram, as shown in Figure 8, and it results in a Nyquist frequency by downsampling the audio signals and hence the sound can be identified (Lieskovska et al., 2021).

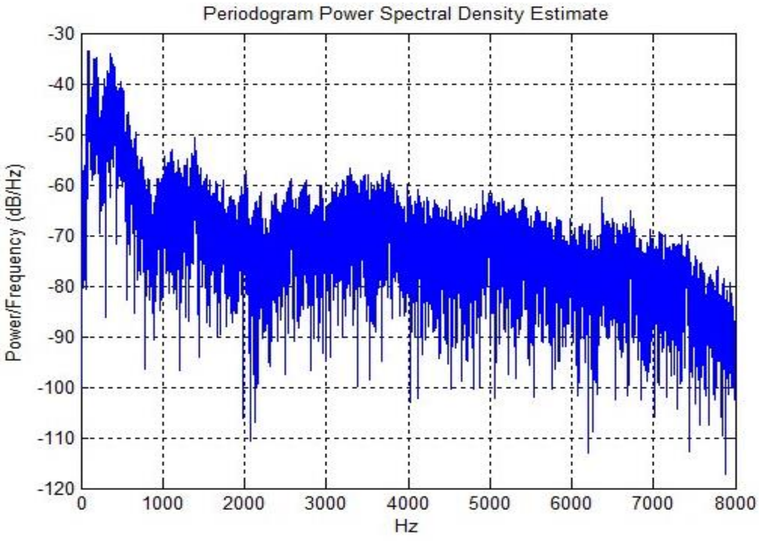


Figure 8. *Periodogram of signal.*

3. Now the periodogram is converted into a spectrogram, which is shown in Figure 9 (Lindasalwa et al., 2010). Different intervals of periodograms are stacked together.

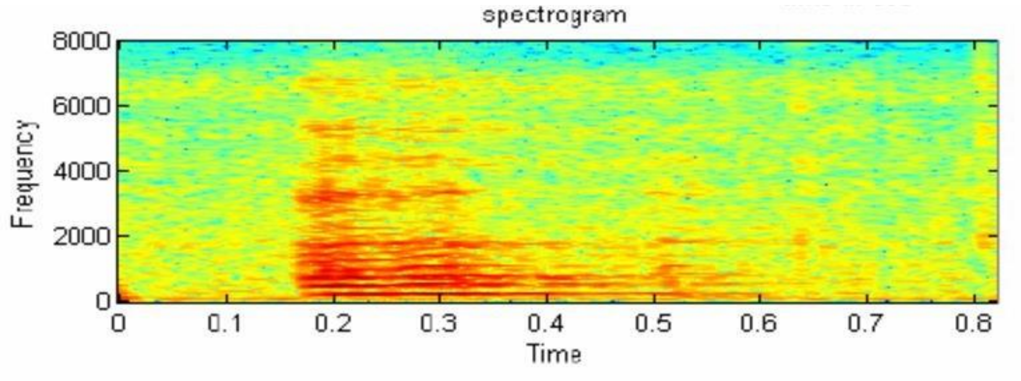


Figure 9. *Spectrogram of speech signal.*

4. A short Fourier transform (SFT) is performed (Lu et al., 2012). The idea behind this is that it helps study the short interval of steady audio signals.
5. Then Hamming windowing is applied (it is an extension of the Hann window that is a raised cosine window of the form) to prevent spectral leakage (a phenomenon which occurs when data are finitely windowed, in general, when we take data and send them through the DFT/FFT algorithm). Figure 10 shows the Hamming window signal (Nassif et al., 2022).

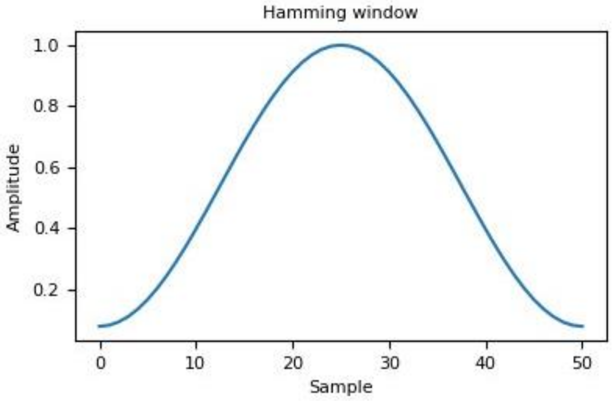


Figure 10. *Hamming window signal.*

6. Again, to convert amplitude to frequency, the FFT is used.
7. Now, the frequency scale is converted into a mel scale by offering a filter that is determined by the machine itself, assisting the machine in learning.
8. All the filter bank energies are logarithmized. It is a collection of band-pass filters and input signals are divided into numerous components, each of which carries a single frequency sub-band of the original signal. Figure 11 shows the mel filter banks of a speech signal.

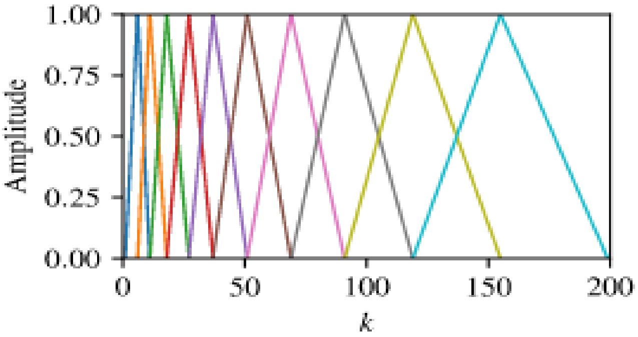


Figure 11. *Mel filter banks.*

9. Log filter bank energies are subjected to Inverse discrete cosine transform (IDCT).
10. The IDCT coefficients 2-13 are retained, while the rest are eliminated (Nassif et al., 2021). 12-13 are considered to be the best.

4.4 Mel scale

The mel scale, which is shown in Figure 12, compares the perceived frequency, or pitch, of a pure tone to its measured frequency (Reddy et al., 2013). Humans are much better at detecting slight pitch fluctuations at low frequencies than they are at high frequencies. Our characteristics become more closely related with what people hear when we use this scale (Rupasinghe et al., 2021).

The frequency to mel scale formula is shown in Equation (1) and Equation (2):

$$\mathrm{M(f)} \ = \ 1125 \ln(1 + \mathrm{f}/700) \tag{1}$$

Where M(f) is the mel scale, f is the frequency of the audio file.

To go from mel scale back to frequency,

$$\mathrm{M} - 1(\mathrm{m}) \ = \ 700\ (\exp(\mathrm{m}/1125) - 1) \tag{2}$$

m-Mels

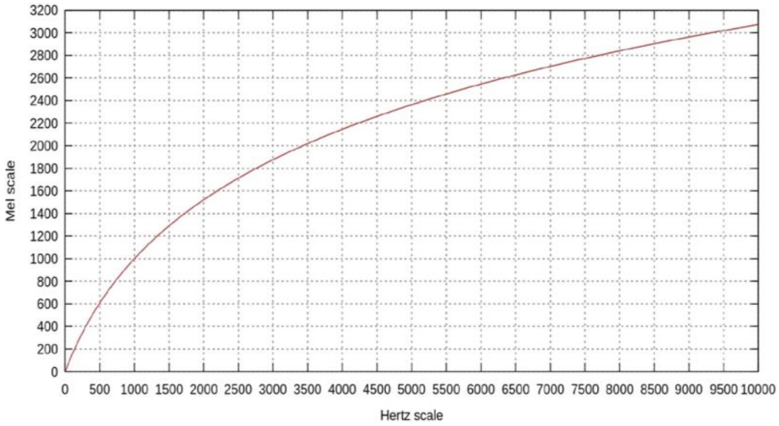


Figure 12. *Mel scale.*

4.5 Calculating MFCC mel filter bank

$$\tilde{S}(l) = \sum_{k=0}^{\frac{N}{2}} S(k)\, M_l(k) \quad l = 0,1,\dots,L-1 \tag{3}$$

Where, $\tilde{S}(l)$ is the mel spectrum, N/2 is half the FFT size, $S(k)$ is the original spectrum, $M_l(k)$ is the l^{th} filter from filter bank, and $k = \left(\frac{kfs}{N}\right) Hz$.

We must first pick a lower and upper frequency to create the filter banks illustrated in Figure 13 (Simantiraki et al., 2016). The lower frequency should be 300 Hz, while the maximum frequency should be 8000 Hz. Naturally, the upper frequency is restricted to 4000 Hz when sample speech is at 8000 Hz (Soury & Laurence, 2013) and the mel filter bank is computed using the formula shown in Equation (3).

Mel-spaced filter bank triangular bandpass filters (Figure 14) are used because of the wide frequency range in the FFT spectrum; furthermore, the vocal signal does not follow a linear scale (Stanek & Sigmund, 2015). To calculate the log energy of each triangle bandpass filter, we multiply the magnitude frequency response by a series of 20 triangular bandpass filters (Tiwari & Darji, 2022). According to the mel frequency, the position of filters is equally spaced, which is linked to the linear frequency *f* by the following Equation (1).

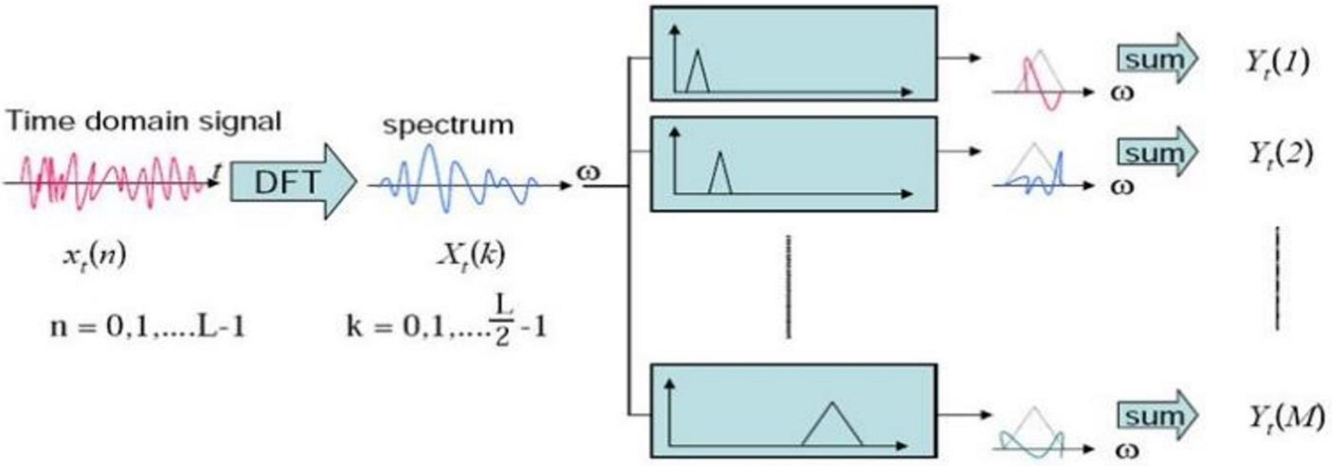


Figure 13. *Mel filter bank.*

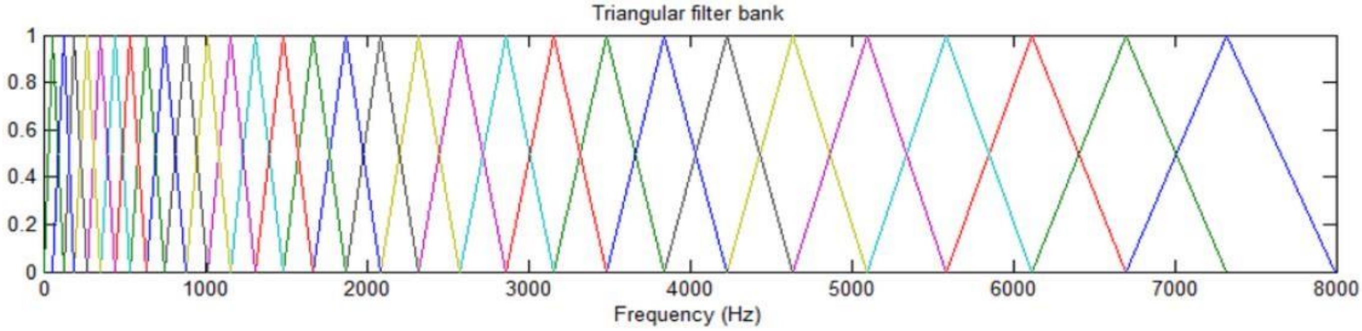


Figure 14. *Triangular filter bank.*

4.6 DCT Step

The following Equation (4) is the DCT to obtain the MFCC.

$$c(i) = \sqrt{\frac{2}{L}} \sum_{m=1}^{L} \log\left(\tilde{S}(m)\right) cos\left(\frac{\pi i}{L}(m - 0.5)\right) \qquad i = 0,1, \dots C-1 \tag{4}$$

Where, $\tilde{S}(m)$ is the mel spectrum, C is the number of cepstral coefficients desired, and m is the mel scale.

Here, DCT is used over IFFT since DCT does not require complex arithmetic operations and DCT uses the redundancy in a real signal to more efficiently accomplish the same function as the FFT (Vaikole et al., 2020). DCT is more computationally efficient.

4.7 Delta cepstrum

Mel frequency cepstral coefficients give smooth and perceptually meaningful estimations of speech spectra through time (Li et al., 2007). Because speech is fundamentally a dynamic signal, it makes sense to look for a representation that incorporates some component of the dynamic character of the short-term cepstrum temporal derivatives (both first and second-order derivatives). The delta cepstrum (first derivative) and delta delta cepstrum (second derivative) parameter sets are the outcome (second derivative). A first difference of cepstral vectors, as indicated in Equation (5), is the simplest technique for obtaining delta cepstrum parameters (Yousefi & John, 2020)

$$\Delta \mathrm{mfccm}[\mathrm{n}] = \mathrm{mfccm}[\mathrm{n}] - \mathrm{mfccm} - 1[\mathrm{n}] \tag{5}$$

Where Δmfccm[n] is the delta cepstrum of the speech signal sample and mfccm[n]- mfccm-1[n] is the local slope of the sample.

The simple difference is an approximation to the first derivative that is not often utilised (Zhou et al., 2001). Instead, a least-squares approximation to the local slope (across an area around the current sample) is utilised, as indicated in Equation (5).

4.8 LSTM model

The first 13 MFCC features are used as input to the LSTM model, since the lower-order coefficients contain most of the information about the overall spectral shape of the source-filter transfer function. Even though the higher-order coefficients represent increasing levels of spectral details, selecting a large number of cepstral coefficients results in more complexity in the models.

The LSTM layer has three gates: a forget gate, an input gate and an output gate. The model is trained with the labelled dataset. The major part of the dataset is used for training. After training the model, testing is done.

5 Results and Discussion

5.1 Probability distribution function

A probability distribution is a statistical function that specifies all possible values and probabilities for a random variable in a given range. The Probability Distribution Function (PDF) for stressed and unstressed audio files are shown in Figures 15 and 16. We can see the variation in the stressed probability density function and in the unstressed probability density function from the above-mentioned plots. So, it is possible to split the stressed and unstressed speech signals using these MFCC features.

The probability distribution function shows variation in the mean and variance of stressed and unstressed speech signals as shown in Table 1. Hence, it provides accurate results between unstressed and stressed people.

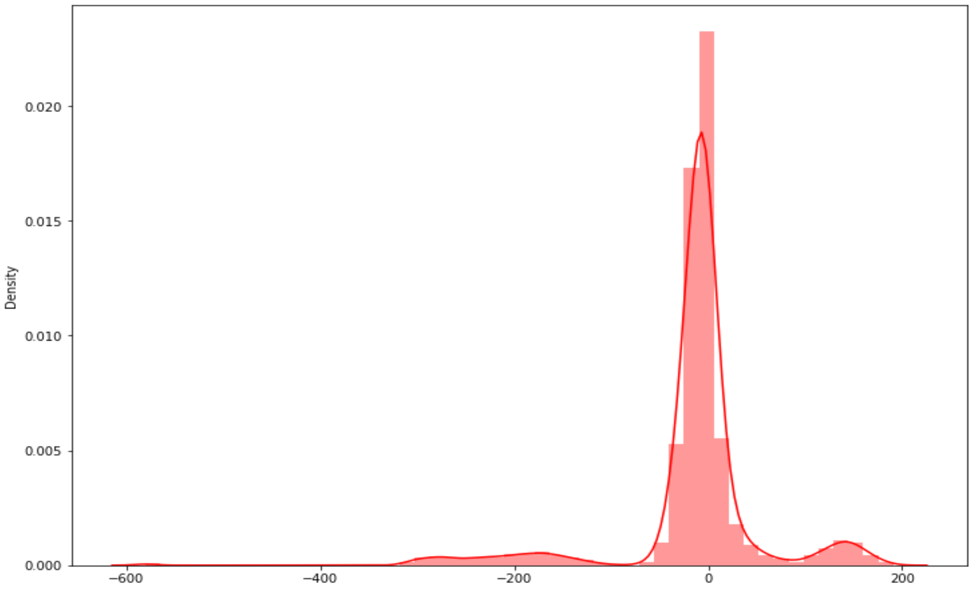


Figure 15. *PDF of unstressed audio.*

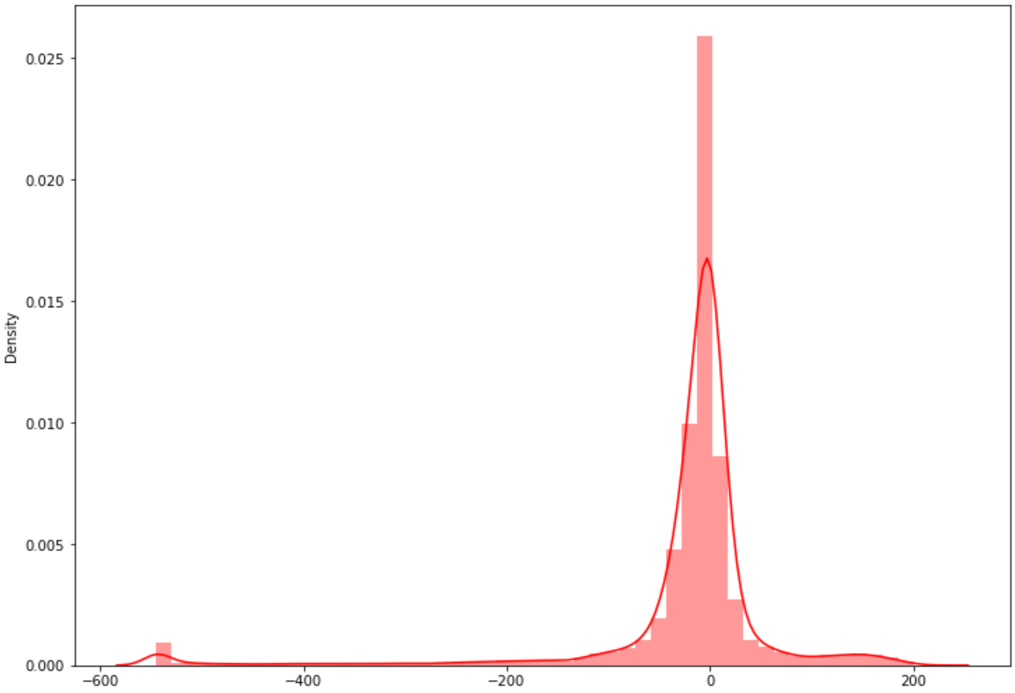


Figure 16. *PDF of stressed audio.*

From Figures 15 and 16, showing the different probability functions, it can be inferred that there is a variation in stressed and unstressed audio files. The file is loaded as a spectrum and those are computed and the mathematical variation of stressed and unstressed audio is identified. The classification is done based on this probability density variation function.

Table 1. *Mathematical deviation of an audio file.*

Parameters	**Values**	
	Stressed	**Unstressed**
Mean	-21.51	-13.77
Variance	9237.44	5478.03
Standard deviation	96.11	74.01

5.2 Spectrogram

A spectrogram is a visual representation of a signal intensity over time at various frequencies contained in a waveform. Spectrograms for stressed and unstressed audio files are shown in Figures 17 and 18.

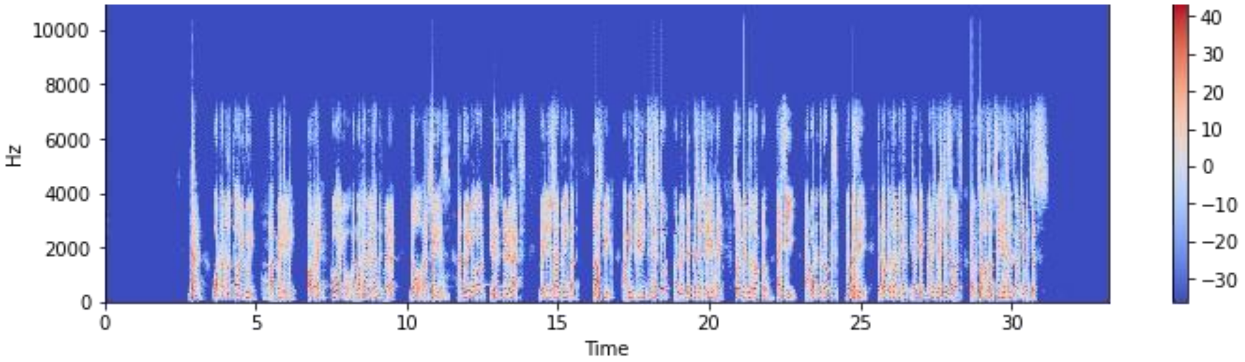


Figure 17. *Spectrogram for stressed audio file.*

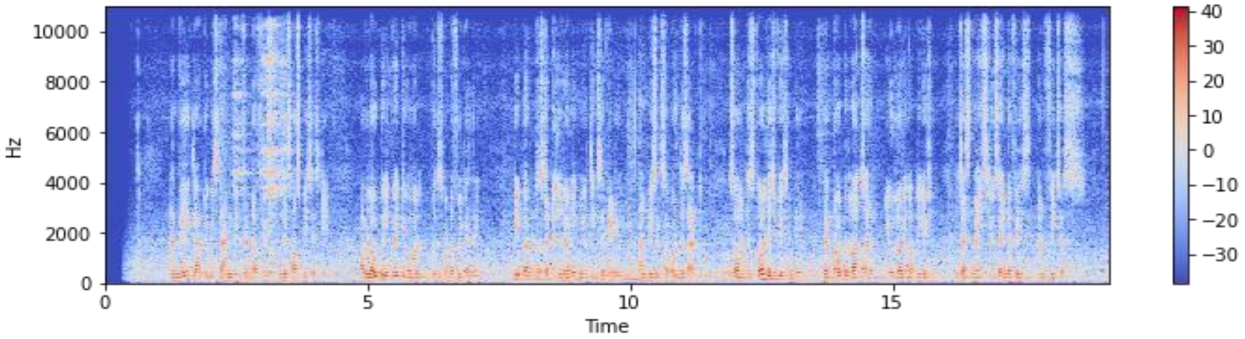


Figure 18. *Spectrogram for unstressed audio file.*

The spectrograms for the unstressed audio file and the stressed audio file show spectrogram equalisation of stressed and unstressed audio. The various decibel levels of different frequencies are shown in the results based on the decibel variation of voice features. The unstressed audio file contains more negative scale and the stressed audio file consists of comparatively negative to positive spectrum.

5.3 Mel spectrum

A mel spectrogram is a spectrogram in which the frequencies are converted to the mel scale. Figures 19 and 20 show the mel spectrum for stressed and unstressed audio files. After converting into Mel scale, the Mel spectrum variations of stressed audio has a more positive Mel coefficient, while the unstressed audio consists of a combination of negative and minimum positive coefficients.

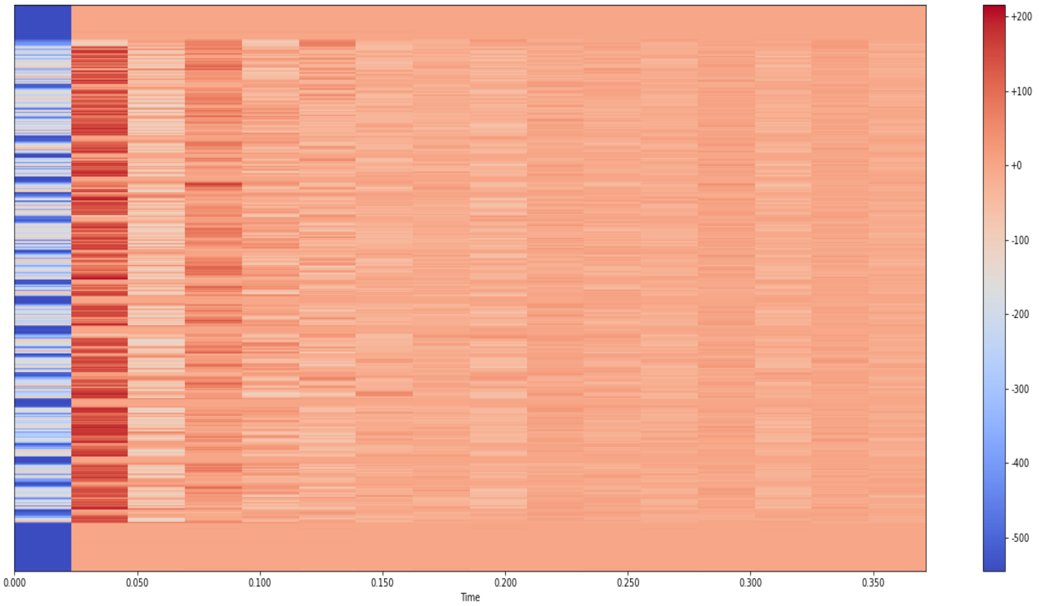


Figure 19. *Mel spectrum for stressed audio file.*

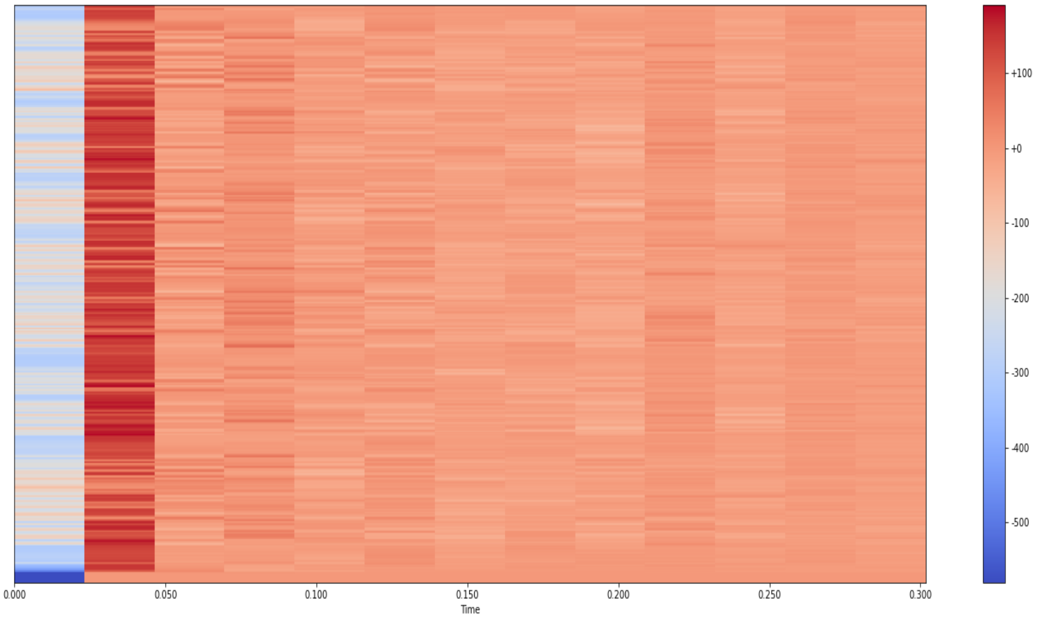


Figure 20. *Mel spectrum for unstressed audio file.*

Figure 21 shows the training of the proposed method and the existing one with respect to epochs, and it can be inferred from the figure that more epochs produce a higher accuracy of the model. Figure 22 shows the results of a comparison between the existing and proposed models. Table 2 presents the result of a comparison between the existing and proposed methods.

Table 2. *Comparison results, existing and proposed method.*

Content	Epochs	Specificity	Sensitivity	Testing accuracy
Existing method	150	95.33	93.23	94.32
Proposed method	150	94.56	96.67	97.87

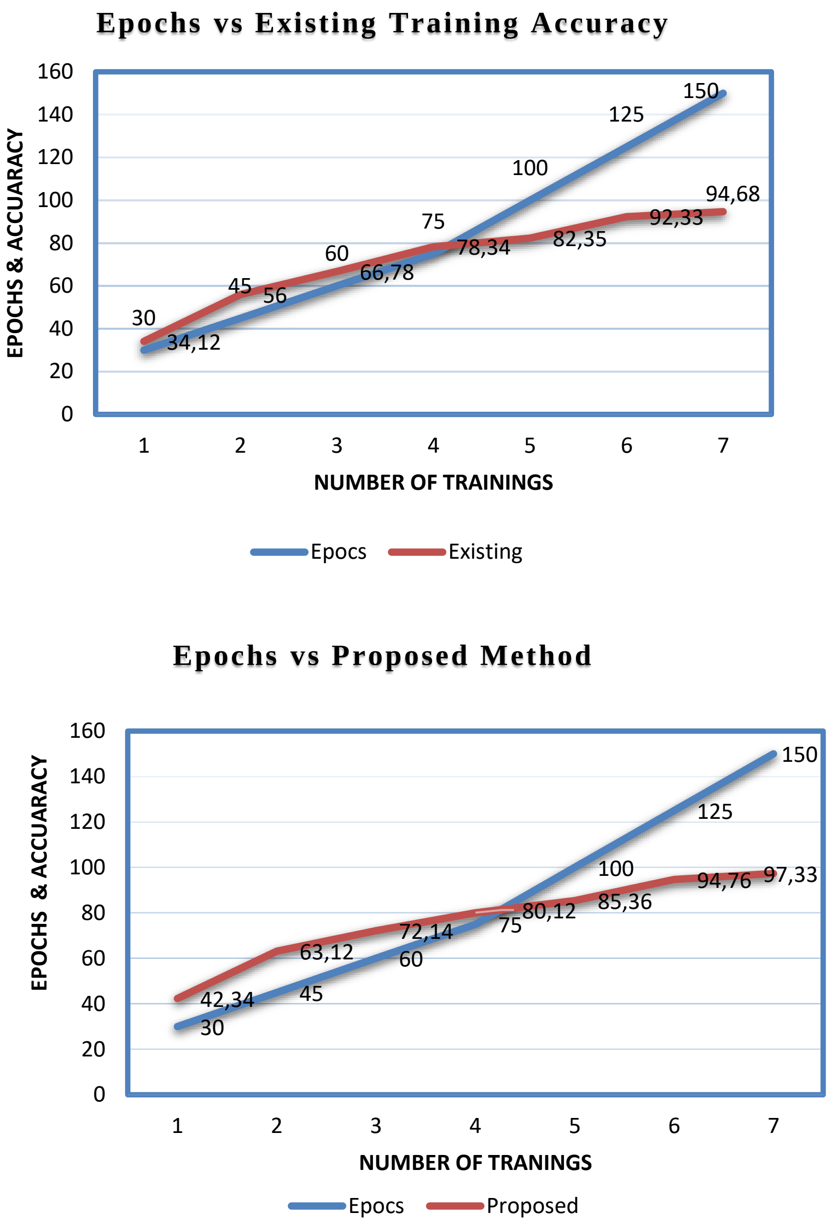


__Figure 21.__ Epochs versus training accuracy.

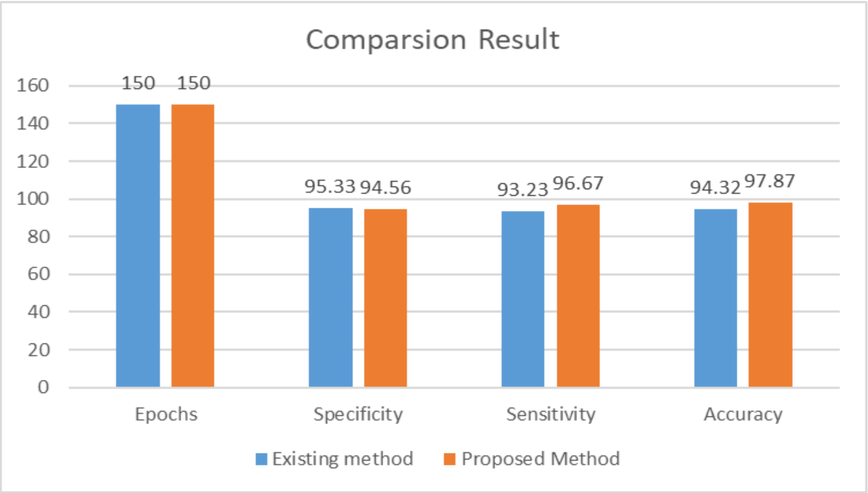


__Figure 22.__ Comparison results, existing and proposed method.

For the same number of epochs in the existing and proposed methods, the specificity, sensitivity and accuracy results are compared. In the comparison, the proposed architecture produces better results than the existing one. The voice features can be used to analyse the stress level better than the sensor method.

6 Conclusion and Future Scope

The probability distribution function and spectrogram of stressed and unstressed speech signal shows that it is possible to differentiate between stressed and unstressed speech signals. As mentioned, 13 MFCC features of stressed and unstressed speech signals were used for the classification. A machine learning model was developed for identifying stress, which has two LSTM layers with two dense and one dropout layers. Besides, stress detection using voice features can be done without the use of sensors. As such, it is a cost-effective method.

The future scope of work is to employ some transformation to enhance the difference between stressed and unstressed MFCC features and develop a deep learning model to detect stress condition of employees with less time consumption and higher accuracy, while also looking for some remedial suggestions for those who are in a stressed condition.

Additional Information and Declarations

Conflict of Interests: The authors declare no conflict of interest.

Author Contributions: I.N.: Conceptualization, Methodology, Writing – review & editing. M.S.: Visualization, Investigation, Supervision, Validation. S.D.: Visualization, Investigation, Supervision. S.E.: Software, Data curation, Writing – original draft. S.K.: Methodology, Data curation, Writing – original draft. S.B.: Investigation, Methodology, Software.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of the external and publicly available dataset.

Data Availability: The data supporting this study's findings are openly available (CREMA-D, 2019).

References

Akçay, M., & Oguz, K. K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116, 56–76. https://doi.org/10.1016/j.specom.2019.12.001

AlShorman, O., Masadeh, M., Heyat, B. B., Akhtar, F., Almahasneh, H., Ashraf, G. M., & Alexiou, A. (2022). Frontal lobe real-time EEG analysis using machine learning techniques for mental stress detection. *Journal of Integrative Neuroscience*, 21(1), 020. https://doi.org/10.31083/j.jin2101020

Archana, V.R., & Devaraju, B.M. (2020). Stress Detection Using Machine Learning Algorithms. *International Journal of Research in Engineering, Science and Management*, 3(8), 251–256.

Bandela, S. R., & Kumar, T. K. (2017). Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. In *International Conference on Computing, Communication and Networking Technologies*. IEEE. https://doi.org/10.1109/icccnt.2017.8204149

Bartusiak, E. R., & Delp, E. J. (2021). Frequency Domain-Based Detection of Generated Audio. *IS&T International Symposium on Electronic Imaging Science and Technology*, 33(4), 273–277. https://doi.org/10.2352/issn.2470-1173.2021.4.mwsf-273

Bou-Ghazale, S. E., & Hansen, J. H. L. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 8(4), 429–442. https://doi.org/10.1109/89.848224

Burrowes, S. A., Goloubeva, O., Stafford, K. A., McArdle, P., Goyal, M., Peterlin, B. M., Haythornthwaite, J. A., & Seminowicz, D. A. (2022). Enhanced mindfulness-based stress reduction in episodic migraine—effects on sleep quality, anxiety, stress, and depression: a secondary analysis of a randomized clinical trial. *Pain*, 163(3), 436–444. https://doi.org/10.1097/j.pain.0000000000002372

CREMA-D. (2019). Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D). https://www.kaggle.com/datasets/ejlok1/cremad

Dhole, N., & Kale, S. (2020). Stress Detection in Speech Signal Using Machine Learning and AI. In *Machine Learning and Information Processing. Advances in Intelligent Systems and Computing* (pp. 11–26). Springer. https://doi.org/10.1007/978-981-15-1884-3_2

Dymecka, J., Gerymski, R., & Machnik-Czerwik, A. (2022). How does stress affect life satisfaction during the COVID-19 pandemic? Moderated mediation analysis of sense of coherence and fear of coronavirus. Psychology Health & Medicine, 27(1), 280–288. https://doi.org/10.1080/13548506.2021.1906436

Fernandes, S. V., & Ullah, M. W. (2021). Development of Spectral Speech Features for Deception Detection Using Neural Networks. In *2021 IEEE 12th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE. https://doi.org/10.1109/iemcon53756.2021.9623077

Firoz S. A., Raji, S. A., & Babu A. P. (2009). Automatic Stress Detection from Speech by Using Discrete Wavelet Transforms. In *Proceedings of National Conference on Information Technology & Business Intelligence*, (pp. 1-5). India. https://www.researchgate.net/publication/200706300_Automatic_Stress_Detection_from_Speech_by_Using_Discrete_Wavelet_Transforms

Gupta M., & Vaikole, S. (2022). Audio Signal Based Stress Recognition System using AI and Machine Learning. *Journal of Algebraic Statistics*, 13(2), 1731–1740.

Hansen, J. H. L. (1996). Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication*, 20(1–2), 151–173. https://doi.org/10.1016/s0167-6393(96)00050-7

He, L., Lech, M., Maddage, N. C., & Allen, N. B. (2011). Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. *Biomedical Signal Processing and Control*, 6(2), 139–146. https://doi.org/10.1016/j.bspc.2010.11.001

Hilmy, M. F., Asnawi, A. L., Jusoh, A., Abdullah, K., Ibrahim, S. F., Ramli, H. a. M., & Azmin, N. F. M. (2021). Stress Classification based on Speech Analysis of MFCC Feature via Machine Learning. In *International Conference on Computer and Communication Engineering*, (pp. 339–343). IEEE. https://doi.org/10.1109/iccce50029.2021.9467176

Kalatzantonakis-Jullien, G., Stefanakis, N., & Giannakakis, G. (2021). Investigation and ordinal modelling of vocal features for stress detection in speech. In *Affective Computing and Intelligent Interaction*. IEEE. https://doi.org/10.1109/acii52823.2021.9597430

Kejriwal, J., Benus, S., & Trnka, M. (2022). Stress detection using non-semantic speech representation. In *2022 32nd International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE. https://doi.org/10.1109/radioelektronika54537.2022.9764916

Kurniawan, H., Maslov, A. V., & Pechenizkiy, M. (2013). Stress detection from speech and Galvanic Skin Response signals. In *Proceedings of IEEE International Symposium on Computer-Based Medical Systems,* (pp. 209–214). IEEE. https://doi.org/10.1109/cbms.2013.6627790

Langari, S., Marvi, H., & Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*, 20, 100424. https://doi.org/10.1016/j.imu.2020.100424

Li, C., Liu, J., & Xia, S. (2007). English sentence stress detection system based on HMM framework. *Applied Mathematics and Computation*, 185(2), 759–768. https://doi.org/10.1016/j.amc.2006.06.081

Li, X., Tao, J., Johnson, M., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007). Stress and Emotion Classification using Jitter and Shimmer Features. In *International Conference on Acoustics, Speech, and Signal Processing*. IEEE. https://doi.org/10.1109/icassp.2007.367261

Lieskovska, E., Jakubec, M., Jarina, R., & Chmulik, M. (2021). A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics*, 10(10), 1163. https://doi.org/10.3390/electronics10101163

Lindasalwa Muda, Mumtaj Begam, & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. https://arxiv.org/abs/1003.4083

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T, Campbell, A.T., Gatica-Perez, D., & Choudhury, T. (2012). Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of ACM conference on ubiquitous computing*, (pp. 351–360). ACM. https://doi.org/10.1145/2370216.2370270

Nassif, A. B., Shahin, I., Elnagar, A., Velayudhan, D., Alhudhaif, A., & Polat, K. (2022). Emotional speaker identification using a novel capsule nets model. *Expert Systems With Applications*, 193, 116469. https://doi.org/10.1016/j.eswa.2021.116469

Nassif, A. B., Shahin, I., Hamsa, S., Nemmour, N., & Hirose, K. (2021). CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. *Applied Soft Computing*, 103, 107141. https://doi.org/10.1016/j.asoc.2021.107141

Prabhu, Ram, N., Meeradevi, T., Vibin, Mammen, Vinod., Gothainayaki, A., Anusha, S., & Agalya, T. (2021) Comparative analysis for offensive language identification of tamil text using SVM and logistic classifier. In *Proceedings of CEUR Workshop Proceedings, 3159*, (pp. 976–983). India.

Reddy, V. R., Maity, S., & Rao, K. J. (2013). Identification of Indian languages using multi-level spectral and prosodic features. *International Journal of Speech Technology*, 16(4), 489–511. https://doi.org/10.1007/s10772-013-9198-0

Robinson, L. E., Valido, A., Drescher, A., Woolweaver, A. B., Espelage, D. L., LoMurray, S., Long, A. C. J., Wright, A. A., & Dailey, M. J. (2023). Teachers, Stress, and the COVID-19 Pandemic: A Qualitative Analysis. *School Mental Health,* 15, 78–89. https://doi.org/10.1007/s12310-022-09533-2

Rupasinghe, L., Alahendra, A.M.A.T., Ranathunge, R.A.D., & Perera, P.D. (2021). Robust Speech Analysis Framework Using CNN. In *2021 3rd International Conference on Advancements in Computing (ICAC),* (pp. 485-490). IEEE. https://doi.org/10.1109/icac54203.2021.9671080

Saeed, S. A., & Gargano, S. P. (2022). Natural disasters and mental health. *International Review of Psychiatry*, 34(1), 16–25. https://doi.org/10.1080/09540261.2022.2037524

Schneiderman, N., Ironson, G., & Siegel, S. D. (2005). Stress and Health: Psychological, Behavioral, and Biological Determinants. *Annual Review of Clinical Psychology*, 1(1), 607–628. https://doi.org/10.1146/annurev.clinpsy.1.102803.144141

Simantiraki, O., Giannakakis, G., Pampouchidou, A., & Tsiknakis, M. (2016). Stress Detection from Speech Using Spectral Slope Measurements. In Pervasive Computing Paradigms for Mental Health. FABULOUS MindCare IIOT 2016 2016 2015, (pp. 41–50). Springer. https://doi.org/10.1007/978-3-319-74935-8_5

Soury, M., & Devillers, L. (2013). Stress Detection from Audio on Multiple Window Analysis Size in a Public Speaking Task. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE. https://doi.org/10.1109/acii.2013.93

Stanek, M., & Sigmund, M. (2015). Finding the Most Uniform Changes in Vowel Polygon Caused by Psychological Stress. *Radioengineering,* 24(2), 604–609. https://doi.org/10.13164/re.2015.0604

Stephanie, W. (2022). Overview of Biofeedback. https://www.webmd.com/pain-management/biofeedback-therapy-uses-benefits

Tiwari, P. K., & Darji, A. D. (2022). Pertinent feature selection techniques for automatic emotion recognition in stressed speech. *International Journal of Speech Technology*, 25(2), 511–526. https://doi.org/10.1007/s10772-022-09978-5

Vaikole, S., Mulajkar, S., More, A., Jayaswal, P., & Dhas, S. (2020). Stress Detection through Speech Analysis using Machine Learning. *International Journal of Creative Research Thoughts*, 8(5), 2239-2246.

Wang, K., An, N., Li, B., Zhang, Y., & Li, L. (2015). Speech Emotion Recognition Using Fourier Parameters. *IEEE Transactions on Affective Computing*, 6(1), 69–75. https://doi.org/10.1109/taffc.2015.2392101

WHO. (2020). *Doing What Matters in Times of Stress: An Illustrated Guide*. WHO. https://www.who.int/publications/i/item/9789240003927

Yousefi, M., & Hansen, J. H. L. (2021). Block-Based High Performance CNN Architectures for Frame-Level Overlapping Speech Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 28–40. https://doi.org/10.1109/taslp.2020.3036237

Zhou, G. J., Hansen, J. H. L., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 9(3), 201–216. https://doi.org/10.1109/89.905995

Editorial record: The article has been peer-reviewed. First submission received on 29 October 2022. Revisions received on 21 December 2022 and 4 February 2023. Accepted for publication on 8 February 2023. The editors coordinating the peer-review of this manuscript were Mazin Abed Mohammed, Seifedine Kadry, and Oana Geman. The editor in charge of approving this manuscript for publication was Zdenek Smutny .

Special Issue: Deep Learning Blockchain-enabled Technology for Improved Healthcare Industrial Systems.

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951