iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XII **Month of publication:** December 2024

DOI: https://doi.org/10.22214/ijraset.2024.65786

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XII Dec 2024- Available at www.ijraset.com

Sentiment Predictor for Stress Detection Using Voice Stress Analysis

Jash Shah[1], Ashutosh Thite[2], Aditya Shinde[4], Shreyas Kulkarni[4]
Dept. of Computer Engineering Pune Institute of Computer Technology Pune, India

Abstract: ***This project proposes a machine learning based approach to predict the sentiment i.e Stressed, Unstressed or Neutral from voice of a human being and also presents a comparative analysis among various machine learning models. This technology seeks to distinguish between stressed and non-stressed outputs in response to stimuli. The dataset used for training in this project comprises audio recordings based on four different situations, each recorded in various emotional states. Given the increased demand for communication between intelligent systems and human, automatic stress detection is becoming an interesting research topic. Even while the level of specific hormones, such cortisol, can be precisely measured to indicate stress, this approach is not feasible for diagnosing stress in interactions between humans and machines. The Sentiment Predictor for Stress Detection Using Voice Stress Analysis utilizes machine learning algorithms such as SVM, Random Forest , Logistic Regression etc. to analyze voice patterns and predict an individual's emotional state based on the changes in their vocal characteristics. The technology can identify changes in pitch, tone, and rhythm, among other factors, to determine an individual's stress level. Overall, the Sentiment Predictor for Stress Detection Using Voice Stress Analysis is a promising technology that can provide valuable insights into an individual's emotional state and support their well-being in various settings.***

Index Terms: Sentiment, Machine learning, Cortisol,MFCCs, Melseptorgam, ZCR, Chroma STFT, RMS, CNN, SVM, Random Forest .

Keywords:

AI	***Artificial Intelligence***
ML	***Machine Learning***
CNN	***Convolutional Neural NetworkANN Artificial Neural Network SVM Support Vector Machine***
MFCCs	***Mel-frequency Cepstral Coefficients***
ZCR	***Zero Crossing Rate***
RMS	***Root Mean Square***
STFT	***Short-Time Fourier Transform***
LSTM	***Long Short Term Memory***

I. INTRODUCTION

Human emotions play a critical role in our daily interac- tions. The ability to recognize the emotional state of an indi- vidual is important for understanding their needs, preferences, and behavior. Emotions can be expressed in many different ways, including through facial expressions, body language, and speech. Of these, speech is a particularly useful modality for identifying emotional states as it is a complex signal that carries information about the speaker, the message, the language, and the emotions.

Stress has become a ubiquitous part of our lives, and it can have a significant impact on an individual's mental and physical health. The early detection of stress is crucial for preventing health-related issues associated with it. Machine learning prediction algorithms can be used to predict stress from a person's voice, making vocal stress analysis technologya promising area of research. The project aims to recognize stress from speech signals by collecting audio data and em- ploying stress-detection models based on machine learningframeworks.

Considerable research in speech-based stress detection has predominantly utilized pre-existing datasets or those collected from a single source or language. In contrast, the dataset in this study was recorded, encompassing diverse situations and Indian accents. This unique dataset contributes to a more comprehensive range of speech signals. Additionally, the study extracted 4-5 features from the audio files, such as Zero Crossing Rate, Chroma STFT, MFCC, and MelSpectogram, known for their effectiveness in detecting emotional statesfrom speech signals.

The stress-detection model employed various classification algorithms, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, and ConvolutionalNeural Networks. These models underwent training and testingon the dataset, and their performance was compared to identify the most suitable one.

A web interface was developed for users to upload an audio file, which the system processed to display the detected emo- tional state, spectrogram, waveplot, providing a comprehensivespeech signal analysis.

The system's applications encompass mental health moni- toring, psychological counseling, and personal wellness, offer-ing early stress detection for improved mental health manage- ment. In workplace scenarios, it can monitor employee stress levels, promoting better work-life balance and productivity.

In summary, the study addressed the gap in speech-based stress detection, utilizing diverse datasets and machine learn- ing algorithms. The proposed system holds potential for ap- plications in mental health monitoring and personal wellness, with scope for further research to expand the dataset andenhance model performance.

II. LITERATURE REVIEW

In this field of research, many classification strategieshave been presented over the years. In the paper 'A Deep Learning-based Stress Detection Algorithm with Speech Sig-nal'[1], a speech signal-based system for deep learning-based psychological stress identification has been developed. The proposed approach first extracts mel-filterbank coefficients from preprocessed speech data, and then uses long short-term memory (LSTM) and feed-forward networks to predictthe stress output status using a binary decision criterion (i.e.,stressed or unstressed). The proposed study just used limitednumber of features and is based on only binary classification.In 'Speech Emotion Recognition using Deep Learning' [4], the advanced version of speech signal-based system [1] is im- plemented. This study aims to identify emotions and classifythem in accordance with voice signals using deep learning and picture classification approaches. The advantage of this system is that various datasets are examined and studied for training an emotion recognition model. The paper [4] is used with Inception Net to identify emotions. With the IEMOCAPdatasets, Inception Net is utilized for emotion recognition. Theapproach also used a limited number of features and audio filesand didn't compare the performance of their proposed modelwith other state-of-the-art speech emotion recognition models.In 'On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks' [2] , this paper emphasizes on improving the robustness and effectiveness of the Speech Recognition model by adding some acoustic noise and testing the model in different room condi- tions. The experimentations are done on iCub robot platform and observes the model's performance in various scenarios. This paper shows large improvements in the model accuracy by improving the robustness using iCub robot and physical testing facilities. The proposed study lacks comparision with other models and used limited dataset recored in controlled environment with limited real-world application.

As Speech Emotion Recognition is a crucial componentof effective Human-Machine interaction, the paper 'Speech emotion recognition in emotional feedback for Human-Robot Interaction'[3] assesses six distinct types of classifiers, in orderto predict six fundamental universal emotions from non-verbalaspects of human speech. Nonverbal indicators including pitch,loudness, spectrum, and speech pace are effective emotional messengers for the majority of people. Within its constraints, the characteristics of a spoken voice likely convey important information on the speaker's emotional state. The approach used limited features and audio dataset was recorded in con-trolled environment which may not represent real-life scenario.In contrast to other approaches, the presented method uti-lized a self-recorded dataset in a natural, real-life environment with a situation-based approach. The audio recordings were conducted in English language with an Indian accent, where a selected dialogue was spoken in different emotions, depicting specific situations. Multiple audio features were extracted from the dataset, and data augmentation was performed toincrease its robustness. To further evaluate the effectivenessof the proposed approach, several machine learning modelswere compared and analyzed. This novel approach is expectedto provide significant contributions towards the development of more accurate and efficient speech emotion recognition systems.

III. PROPOSED METHODOLOGY

The proposed methodology for our research paper on the Sentiment Predictor for Stress Detection Using Voice Stress Analysis is as follows:

1) Data Collection: Recorded a dataset based on four situations, where a dialogue per situation and recorded it in different emotions. The recordings were grouped into three categories: stressed, unstressed, and neutral.
2) Data Preprocessing: The audio files were preprocessed by removing blank files and trimming the audio part where the amplitude was less than the threshold.

Some visualizations were made to check for any abnormalities in the data. Also data augmentation techniques were used like adding noise to make the model more robust.

- Data augmentation : It is a technique employed in machine learning to artificially expand a dataset by applying various transformations to the existing data. In the context of speech emotion recogni- tion, data augmentation involves creating additional training samples by introducing variations in pitch, speed, or adding background noise to the original audio recordings. This process helps improve the model's robustness by exposing it to a wider range of potential input variations, ultimately enhancingits ability to generalize and perform better on unseendata.

3) Feature Extraction: Various features were extraced from the preprocessed audio files, including Mel-frequency cepstral coefficients (MFCC), zero-crossing rate, Mel- Spectrogram, RMS value, and chroma STFT.

- MFCC : Mel-frequency cepstral coefficients(MFCC) are a type of feature extraction technique widely used in speech and audio signal processingto represent the spectral envelope of a signal.
- ZCR : Zero-crossing rate (ZCR) is a feature that calculates the number of times a signal crosses the horizontal axis (zero amplitude) per unit of time, used for pitch estimation and onset detection.
- MelSpectrogram : It is a visual representation ofthe spectrum of a signal, with the frequency scale warped to match the non-linear human perceptionof pitch.
- RMS : Root Mean Square value is a measure of the overall power of a signal that is commonly used for loudness normalization and compression.
- Chroma STFT : It is a feature that calculates the energy distribution of audio across the frequency spectrum.

4) Data Preparation: Prepared the data for model training by scaling the features and splitting the data into training and testing sets.

5) Model Training: Used different machine learning mod- els, including SVM, Random Forest, Logistic Regres- sion, Decision Tree, and CNN, to train our dataset. Also used grid search to choose the hyperparameters for Decision Tree and Random Forest models.

6) Model Evaluation: Evaluated the performance of each model by calculating metrics such as accuracy, preci- sion, recall, F1-score, and confusion matrix. And also compared the performance of the models and identified the best-performing model.

7) Conclusion: Conclusions were drawn based on our results and discuss the potential applications of the Sentiment Predictor for Stress Detection Using Voice Stress Analysis.

IV. SYSTEM ARCHITECTURE

The proposed system architecture for the Sentiment Predic- tor for Stress Detection Using Voice Stress Analysis is shown in Fig. 2. The system consists of two main components: the web interface and the machine learning model.

The user will upload an audio file to the web interface for analysis of the emotion state. The uploaded audio file will be preprocessed, and the machine learning model will be used in the backend to generate an output. The output will consist of a spectrogram, waveplot, and detected emotion state.

For building the machine learning model, a dataset wascreated based on four situations, with a dialogue per situation recorded in different emotions. The dataset includes audio fileswith emotions such as sad, happy, angry, and neutral, but theseemotions were categorized into three main categories: stressed(including sad and angry emotions), unstressed (including happy emotions), and neutral. In the preprocessing step, blank audio files were removed, and the audio parts with amplitude lower than a threshold were trimmed. Visualization and data exploration were performed, and data augmentation was carried out by adding noise to the audio files to make the model more robust. The librosa librarywas used for visualization of audio waveform and spectrogram.

Fig. 1. Flow Diagram

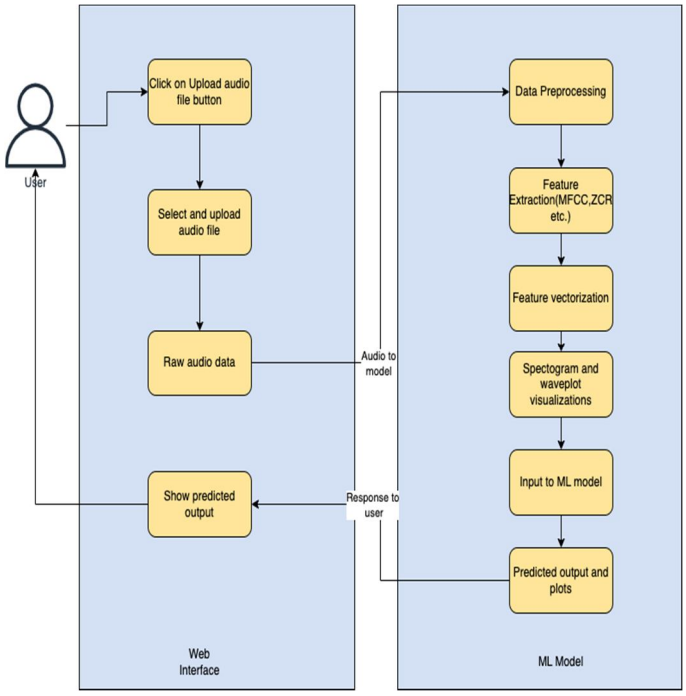


Fig. 2. System Architecture

Next, feature extraction was performed using the librosa library to extract features such as zero crossing rate, chroma stft, MFCC, and MelSpectogram from the audio files in the dataset. A feature vector was created from the extracted features, which was used as input to the classifiers.

In the next step, the data was prepared for model training and testing using python libraries such as scikit-learn, keras, and tensorflow. Classifiers such as Logistic Regressor, SVM, Decision Tree, Random Forest, and CNN were used to train the model. The performance of the models was compared to select the one best suited for the application.

The web application was built using Streamlit, and the finalgenerated model was integrated with the web application to produce the results. The user uploads an audio file to the web application, which is preprocessed and feature-extracted. The feature vector is then input to the model, and the final results are generated.

Overall, the proposed system architecture provides an end- to-end solution for the Sentiment Predictor for Stress Detection Using Voice Stress Analysis, with a user-friendly web inter- face and a robust machine learning model for emotion state analysis.

V. DATASET DESCRIPTION

The dataset used in this project is recorded by the authors ofthe paper itself. The dataset is based on situational approach. Four situations are selected with a corresponding dialogue as follows:

1) Situation: Passing a exam.
Dialogue: ”Finally I passed the exam.”
2) Situation: Someone fell from cycle.
Dialogue: ”Look, someone fell from bicycle there.”
3) Situation: Getting admission in PICT. Dialogue: ”I got admission in PICT.”
4) Situation: Welcoming guests to home.
Dialogue: ”It’s so nice to see you after such a long time.”

This dataset contains 256 files such that each of the 4 actors (male) have recorded 4 dialogues each for a particularsituation each in 4 different emotions and each in 4 different variations. The statements are in English language and spoken in Indian accent. Speech includes happy, sad, angry and neutral expressions. All the audio files have single channel withframerate of 44.1 kHz or 48 kHz recorded in .wav format.

Each file is named in the following format : (Name)(VariationNo)(SituationNo) (EmotionName).wav

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XII Dec 2024- Available at www.ijraset.com

VI. RESULTS

Twenty percent of the data was allocated for testing, while the remaining 80% was utilized for training the models. Logistic Regression, SVM, Decision Tree, Random Forest,and CNN models were employed for classification, and their performance was subjected to comparative analysis. The test set has a shape of 103 x 162, while the training set has a shapeof 409 x 162.

TABLE I
ACCURACY OF DIFFERENT ML MODELS

Model	Stressed	Neutral	Unstressed	Accuracy
Logistic Regression	0.72	0.63	0.88	74.76%
SVM	0.71	0.67	0.88	74.76%
Decision Tree	0.83	0.65	0.84	77.67%
Random Forest	0.91	1.00	0.94	94.20%
CNN	0.84	0.81	0.97	87.00%

The Random Forest and CNN models demonstrated superiorperformance, achieving 87% and 94.20% accuracy, respec- tively. The CNN model's accuracy was relatively lower dueto insufficient data for training a neural network. Notably, all models exhibited higher accuracy in predicting the unstressed class.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 160, 64)           256

 max_pooling1d (MaxPooling1D  (None, 80, 64)           0
 )

 conv1d_1 (Conv1D)           (None, 78, 128)           24704

 max_pooling1d_1 (MaxPooling  (None, 39, 128)          0
 1D)

 conv1d_2 (Conv1D)           (None, 37, 256)           98560

 max_pooling1d_2 (MaxPooling  (None, 18, 256)          0
 1D)

 conv1d_3 (Conv1D)           (None, 16, 512)           393728

 max_pooling1d_3 (MaxPooling  (None, 8, 512)           0
 1D)

 flatten (Flatten)           (None, 4096)              0

 dense (Dense)               (None, 512)               2097664

 dropout (Dropout)           (None, 512)               0

 dense_1 (Dense)             (None, 3)                 1539

=================================================================
Total params: 2,616,451
Trainable params: 2,616,451
Non-trainable params: 0
_________________________________________________________________
```

Fig. 3. CNN Model Summary

VII. CONCLUSION AND FUTURE SCOPE

A. Conclusion

In summary, the stress-detection models using various ma- chine learning frameworks effectively classified audio signals into stressed, neutral, and unstressed categories. Employing classifiers like logistic regression, SVM, decision tree, ran- dom forest, and CNN revealed the random forest classifier's superior performance with a 93% accuracy, while the CNN model achieved an 87% accuracy.

Several features, including MFCC, ZCR, melSeptogram, and RMS extracted from the audio data, contributed signifi- cantly to accurate signal classification. The application of suchfeatures holds substantial potential in psychology, healthcare, and security.

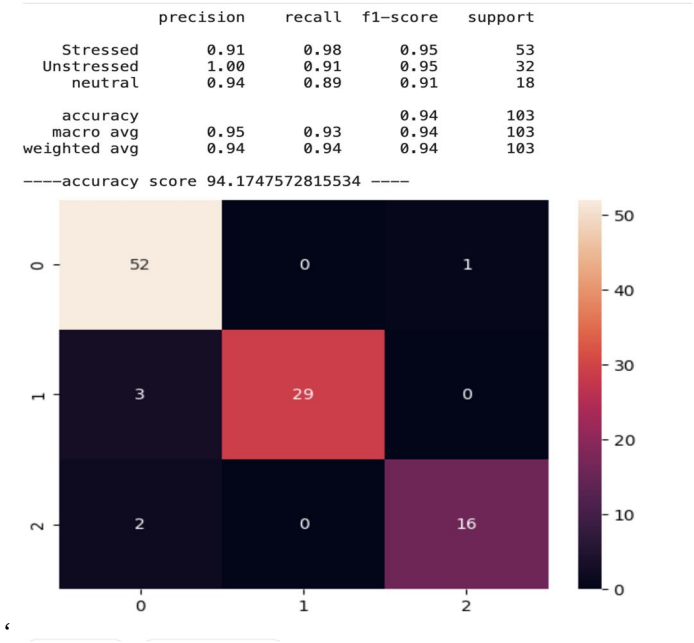


Fig. 4. Random Forest Results

Subsequent research could explore the capabilities of deep learning models in stress detection from audio signals and consider integrating other modalities like physiological signals to enhance model accuracy. Extending these models to real- time stress detection scenarios could be pivotal for effective stress management and prevention.

B. Future Scope

To enhance detection accuracy, exploring a conceptual multi-modal approach is a potential avenue for improvement. Expanding the dataset could facilitate better training for CNN,thereby yielding improved results. The application of voice stress analysis extends to clinical settings, offering potential benefits in diagnosing and treating conditions like anxiety disorders and depression. Future endeavors may center ondeveloping voice stress analysis tools designed specifically forclinical applications.

VIII. ACKNOWLEDGMENT

It is our pleasure to present the research paper titled ”Sentiment Predictor For Stress Detection Using Voice Stress Analysis.” We extend our heartfelt gratitude to our guide, Dr. B.A. Sonkamble, for his invaluable assistance and guidance throughout this research endeavor. Our sincere appreciation goes to Dr. Sarang Joshi, our esteemed reviewer, for providing valuable suggestions that significantly enhanced the quality of this work. We express our gratitude to Dr. Geetanjali Kale, Head of the Department of Computer Engineering at Pune Institute of Computer Technology, for her indispensable support and insightful suggestions.

REFERENCES

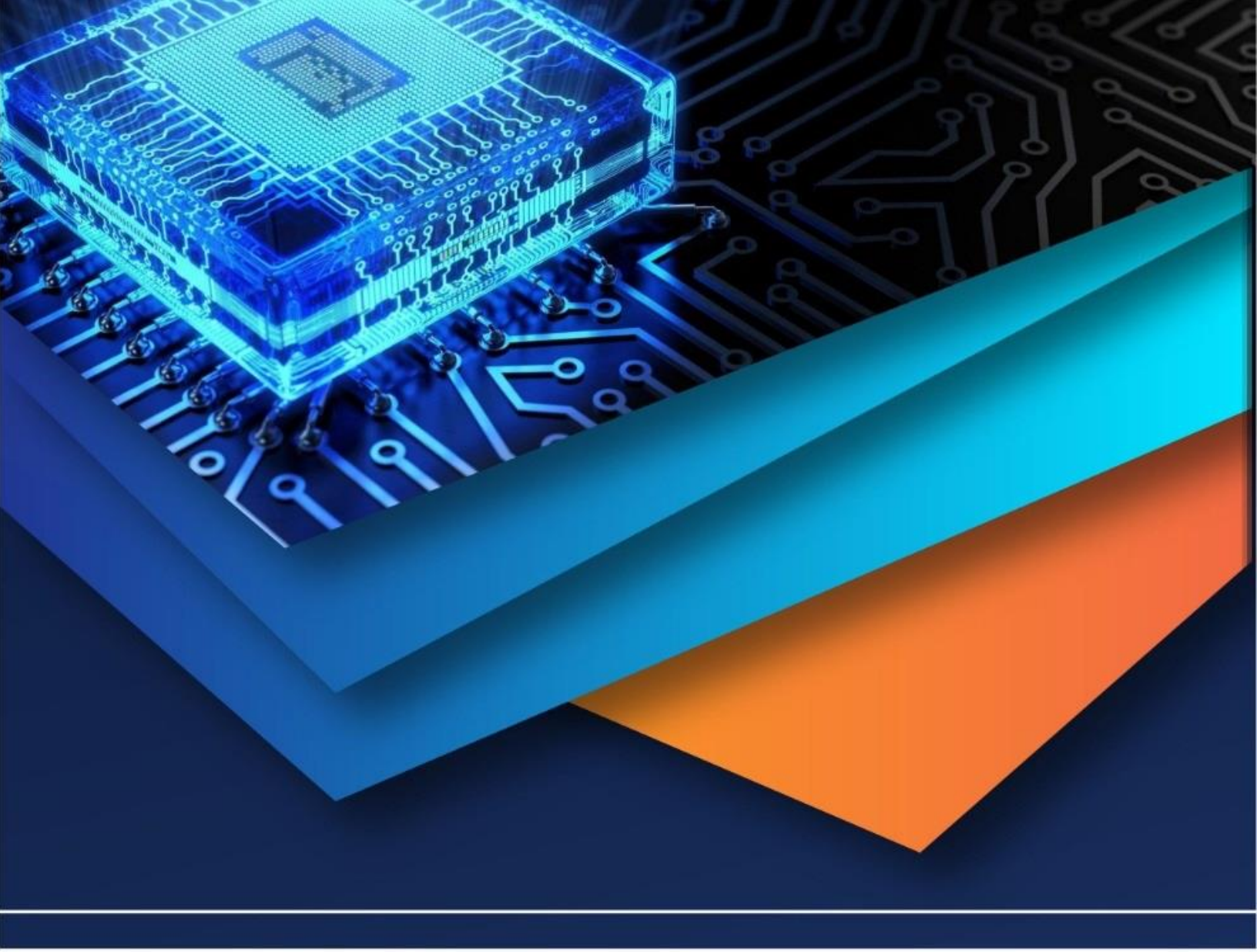
[1] Han, Hyewon Byun, Kyunggeun Kang, Hong-Goo. (2018). A Deep Learning-based Stress Detection Algorithm with Speech Signal. 11-15. 10.1145/3264869.3264875.
[2] Robustness of Speech Emotion Recognition for Human-Robot Inter- action with Deep Neural Networks.” 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE Press, pp. 854–860
[3] Rázuri, Javier Francisco Sundgren, David Rahmani, Rahim Lars- son, Aron Moran, Antonio Bonet, Isis. (2015). Speech emotion recognition in emotional feedback for Human-Robot Interaction. In-ternational Journal of Advanced Research in Artificial Intelligence. 4. 10.14569/IJARAI.2015.040204.
[4] Roopa, S. Prabhakaran, M. Betty, P.. (2019). Speech emotion recog- nition using deep learning. International Journal of Recent Technology and Engineering. 7. 247-250.
[5] Arushi, R. Dillon and A. N. Teoh, ”Real-time Stress Detection Model and Voice Analysis: An Integrated VR-based Game for Training Public Speaking Skills,” 2021 IEEE Conference on Games (CoG), 2021, pp. 1-4
[6] Tripathi, Samarth and Homayoon S. M. Beigi., 2018, “Multi-Modal emotion recoggnition on IEMOCAP with neural networks.”, eprint arXiv:1804.05788, April 201
[7] Tomba, K., Mugellini, E., ”Stress Detection Through Speech Analysis”, 15th International Joint Conference on e-Business and Telecommunica- tions - ICETE , pp. 394- 398, January 2018.

[8] Han, Kun Yu, Dong Tashev, Ivan. (2014). Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine. Proceed- ings of the Annual Conference of the International Speech Communi- cation Association, INTERSPEECH. 10.21437/Interspeech.2014-57.

[9] Yu, C., Tian, Q., Cheng, F., Zhang, S. (2011). Speech Emotion Recog- nition Using Support Vector Machines. In: Shen, G., Huang, X. (eds) Advanced Research on Computer Science and Information Engineering. CSIE 2011. Communications in Computer and Information Science, vol 152. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-21402-8 35

[10] Iqbal, Aseef Barua, Kakon. (2019). A Real-time Emotion Recognition from Speech using Gradient Boosting. 1-5. 10.1109/ECACE.2019.8679271.

[11] M. G. de Pinto, M. Polignano, P. Lops and G. Semeraro, "Emotions Understanding Model from Spoken Language using Deep Neural Net- works and Mel-Frequency Cepstral Coefficients," 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Bari, Italy, 2020, pp. 1-5, doi: 10.1109/EAIS48028.2020.9122698.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY