VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Project Report on

Audio Emotion Recognition using Digital Stethoscope and Deep Learning Techniques

In partial fulfilment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2023-24

Submitted by
Ritika Hotwani (D17B 27)
Vivek Balani (D17B 05)
Yash Pahlani (D17B 49)
Prabha Pamula (D17B 51)

Project Mentor
Mrs. Mannat Doultani

(2023-24)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
An Autonomous Institute Affiliated to University of Mumbai
Department of Computer Engineering

Certificate

This is to certify that ***Ritika Hotwani, Vivek Balani, Yash Pahlani, Prabha Pamula*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "***Audio Emotion Recognition Using Digital Stethoscope and Deep Learning Techniques***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Prof.Mannat Doultani*** in the year 2023-24 .

This thesis/dissertation/project report entitled (***Audio Emotion Recognition Using Digital Stethoscope and Deep Learning Techniques***) by (***Ritika Hotwani, Vivek Balani, Yash Pahlani, Prabha Pamula***) is approved for the degree of Bachelor of Engineering in Computer Engineering.

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12 PSO1, PSO2	

Date:

Project Guide:

Project Report Approval
For
B. E (Computer Engineering)

This project report entitled (***Audio Emotion Recognition Using Digital Stethoscope and Deep Learning Techniques***) by (***Ritika Hotwani, Vivek Balani, Yash Pahlani, Prabha Pamula***) is approved for the degree of Bachelor of Engineering in Computer Engineering.

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---	---
Ritika Hotwani (27)	Vivek Balani (05)
---	---
Yash Pahlani (49)	Prabha Pamula (51)

Date:

ACKNOWLEDGEMENT

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Prof. Mannat Doultani** for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.) Nupur Giri** and our Principal **Dr. (Mrs.) J.M. Nair ,** for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is a great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering.

We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement several times.

Computer Engineering Department
COURSE OUTCOMES FOR B.E PROJECT

Learners will be to,

Course Outcome	Description of the Course Outcome
CO 1	Able to apply the relevant engineering concepts, knowledge and skills towards the project.
CO2	Able to identify, formulate and interpret the various relevant research papers and to determine the problem.
CO 3	Able to apply the engineering concepts towards designing solutions for the problem.
CO 4	Able to interpret the data and datasets to be utilised.
CO 5	Able to create, select and apply appropriate technologies, techniques, resources and tools for the project.
CO 6	Able to apply ethical, professional policies and principles towards societal, environmental, safety and cultural benefit.
CO 7	Able to function effectively as an individual, and as a member of a team, allocating roles with clear lines of responsibility and accountability.
CO 8	Able to write effective reports, design documents and make effective presentations.
CO 9	Able to apply engineering and management principles to the project as a team member.
CO 10	Able to apply the project domain knowledge to sharpen one's competency.
CO 11	Able to develop a professional, presentational, balanced and structured approach towards project development.
CO 12	Able to adopt skills, languages, environment and platforms for creating innovative solutions for the project.

Index

Title: ***Audio Emotion Recognition Using Digital Stethoscope and Deep Learning Techniques***

Chapter No.	Title	Page No.
	Abstract	**11**
1	**Introduction**	12
	1.1 Introduction	12
	1.2 Motivation	13
	1.3 Problem Definition	13
	1.4 Existing Systems	13
	1.5 Lacuna of the existing systems	14
	1.6 Relevance of the Project	14
2.	**Literature Survey**	15
	A. Brief Overview of Literature Survey	15
	B. Related Works	15
	2.1 Research Papers Referred	16
	a. Abstract of the research paper	
	b. Inference drawn	
	2.2 Patent search	24
	Links 1. European Patent	
	2. US patent	
	2.3. Inference drawn	25
	2.4 Comparison with the existing system	25
3.	**Requirement Gathering for the Proposed System**	26
	3.1 Introduction to requirement gathering	26
	3.2 Functional Requirements	26
	3.3 Non-Functional Requirements	27
	3.4.Hardware, Software , Technology and tools utilised	27
	3.5 Constraints	28
4.	**Proposed Design**	29
	4.1 Block diagram of the system	29
	4.2 Modular design of the system	30
	4.3 Detailed Design	31
	4.4 Project Scheduling & Tracking using Timeline / Gantt Chart	32

5.	Implementation of the Proposed System 5.1. Methodology employed for development 5.2 Algorithms and flowcharts for the respective modules developed 5.3 Datasets source and utilisation	33 33 34 36
6.	Testing of the Proposed System 6.1 . Introduction to testing 6.2. Types of tests Considered 6.3 Various test case scenarios considered 6.4. Inference drawn from the test cases	37 37 37 37 39
7.	Results and Discussion 7.1. Screenshots of User Interface (UI) for the respective module 7.2. Performance Evaluation measures 7.3. Input Parameters / Features considered 7.4. Graphical and statistical output 7.5. Comparison of results with existing systems 7.6. Inference drawn	40 40 41 41 42 43 43
8.	Conclusion 8.1 Limitations 8.2 Conclusion 8.3 Future Scope	44 44 44 45

List of Figures:

Figure No.	Page No.	Figure Name	Description
1	16	The structure of speech emotion recognition system	The structure of speech emotion recognition system for the 'Study on Speech Emotion Recognition System in E-Learning' paper.
2	16	Comparison of the Results	Results comparison graph for the 'Study on Speech Emotion Recognition System in E-Learning' paper.
3	17	Network Structure	Network Structure for the 'Improvement of multilingual emotion recognition method based on normalised acoustic features using CRNN' paper.
4	17	Speech emotion recognition accuracy	Accuracy distribution for the 'Improvement of multilingual emotion recognition method based on normalised acoustic features using CRNN' paper
5	18	Architecture of the program for emotion recognition by speech segments	Architecture of the emotion recognition program by speech segments.
6	19	Proposed LSTM-Transformer model	Propose model of the 'Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files' paper.
7	20	Flow chart for heart sound signal acquisition, processing and analysis	Heart sound signal flow chart for the 'The electronic stethoscope' paper.
8	20	Structure of adaptive noise canceller.	Noise canceller for the 'The electronic stethoscope' paper.
9	29	Block Diagram of the Audio emotion recognition system	Working flow of the system
10	30	Modular Diagram of Audio emotion recognition system	Modular flow of the system
11	31	Hardware Representation	Diagram representation of hardware components
12	31	Hardware Implementation	Aggregation of hardware components
13	32	Gantt Chart	Timeline of the project workflow
14	35	Flow chart for the proposed system	Frontend-Backend work flow of the

			system.
15	40	Home page UI	Results page of the website
16	40	Prediction page UI	Results page from the website
17	41	Classification report of MLP classifier	Classification results for MLP classifier
18	41	Classification report of KNN classifier	Classification results for KNN classifier
19	42	Graphical output on UI	Predicted emotions probabilities output visualisation in the form of graphs

List of Abbreviations

1. ER: Emotion Recognition
2. SER: Speech Emotion Recognition
3. SAD: Sadness
4. ANG: Anger
5. SUR: Surprise
6. HAP: Happiness
7. NEU: Neutral

Abstract

Emotion Recognition is a trending research topic these days, with its main motive to improve human-machine interaction. As human beings, speech is the most natural way to express oneself. Emotions can be determined considering various factors like speech, facial expressions, and body language. Also, Auscultation tests are performed to listen to the sounds of internal organs like lungs, heart, etc using a stethoscope. The existing classifiers are found to be good at end-to-end audio classification models, aiming to learn a direct mapping from audio spectrograms to corresponding labels.

The proposed idea is to define an Emotion Recognition (ER) system as a collection of methodologies by correlating audio signals and heartbeats of an individual for the detection of various emotions. The system aims to extract the most prominent features from the input audio file imported using a microphone and the sounds of the heart is measured via a digital stethoscope. As emotions bring a greater impact on the functionality of heart rate, the recognizer is allowed to view the auscultation signal in the form of wave plots (heart rate), the pitch diagram (for input speech), and the resultant of both will help in the classification of individual emotions using deep learning techniques. Such a system can be used in a wide variety of application areas like interactive voice-based-assistant or caller-agent conversation and analysis of human mental and heart health prediction and human-robot interaction.

Chapter 1: Introduction

1.1 Introduction

Emotion Recognition (ER) can be defined as the extraction of the emotional state of the speaker. Emotions play a vital role in communication, and the detection and analysis of the same are of vital importance in today's digital world of remote communication. Emotion detection(via audio) is a challenging task because emotions are subjective. There is no common consensus on how to measure or categorise them. Most acoustic features that have been used for emotion recognition can be divided into two categories: prosodic and spectral. Prosodic features have been shown to deliver recognition, including intonation, accent, pitch, mute, and rate of speech. Spectral features convey the frequency content of the speech signal and provide complementary information to prosodic features. When a speaker expresses an emotion while adhering to an inconspicuous intonation pattern, human listeners can nevertheless perceive the emotional information through the pitch and intensity of speech.

It is well known that physiological signals are affected by the human endocrine system and the autonomic nervous system. These systems are less affected by human subjective consciousness and can reflect the real emotional state more objectively and accurately. Compared with facial expressions and limb movements, the heart rate based result is more objective and the actual emotions are not easy to be hidden. Studies have shown that the heart rate varies with mood changes. It proved that physiological signals had unique responses to different emotions. The heart rate increased significantly when people were angry or scared. The effects of relaxation and fear on heart rate were significantly different, and the average heart rate during happiness was lower than that while in a sad state. From this perspective, emotion recognition based on physiological signals makes the results more objective.

We aim to capture the diverse acoustic cues that are in the speech signal and analyse their mutual relationship to the speaker's emotion also considering the physiological factor (Heart rate). We propose a technique to recognize several basic emotions, namely sadness (SAD),

anger (ANG), surprise (SUR), happiness (HAP), and neutral (NEU), based on the analysis of phonetic and acoustic properties.

1.2 Motivation

This project is motivated by the growing importance of emotion recognition in digital communication. In an era of remote interactions, accurately detecting and interpreting emotions through speech and physiological signals is crucial for improving customer service, education, healthcare, and more. By combining acoustic and physiological data, we aim to create a reliable emotion recognition system that overcomes subjectivity, advances technology, and addresses ethical concerns, ultimately enhancing the quality of digital interactions.

1.3 Problem Definition

Emotion Recognition (ER) is a challenging task in the field of digital communication due to the subjective nature of emotions and the lack of a common consensus on their measurement and categorization. Current approaches often rely on acoustic features, specifically prosodic and spectral cues, to detect emotions from audio. However, these methods may not fully capture the emotional state of the speaker, leading to potentially inaccurate results. To address this limitation, we propose a technique that incorporates physiological signals, such as heart rate, which can provide a more objective and accurate representation of the speaker's emotional state. The goal of this project is to develop a robust emotion recognition system capable of identifying emotions (sadness, anger, surprise, happiness, and neutral) by analysing phonetic and acoustic properties of the speech signal while considering the speaker's physiological factors, ultimately enhancing the effectiveness of emotion detection in remote communication scenarios.

1.4 Existing Systems

Existing systems for emotion recognition encompass a variety of approaches, ranging from rule-based systems to sophisticated machine learning models. Many of these systems focus on analysing speech signals to detect emotional cues, utilising features such as pitch, intensity, and duration to classify emotions. Facial expression analysis is another common modality, with systems employing computer vision techniques to identify facial muscle movements indicative of different emotions. Additionally, some systems integrate

physiological signals, such as heart rate and galvanic skin response, to provide a more comprehensive understanding of the user's emotional state. These existing systems often achieve reasonable accuracy in controlled laboratory settings but may struggle to generalise effectively to real-world environments due to factors such as variability in emotional expression and noise in input signals.

Despite advancements in emotion recognition technology, existing systems face several challenges. One major limitation is their reliance on single modality data, which may not capture the full spectrum of emotional cues present in human communication. Furthermore, many existing systems are designed and evaluated primarily on datasets collected from Western populations, potentially limiting their applicability to diverse cultural and linguistic contexts. Addressing these limitations requires ongoing research and development efforts aimed at improving the robustness, generalizability, and ethical soundness of existing emotion recognition systems.

1.5 Lacuna of the existing systems

Many existing emotion recognition systems heavily depend on contextual information, such as the content of the speech or the speaker's identity. Most of them does not take physiological cues into account. In applications requiring real-time emotion recognition, existing systems may face computational challenges, leading to latency issues. Noise is the major concern which leads to inaccuracies in prediction.

1.6 Relevance of the Project

1. Improves the quality of remote communication by enabling a more profound understanding of emotions beyond words.
2. Emotion recognition technology has wide-ranging applications in artificial intelligence, chatbots, virtual assistants, and customer service. By incorporating emotional understanding into AI systems, we can create more empathetic and effective digital interactions, enhancing user experience and customer satisfaction.
3. The ability to detect emotions through physiological signals, such as heart rate, can have significant implications for mental health monitoring. It can aid in early detection of emotional distress, stress, or anxiety, enabling timely interventions and support.
4. While subjective assessments of emotions can vary from person to person, physiological signals like heart rate offer a more objective measure of emotional

states. This objectivity can be especially valuable in contexts where reliable emotion recognition is essential, such as in law enforcement or healthcare.

Chapter 2: Literature Survey

A. Brief Overview of Literature Survey

Speech Emotion Recognition (SER) is a field that analyses speech data to identify the emotions conveyed. Researchers are constantly developing new methods to improve accuracy.

Two of the papers (1 & 4) explore using a combination of audio and text data for SER. Paper 1 introduces a model with separate analysis of audio and text data using dual recurrent neural networks, achieving good results (71.8% accuracy on 4 emotions). Paper 4 goes a step further with a "multi-hop attention mechanism" that analyses the connection between audio and text, leading to a significant improvement (6.5% increase) in accuracy.

On the other hand, papers 2 and 3 focus on using audio data only. Paper 2 utilises deep learning (CNNs) with audio features like MFCCs to achieve an overall F1 score of 0.85 on 8 emotions. Paper 3, a survey paper, highlights methods like Support Vector Machines (SVMs) with MFCCs reaching over 95% accuracy in some cases. It also identifies limitations of current methods and promising areas for future research.

B. Related Works

"Emotion recognition from speech using gender-specific models" by Schuller et al. (2010): This study investigates the impact of gender-specific models on speech emotion recognition. The authors found that using gender-specific models improved the recognition accuracy, especially for female speakers.

"Speech emotion recognition based on fusion of spectral and prosodic features" by Yella et al. (2020): This study proposes a method for fusing spectral and prosodic features to improve speech emotion recognition. The authors achieved high recognition rates on several datasets.

"Speech emotion recognition using ensemble classifiers" by Liu et al. (2018): This study explores the use of ensemble classifiers, such as random forests and AdaBoost, for speech emotion recognition. The authors achieved high recognition rates on several datasets.

"A comparative study of feature extraction techniques for speech emotion recognition" by Patil et al. (2019): This study compares the performance of several feature extraction techniques, such as Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) features, for speech emotion recognition. The authors found that MFCCs performed better than other techniques on several datasets.

2.1. Research Papers Referred

[1] A. Zhu and Q. Luo, "Study on Speech Emotion Recognition System in E-Learning", presented at International Conference on Human-Computer Interaction, vol 4552, Springer, Berlin, Heidelberg, 2007, https://doi.org/10.1007/978-3-540-73110-8_59. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-3-540-73110-8_59.pdf?pdf=inline%20link

Abstract: The article discusses a study on speech emotion recognition systems in E-Learning. The authors propose a system to address the emotion deficiency in current E-Learning systems by collecting a corpus of emotional speech from various subjects speaking different languages. The potential prosodic features are extracted from each utterance for the computational mapping between emotions and speech patterns. The discriminatory power of these features is then analysed using a systematic approach. The study shows that prosodic cues are very powerful signals of human vocal emotions, and the proposed system has demonstrated that the emotional expression in speech can be identified beyond the language boundaries. The article concludes by suggesting that multi-module emotion recognition, including facial and other features such as gesture, will be studied in the future to obtain higher emotion recognition accuracy.

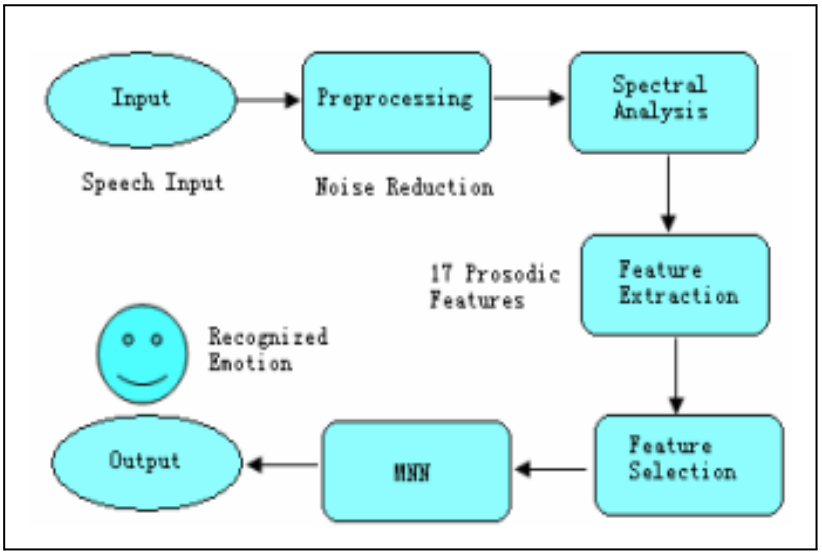

Fig. 1 The structure of speech emotion recognition system

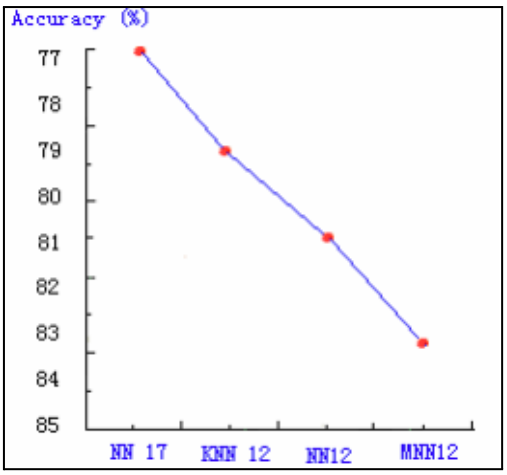

Fig.2 Comparison of the Results

Inference: Study on speech emotion recognition system in E-Learning. The authors propose a system to address the emotion deficiency in current E-Learning systems by collecting a corpus of emotional speech from various subjects speaking different languages. The potential prosodic features are extracted from each utterance for the computational mapping between emotions and speech patterns. The discriminatory power of these features is then analysed using a systematic approach. The study shows that prosodic cues are very powerful signals of human vocal emotions, and the proposed system has demonstrated that the emotional expression in speech can be identified beyond the language boundaries. The article concludes by suggesting that multi-module emotion recognition, including facial and other features such as gesture, will be studied in the future to obtain higher emotion recognition accuracy.

[2] M. Suzukia and J. Qib, "Improvement of multilingual emotion recognition method based on normalised acoustic features using CRNN", Procedia Computer Science, vol. 207, pp. 684-691, 2022, https://doi.org/10.1016/j.procs.2022.09.123. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922010043

Abstract: The Research paper titled Investigation of Speech Emotion Recognition Using Synthesised Speech Data. The paper discusses the effects of sex-based differences and length differences between emotional and synthesised speeches on speech emotion recognition. The experiment used female speakers for all synthesised speech data and male speakers for emotional speech data. The paper presents the results of the investigation, including the percentage of male speakers in recognized speech and the relationship between speech length and recognition results. The study found that even if the gender of the synthesised speech differs from that of the emotional speech, normalisation is not significantly affected. However, if the lengths of the frames are significantly different between emotional and synthesised speeches, unreasonable completion may occur, and correct normalisation may not be performed.

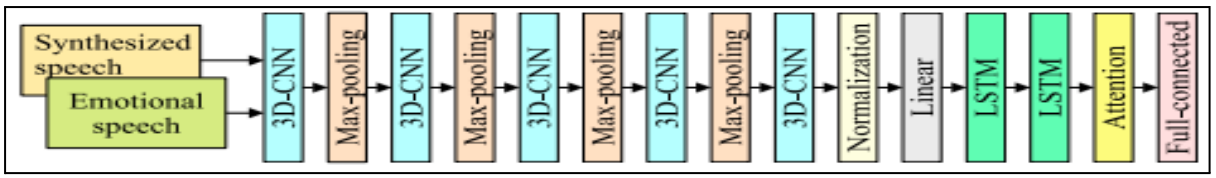

Fig. 3 Network Structure

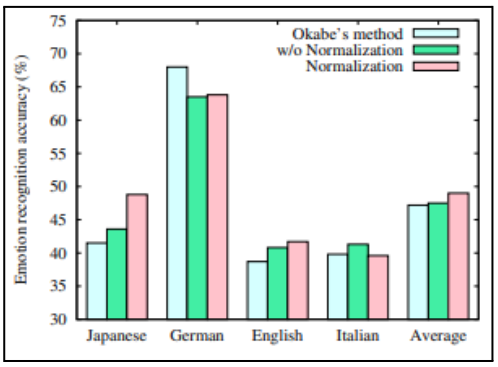

Fig. 4 Speech emotion recognition accuracy

Inference: The article "Research on a voice emotion recognition system in online learning" proposes a voice emotion recognition system to overcome the lack of emotions in current online learning systems. The paper identifies 17 potential prosodic features and uses a sequential forward selection (SFS) method with a general recurrent neural network (GRNN) to determine the discriminative ability of each feature. The paper also presents a consistency-based approach to evaluate the suitability of features around and beyond the minimum error point. The paper concludes by proposing a modular neural network (MNN) architecture for emotion recognition in speech, consisting of six subnetworks, each specialising in a specific emotion class. The paper reports an overall correct recognition rate of 80.69% on 12 selected features using the proposed method.

[3] A.V.Tsaregorodtseva et al.,"The architecture of the emotion recognition program by speech segments", Procedia Computer Science, vol. 213, pp. 338-345, 2022, https://doi.org/10.1016/j.procs.2022.09.123. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922017677

Abstract: The article discusses the issues of automatic recognition of the emotional colouring of a person's utterance using machine learning methods. In the software being developed, a change in the basic tone of a person's voice is taken as a basis. The architecture of the program is based on a multilayer perceptron. The result of the classifier is the choice of one of eight possible classes of emotions. In the course of the study, the accuracy of the classification of emotions was determined and the main directions of the development of the program were outlined.

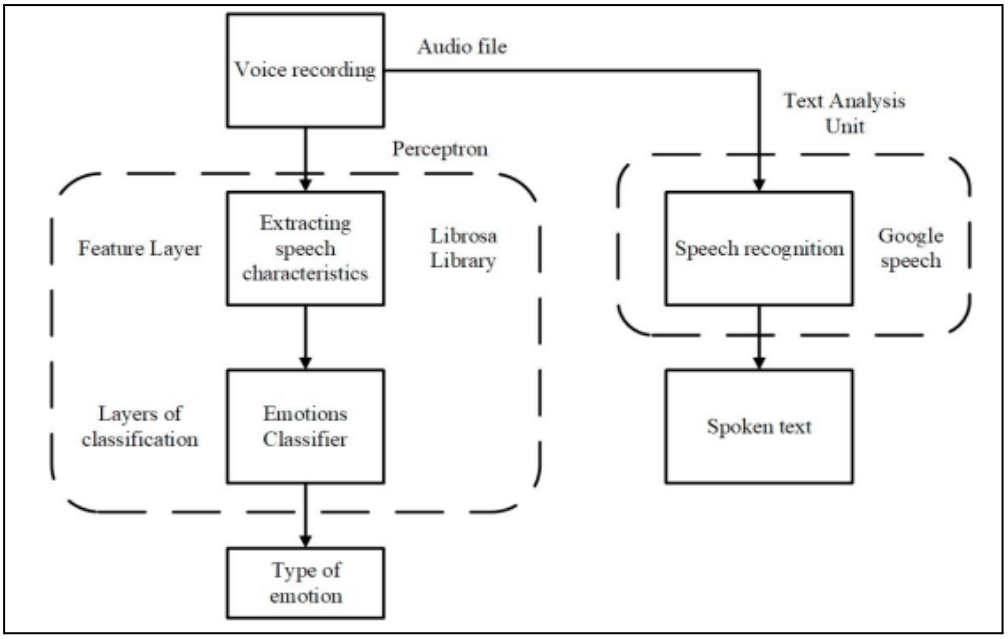

Fig. 5 Architecture of the program for emotion recognition by speech segments

Inference: The research paper introduces a speech emotion recognition program using machine learning, focusing on the significance of recognizing emotions in various applications. It employs a multilayer perceptron (MLP) for classification, with a two-part architecture comprising voice recording and speech recognition. While successfully recognizing eight emotions namely “calm”; “happy”; “sad”; “angry"; “scared"; “disgust"; “surprise“ and ”neutral voice“, the paper acknowledges challenges, including sensitivity to noise and individual speech characteristics, speech tempo effects, and processing speed. Overall, the program demonstrates potential for emotion recognition but requires further refinement to address limitations and improve performance.

[4] F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files" in IEEE Access, vol. 10, pp. 36018-36027, 2022, doi: 10.1109/ACCESS.2022.3163856. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9745599

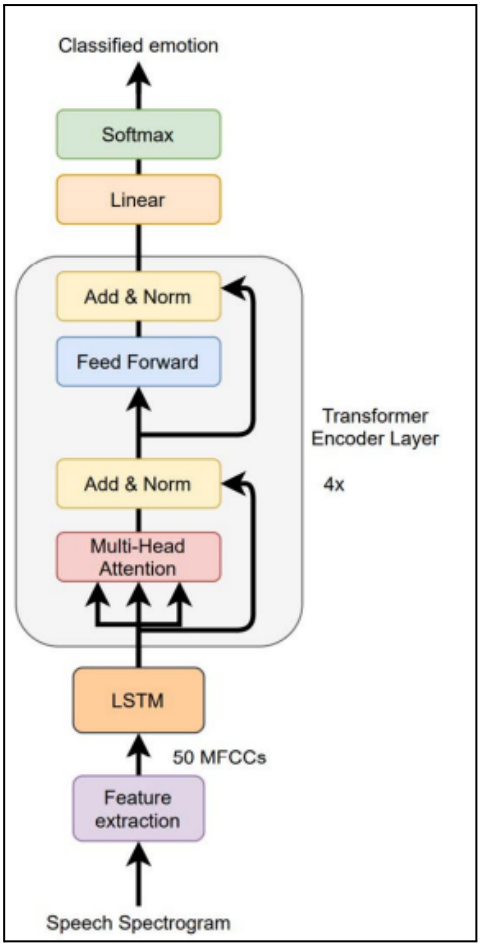

Fig. 6 Proposed LSTM-Transformer model

Abstract: Emotion is a vital component in daily human communication and it helps people understand each other. Emotion recognition plays a crucial role in developing human-computer interaction and computer-based speech emotion recognition. In a nutshell, Speech Emotion Recognition (SER) recognizes emotion signals transmitted through human speech or daily conversation where the emotions in a speech strongly depend on temporal information. Despite the fact that much existing research showed that a hybrid system performs better than traditional single classifiers used in SER, there are some limitations in each of them. As a result, this paper discussed a proposed hybrid Long Short-Term Memory (LSTM) Network and Transformer Encoder to learn the long-term dependencies in speech signals and classify emotions. Speech features are extracted with Mel Frequency Cepstral Coefficient (MFCC) and fed into the proposed hybrid LSTM-Transformer classifier. A range of performance evaluations was conducted on the proposed LSTM-Transformer model. The results indicate that it achieves a significant recognition improvement compared with existing models offered by other published works. The proposed hybrid model reached 75.62%, 85.55%, and 72.49% recognition success with the RAVDESS, Emo-DB, and language-independent datasets.

Inference: The research paper presents a hybrid Long Short-Term Memory (LSTM) and Transformer Encoder model for Speech Emotion Recognition (SER). The model achieves significant recognition improvements across different datasets, including RAVDESS,

Emo-DB, and a language-independent dataset. It attains recognition rates of 75.62% (RAVDESS), 85.55% (Emo-DB), and 72.49% (language-independent dataset).

[5] S. Leng, R. Tan, K. Chai et al., "The electronic stethoscope", BioMed Eng OnLine 14, 66, 2015, doi: 10.1186/s12938-015-0056-y. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26159433/

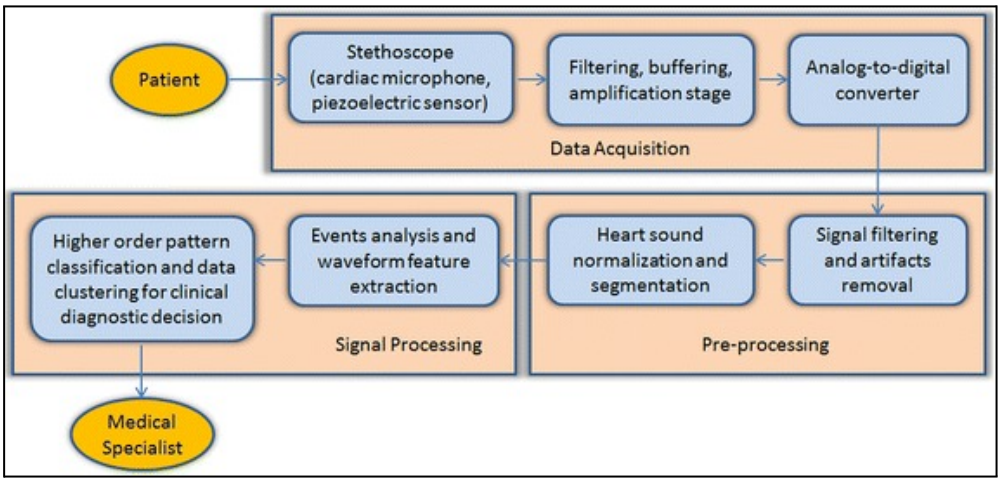

Fig. 7 Flow chart for heart sound signal acquisition, processing and analysis

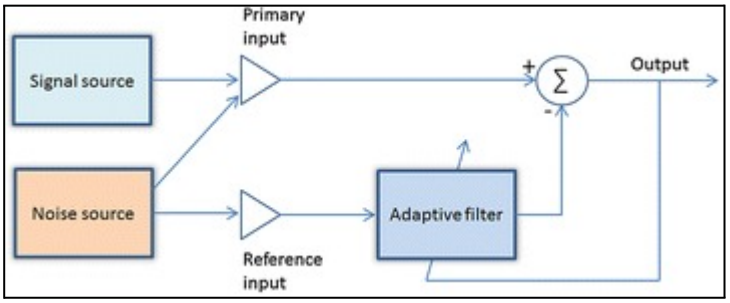

Fig. 8 Structure of adaptive noise canceller.

Abstract: Most heart diseases are associated with and reflected by the sounds that the heart produces. Heart auscultation, defined as listening to the heart sound, has been a very important method for the early diagnosis of cardiac dysfunction. Traditional auscultation requires substantial clinical experience and good listening skills. The emergence of the electronic stethoscope has paved the way for a new field of computer-aided auscultation. This article provides an in-depth study of (1) the electronic stethoscope technology, and (2) the methodology for diagnosis of cardiac disorders based on computer-aided auscultation. The paper is based on a comprehensive review of (1) literature articles, (2) market (state-of-the-art) products, and (3) smartphone stethoscope apps. It covers in depth every key component of the computer-aided system with electronic stethoscope, from sensor design, front-end circuitry, denoising algorithm, heart sound segmentation, to the final machine learning techniques. Our intent is to provide an informative and illustrative presentation of

the electronic stethoscope, which is valuable and beneficial to academics, researchers and engineers in the technical field, as well as to medical

Inference: It highlights the shift from traditional auscultation, which relies on clinical experience, to computer-aided auscultation enabled by electronic stethoscopes. The paper thoroughly reviews the technology and methodology involved in this approach, covering aspects such as sensor design, noise reduction algorithms, heart sound segmentation, and machine learning techniques. The intent is to provide a detailed understanding of electronic stethoscopes, benefiting both technical experts and medical professionals. Furthermore, the paper emphasises the potential for developing real-time integrated systems for heart sound detection and quantification, with implications for improved cardiac diagnosis and healthcare.

[6] A. A. Varghese, J. P. Cherian and J. J. Kizhakkethottam, "Overview on emotion recognition system," 2015 International Conference on Soft-Computing and Networks Security (ICSNS), 2015, pp. 1-5, doi: 10.1109/ICSNS.2015.7292443. [Online]. Available: https://ieeexplore.ieee.org/document/7292443

Abstract: Human emotion recognition plays an important role in the interpersonal relationship. The automatic recognition of emotions has been an active research topic from early eras. Therefore, there are several advances made in this field. Emotions are reflected from speech, hand and gestures of the body and through facial expressions. Hence extracting and understanding of emotion has a high importance of the interaction between human and machine communication. This paper describes the advances made in this field and the various approaches used for recognition of emotions. The main objective of the paper is to propose real time implementation of emotion recognition systems.

Inference: In the paper “Overview on Emotion Recognition System”, various techniques are explored. Matthew S. Ratliff and Eric Patterson used Active Appearance Models (AAM) to classify emotions based on facial expressions. They trained AAM on a database of faces and used feature classification schemes to identify basic emotions. Michel F. Valstar and Maja Pantic proposed a method for recognizing a broader range of facial behaviours using Action Units (AUs) and temporal characteristics. This method can recognize 22 AUs and temporal segments, but it has limitations with face view angles and occlusions. Another approach combines facial expressions and speech for emotion recognition, improving accuracy and robustness. It uses a database of posed expressions, limiting real-time applications. Ginevra Castellano et al. presented a multimodal approach combining facial expressions, body gestures, and speech to recognize eight emotions, achieving a 10% improvement over unimodal systems. However, certain prerequisites and limitations apply to the actors' appearance and environment

[11] Cheng Xiefeng, Yue Wang, Shicheng Dai, Pengjun Zhao & Qifa Liu Scientific Reports volume 9, Article number: 6486 (2019) “Heart sound signals can be used for emotion recognition”
Available: https://www.nature.com/articles/s41598-019-42826-2

Abstract : This article studies whether heart sound signals can be used for emotion recognition. First, we built a small emotion heart sound database, and simultaneously recorded the participants' ECG for comparative analysis. Second, according to the characteristics of the heart sound signals, two emotion evaluation indicators were proposed: HRV of heart sounds (difference between successive heartbeats) and DSV of heart sounds (the ratio of diastolic to systolic duration variability). Then, we extracted linear and nonlinear features from two emotion evaluation indicators to recognize four kinds of emotions. Moreover, we used valence dimension, arousal dimension and valence-arousal synthesis as evaluation standards. The experimental results demonstrated that heart sound signals can be used for emotion recognition. It was more effective to achieve recognition results by combining the features of HRV and DSV of heart sounds. Finally, the average accuracy of four emotion recognitions on valence dimension, arousal dimension and valence-arousal synthesis was up to 96.875%, 88.5417% and 81.25%, respectively.

Inference: this paper explores the potential of heart sound signals for emotion recognition, emphasising their non-invasive acquisition and ability to reflect both chronotropic and inotropic aspects of heart activity. It highlights the lack of empirical validation for heart sound-based emotion recognition and proposes novel indicators, heart sound HRV and heart sound DSV, alongside an experimental plan to construct an emotion heart sound database and record simultaneous ECG data. The survey outlines the methodology for feature extraction from these indicators and presents a block diagram of the proposed emotion recognition system, underscoring its potential application in mental health assessment and signalling a promising direction for future research and validation.

[13] Wang L, Hao J, Zhou TH. ECG Multi-Emotion Recognition Based on Heart Rate Variability Signal Features Mining. Sensors (Basel). 2023 Oct 22;23(20):8636. doi: 10.3390/s23208636. PMID: 37896729; PMCID: PMC10610830.
Available:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10610830/#B26-sensors-23-08636

Abstract : Heart rate variability (HRV) serves as a significant physiological measure that mirrors the regulatory capacity of the cardiac autonomic nervous system. It not only indicates the extent of the autonomic nervous system's influence on heart function but also unveils the connection between emotions and psychological disorders. Currently, in the field of emotion recognition using HRV, most methods focus on feature extraction through the comprehensive analysis of signal characteristics; however, these methods lack in-depth analysis of the local features in the HRV signal and cannot fully utilise the information of the HRV signal. Therefore, we propose the HRV Emotion Recognition (HER) method, utilising the amplitude level quantization (ALQ) technique for feature extraction. First, we employ the emotion quantification analysis (EQA) technique to impartially assess the semantic resemblance of emotions within the domain of emotional arousal. Then, we use the ALQ method to extract rich local information features by analysing the local information in each frequency range of the HRV signal. Finally, the extracted features are classified using a logistic regression (LR) classification algorithm, which can achieve efficient and accurate emotion recognition. According to the experiment findings, the approach surpasses existing techniques in emotion

recognition accuracy, achieving an average accuracy rate of 84.3%. Therefore, the HER method proposed in this paper can effectively utilise the local features in HRV signals to achieve efficient and accurate emotion recognition. This will provide strong support for emotion research in psychology, medicine, and other fields.

Inference: The paper investigates the relationship between emotions and Heart Rate Variability (HRV) signals in Electrocardiogram (ECG) data, proposing a novel HRV Emotion Recognition (HER) model for accurate emotion classification. Emotions, crucial for human interaction and physiological regulation, are analysed through the autonomic nervous system (ANS), with HRV serving as a key indicator of emotional states. The HER model, utilising rule extraction and amplitude level quantization, achieves an 84.3% accuracy rate in classifying 17 emotions, validated against the DREAMER dataset. The study suggests promising applications in healthcare, human-computer interaction, and psychological counselling, underscoring the potential of HRV-based emotion recognition for various domains. Future research may focus on further refining emotion classification precision and exploring additional physiological indicators for enhanced emotion recognition systems.

[10] Luis Alberto Barradas Chacon, Artem Fedoskin, Ekaterina Shcheglakova, Sutthida Neamsup and Ahmed Rashed, "Emotion Analysis using Heart Rate Data", Information Systems and Machine Learning Lab, University of Hildesheim, Germany.
Available:
https://www.researchgate.net/publication/335232482_Emotion_Analysis_Using_Heart_Rate_Data

Abstract：This paper describes the attempt to classify human emotions without including any sensors measuring biosignals, using the DEAP dataset. Basing our research on the original paper [Koelstra et al, 2012], we claim that emotions (Valence and Arousal scores) can be classified with only one pulse detecting sensor with a comparable result to the classification based on the EEG signal. Moreover, we propose a method to classify emotions avoiding any sensors by extracting the pulse of the person from the video based on the head movements. Using the Lucas-Kanade algorithm for optical flow, we extract movement signals, filter them, extract principal components, recreate HR signals and then use them in the emotion classification. First part of the project was conducted on the dataset containing 32 participants' HR values (1280 activation cases), while the second part was based on the frontal videos of the participants (874 videos). Results support the idea of one-sensor emotion classification and deny the possibility of zero-sensor classification with the proposed method.

Inference: The paper presents a comparative study between EEG and heart rate data for emotional assessment, demonstrating the effectiveness of heart rate-based models in classifying emotional states, particularly valence and arousal. It highlights the superior performance of models utilising features extracted from heart rate data, such as BPM, HRV, and Pulse, with the Feedforward Neural Network showing promising results in valence classification. However, challenges arise in the zero-sensor approach, aiming to classify emotions solely from facial video data, due to difficulties in accurately extracting heart rate signals. The study suggests future directions, including exploring colour differences in video channels, incorporating infrared cameras, and utilising CRNN for automated pulse signal

learning, to enhance emotional assessment methods. Overall, the findings suggest the potential of heart rate-based approaches as accessible alternatives to EEG-based methods in emotional assessment, while also identifying areas for further refinement and development in emotion recognition technology.

2.2 Patent search

1.Emotion recognition method and device

Inventor: Pierre-Yves Oudeyer

Current Assignee: Sony France SA

Abstract: Emotion recognition is performed by extracting a set comprising at least one feature derived from a signal, and processing the set of extracted feature(s) to detect an emotion therefrom. The voice signal is low pass filtered prior to extracting therefrom at least one feature of the set. The cut-off frequency for the low pass filtering is typically centred around 250 Hz. The features are e.g. statistical quantities extracted from sampling a signal of the intensity or pitch of the voice signal.

2.Emotion recognition system

Inventor: Shrikanth S. Narayanan

Current Assignee: University of Southern California USC

Abstract: An emotion recognition system for assessing human emotional behaviour from communication by a speaker includes a processing system configured to receive signals representative of the verbal and/or non-verbal communication. The processing system derives signal features from the received signals. The processing system is further configured to implement at least one intermediate mapping between the signal features and one or more elements of an emotional ontology in order to perform an emotion recognition decision. The emotional ontology provides a gradient representation of human emotional behaviour.

2.3 Inference Drawn

Speech emotion recognition systems typically involve the use of signal processing techniques and feature extraction methods to analyse and interpret various aspects of speech, such as pitch and intensity, in order to classify emotional states. Additionally, many of these systems incorporate machine learning algorithms to improve their accuracy and robustness in recognizing emotions from speech signals.

2.4 Comparison with the existing system

Existing System	Our System
Existing System is mostly focused on English language	Our System is focused on Hindi language
In Existing System, standard datasets like Emo-DB (Berlin Database of Emotional Speech) have been used	Our system uses Hindi Dataset which is created by considering various sources such as TV serials, YouTube, and other online platforms.
In Existing System, most of the websites/System does not have proper visualisation of the predicted emotion	Our system provides visual representations of predicted emotions through graphs.
The existing system requires high processing power to analyse large datasets and extract features	Our system also requires processing power but less than the existing system due to the use of deep learning techniques that extract features automatically

Chapter 3: Requirement Gathering for the Proposed System

In this chapter we are going to discuss the resources we have used and how we analysed what the user actually needs and what we can provide. We will also discuss the functional and non-functional requirements and finally the software and hardware used.

3.1 Introduction to requirement gathering

The Requirement Gathering is a process of requirements discovery or generating list of requirements or collecting as many requirements as possible by end users. It is also called as requirements elicitation or requirement capture.

The requirements gathering process consists of six steps :

- Identify the relevant stakeholders
- Establish project goals and objectives
- Elicit requirements from stakeholders
- Document the requirements
- Confirm the requirements
- Prioritise the requirements

3.2 Functional Requirements

3.2.1 Emotion Recognition (ER) Module:

The system must be capable of recognizing and classifying emotions from speech input and physiological data (heart rate) using deep learning techniques.

3.2.2 Audio and Heart Rate Data Integration:

The system should seamlessly integrate audio data captured from a microphone and heart rate data collected via a digital stethoscope.

3.2.3 Feature Extraction:

The system must extract relevant features from audio and heart rate signals, including wave plots, pitch diagrams, and other acoustic properties.

3.2.4 Emotion Classification:

The system should classify emotions into predefined categories, including sadness (SAD), anger (ANG), surprise (SUR), happiness (HAP), and neutral (NEU).

3.2.5 Real-time Processing:

The system should process audio and heart rate data in real-time, ensuring timely emotion recognition.

3.2.6 User Interface:

Provide a user-friendly interface for input and visualisation of emotion recognition results.

3.3 Non-Functional Requirements

3.3.1 Accuracy:

The system must achieve high accuracy in emotion recognition, with a focus on minimising false positives and false negatives.

3.3.2 Speed:

Real-time processing of data should be efficient, with low latency for emotion classification.

3.3.3 Robustness:

The system should be resilient to noise and variations in input data, ensuring consistent performance.

3.3.4 Security:

Ensure the privacy and security of user data, especially heart rate information.

3.3.5 Scalability:

The system should be scalable to handle a growing dataset and adapt to different environments

3.4.Hardware, Software , Technology and tools utilised

3.4.1 Hardware:

1. Raspberry Pi 4: Main computing platform for processing audio data and running deep learning models.
2. SD Card: Storage medium for the Raspberry Pi's operating system and project files.
3. Microphones: Capture audio data from the stethoscope for emotion recognition.
4. Stethoscope: Collects physiological audio signals for analysis.
5. Bread Board: Used for connecting and interfacing various components.
6. Connecting Wires: Necessary for establishing connections between components.
7. Charger: Provides power to the Raspberry Pi and other devices in the setup.

3.4.2 Software:

1. Raspberry Pi OS: Install a compatible operating system for Raspberry Pi, such as Raspberry Pi OS (formerly Raspbian), to serve as the project's main computing platform.
2. Python: Python is commonly used for deep learning and data processing tasks.
3. Deep Learning Frameworks: Deep learning frameworks like TensorFlow, Keras, or PyTorch for building and training your emotion recognition models.
4. Audio Processing Libraries: Utilise libraries like NumPy, SciPy, and librosa for audio signal processing and feature extraction.
5. Audio I/O Libraries: libraries like PyAudio or ALSA for handling audio input/output on the Raspberry Pi.
6. Arduino Shell:Arduino shell for coding in Raspberry.

3.5 Constraints

3.5.1 Data Availability:

The system's performance may depend on the availability of high-quality training data for both speech and heart rate signals.

3.5.2 Processing Power:

The system may require significant computational resources, especially when performing real-time analysis.

3.5.3 Environment:

The system’s prediction would depend on the environment i.e the room in which the recording is being done must be quiet.

Chapter 4: Proposed Design

In this chapter, we outline our proposed design for an emotion recognition system that integrates both speech and heart rate signals. Emotion recognition, essential for effective communication and understanding, poses significant challenges due to its subjective nature. Our approach combines advanced signal processing techniques with machine learning algorithms to extract key features from the input signals. Our modular framework ensures scalability and flexibility, allowing for efficient processing and classification of emotions. Through this design, we aim to develop a system that not only accurately detects emotions but also considers practical considerations such as usability and ethical implications, paving the way for enhanced human-computer interaction and mental health monitoring.

4.1 Block diagram of the system

Block Diagram

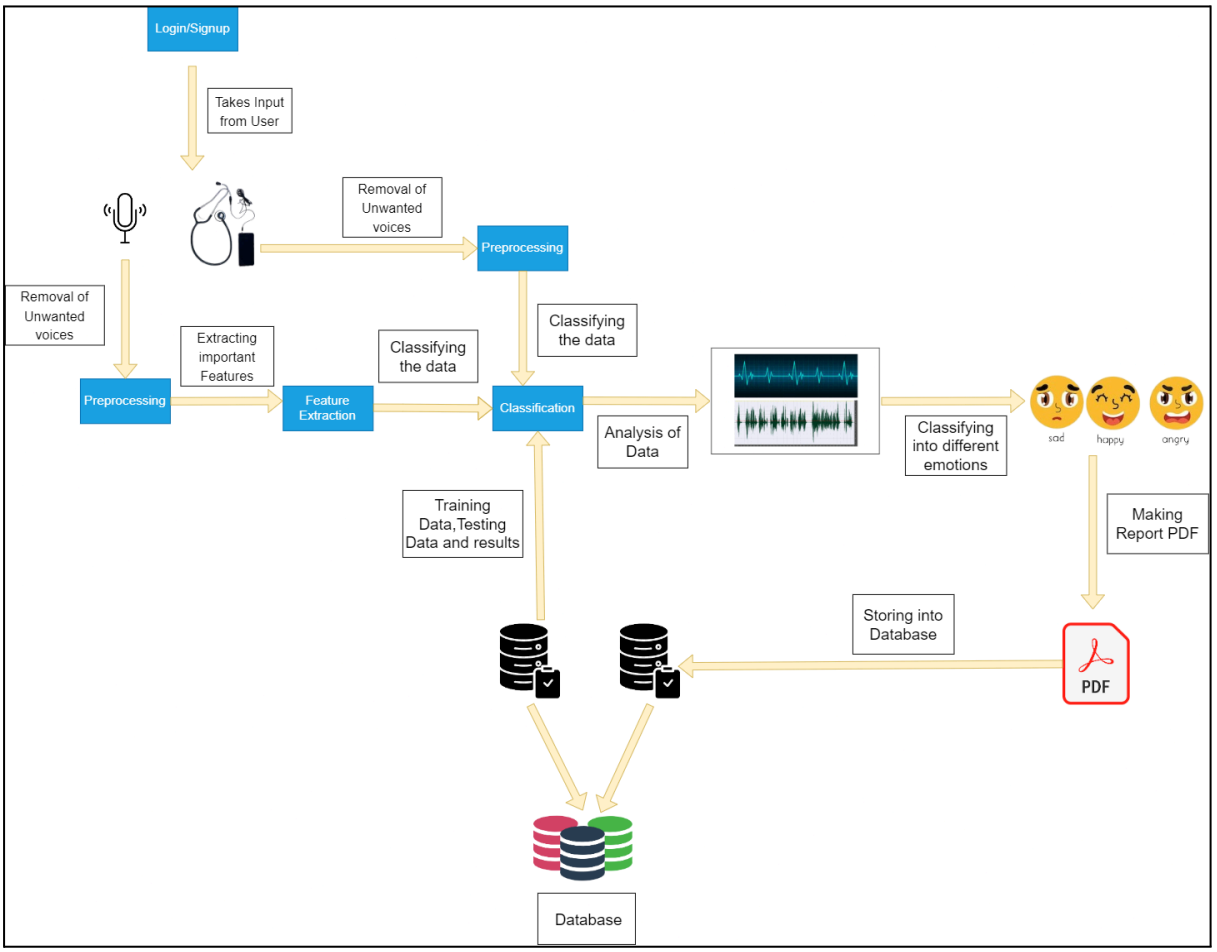

Fig 9. Block Diagram of the Audio emotion recognition system

Fig9 represents a Block diagram for our project. Firstly we are taking two different simultaneous audio signals (Speech audio and Heart beat audio) as an input. Preprocessing is a feature enhancement operation which can help us to detect better emotions. In Feature

extraction we extract the features from that audio, like what component, what factor of audio will affect the emotion. And the predicted emotions will be detected using the classifier. Finally, a report will be generated in the form of a pdf which is stored in a database for future analysis.

4.2 Modular design of the system

Modular Diagram

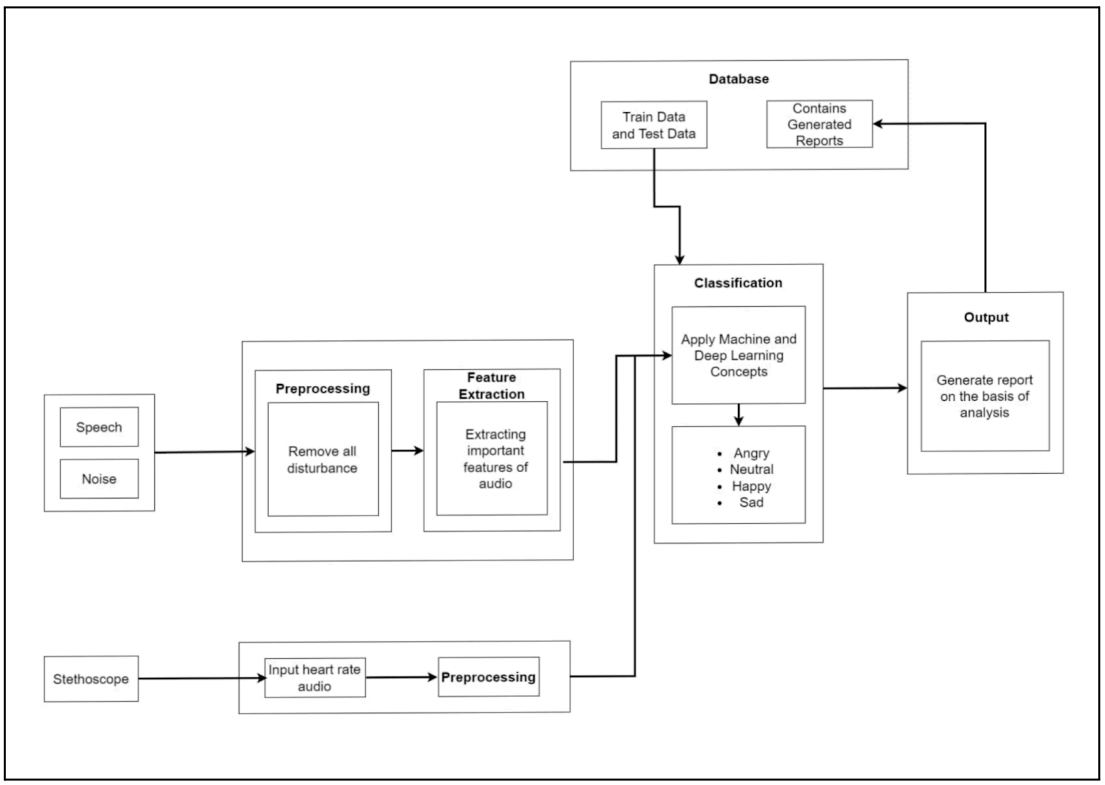

Fig 10. Modular Diagram of Audio emotion recognition system

Fig10 represents the modular diagram.The sample audio is provided as input. Here we are also taking heart rate as an input to increase the accuracy of the system. First the input signals we go through preprocessing where the input signals are enhanced by pre-emphasis filter, framing, windowing and noise removal. Then we extract features of the pure signals of Speech audio using the MFCC model and used the Heartpy module for heart audio. Then it will be classified into 5 different emotions like SAD, ANG, SUR, HAP, NEU with the help of MLP model and KNN classifier. The classified emotions will be displayed as an output to the user.

4.3 Detailed Design

Hardware Representation

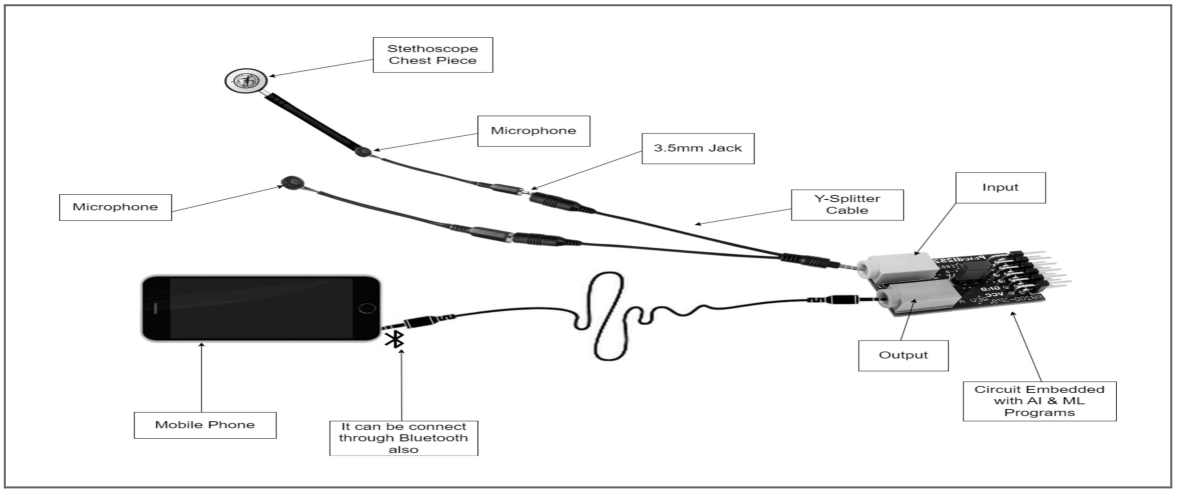

Fig 11. Hardware Representation

Fig 12. Hardware Implementation

Figure shows Raspberry Pi 4B: The central computing unit of the setup, equipped with various ports and functionalities. Power Source: Provides electrical power to the Raspberry Pi 4B via a USB Type-C connection. Display Connection: Utilises an HDMI cable connected to the micro HDMI port on the Raspberry Pi 4B to interface with a display for visual output. Internet Connectivity: Establishes network connectivity through the LAN port, enabling access to the internet and local network resources.

USB Ports:

- USB-A Ports (x4): Provides connectivity for external peripherals and devices.
- Keyboard and Mouse: Two USB-A ports are allocated for connecting a keyboard and a mouse, facilitating user input and interaction with the system.
- Microphone: One USB-A port is dedicated to connecting a standard microphone for recording speech audio.
- Digital Stethoscope: One USB-A port is designated for connecting a digital stethoscope, enabling the recording of heart rate and related physiological data.

4.4 Project Scheduling & Tracking using Timeline / Gantt Chart

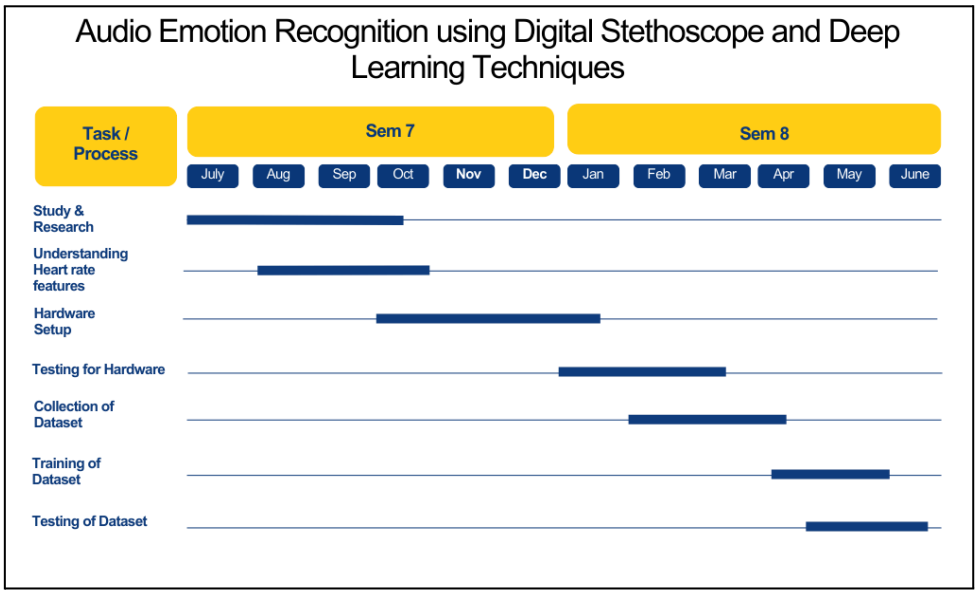

Fig 13. Gantt Chart

The project consists of seven tasks/processes that span from July to June. The first task is to study and research the existing literature and methods on audio emotion recognition and heart rate features. The second task is to understand the heart rate features that are relevant for emotion recognition and how to extract them from audio signals. The third task is to set up the hardware, such as the digital stethoscope, the microphone, and the computer, that will be used for data collection and processing. The fourth task is to collect the dataset of audio recordings from different subjects with different emotions and label them accordingly. The fifth task is to train a deep learning model on the dataset using various techniques and parameters. The sixth task is to test the model on new data and evaluate its performance and accuracy. The final task is to complete the project report and presentation for the eighth semester.

Chapter 5: Implementation of the Proposed System

5.1. Methodology employed for development

Data collection:

Data is gathered for the proposed system in two different ways. One method is by considering the existing (open source) audio signals of both speech and heart rate. The alternate method is to take real time audio of speech and heart rate simultaneously, which is captured using microphone and stethoscope respectively as vibrations hitting the chestpiece are funnelled into a narrow tube, instead of being allowed to travel outward at will, most of them reach the device making it amplified both at the same time. These input audios are then applied for cleaning, feature extraction and classification process.

Data Preprocessing:

Once the raw data is gathered, the next step is to remove the unwanted noises for feature enhancement operation. The main purpose of preprocessing is to provide a set of clean and levelled audio assets from the noisy audio. It processes the raw audio signals to a common standard format. It also removes the unwanted sections, such as chatter between talks, coughs, sneezes and any aberrant peaks. This will help us to detect emotions better. In preprocessing step, raw data is enhanced by different preprocessing techniques as mentioned below:

Pre-emphasis filter:

Preemphasis increases the magnitude of energy in the higher frequency. When we look at the frequency domain of the audio signal for the voiced segments like vowels, it is observed that the energy at a higher frequency is much lesser than the energy in lower frequencies. It basically increases the amplitude of high frequency bands and decreases the amplitudes of lower bands.

Framing :

The next step is known as signal framing. It is also alluded to as speech segmentation and is the way toward apportioning constant speech signals into fixed length sections to surpass a few ER difficulties. Emotions often tend to vary during a speech as a result of the signals being non-stationary. Despite this fact, the speech remains invariant even though it is for a very short period, such as 20 to 30 milliseconds. Speech signal, when framed, helps to estimate the semi-fixed and local features

Windowing:

During Fast Fourier Transform (FFT) of information, leakages occur due to discontinuities at the edge of the signals, henceforth reduced by the windowing function. Windowing is a classical method in signal processing and it refers to splitting the input signal into temporal segments. The borders of segments are then visible as discontinuities, which are incongruent with the real-world signal. To reduce the impact of segmenting on the statistical properties of the signal, we apply windowing to the temporal segments. Windowing functions are smooth functions and go to zero at the borders. By multiplying the input signal with a window function, the windowing function also goes to zero at the border such that the discontinuity at

the border becomes invisible. Windowing does thus change the signal, but the change is designed such that its effect on signal statistics is minimised.

Noise removal:

Noise reduction techniques exist for audio and images. All signal processing devices, both analog and digital, have traits that make them susceptible to noise. Noise reduction helps in reducing noise in time-domain signals like speech, bioacoustics, and physiological signals.

Feature Extraction:

The speech signal contains a large amount of hidden information which reflects the emotional characteristics. Feature extraction is considered as an important phase of the system as extraction of the relevant and significant features heavily impact on the final decisions of the classifier model. The extracted features are pitch, amplitude of the signals and so on. Also, balancing time-frequency ranges by converting digital and analog signals using Mel-Frequency Cepstral Coefficients (MFCC), which is good in error reduction and able to produce a robust feature when the signal is affected by noise. This step will help us to make a robust classifier for predicting the emotion of the speaker.

Classifier Model and Report Generator:

The classifier model aims to predict emotional labels using the most relevant set of extracted features from audio and heart rate signals. The processed dataset is further divided into training and development datasets in the ratio of 80:20. To train the model, various deep learning algorithms are used to classify into different emotions like SAD(Sadness), ANG(Anger), HAP(Happy), Surprise (SUR), NEU(Neutral) of an individual. Unknown samples are used to evaluate the model and the resultant data is stored in the database for future analysis. Eventually, reports are generated in pdf format, composed of resultant graphs in the form of pitch diagrams, wave plots and emotion tags of an individual user.

5.2 Algorithms and flowcharts for the respective modules developed

Algorithms utilised in the existing systems

1. Convolutional Neural Networks (CNNs): CNNs are employed for audio and heart rate feature extraction.
2. Recurrent Neural Networks (RNNs): RNNs may be used for sequential data analysis.
3. Multi Layer Perceptron (MLP)

Flowchart for the proposed system

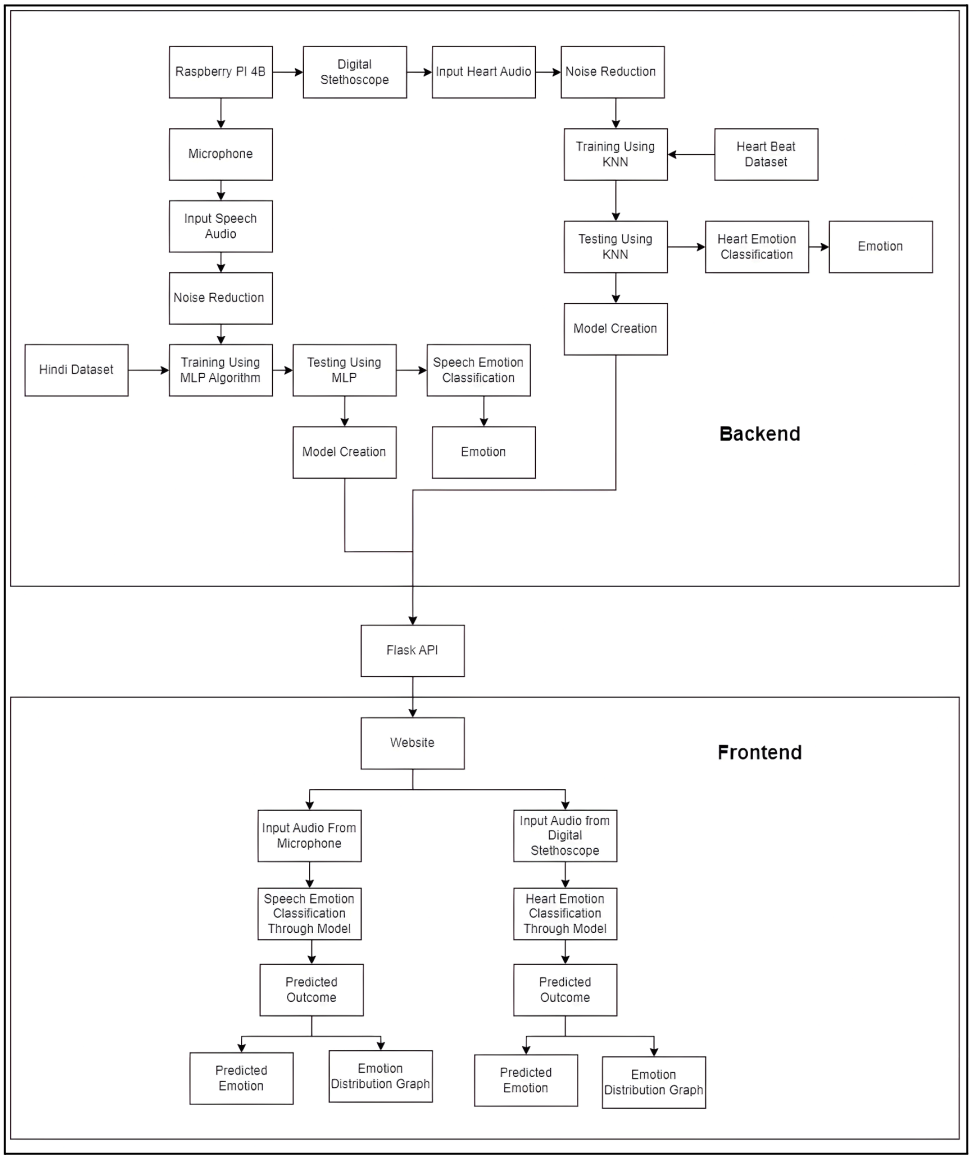

Fig 14. Flow chart for the proposed system

The flowchart depicts a system for speech and heart emotion classification, consisting of a backend and a frontend. In the backend, a Raspberry Pi 4B captures audio input from either a microphone or a digital stethoscope. This input audio undergoes noise reduction processing. The noise-reduced audio is then sent through two separate pipelines for speech emotion classification and heart emotion classification.

For speech emotion classification, a Hindi dataset is used for training and testing an MLP (Multilayer Perceptron) algorithm. This process results in the creation of a speech emotion classification model. On the other hand, for heart emotion classification, a heart beat dataset is employed for training and testing a KNN (K-Nearest Neighbors) algorithm, leading to the development of a heart emotion classification model. Both these models are integrated into the backend system.

The backend component exposes a Flask API, which serves as an interface for communication with the frontend. The frontend is a website that accepts audio input from

either a microphone or a digital stethoscope. This audio input is processed by the respective models (speech emotion classification or heart emotion classification) in the backend.
The frontend then displays the predicted emotion outcomes, including the predicted emotion labels and emotion distribution graphs. These visualisations provide insights into the emotional state detected from the speech or heart audio input.

5.3 Datasets source and utilisation

In the development of the proposed emotion recognition system, careful attention was given to the acquisition and utilisation of datasets, ensuring both diversity and relevance to the target application domain. For the speech audio component, a bespoke dataset was meticulously curated from a variety of Hindi-language sources, including YouTube videos and television shows. This approach ensured a rich and diverse collection of audio samples encompassing various contexts and emotional expressions encountered in real-life interactions. Each audio sample was meticulously labelled with the corresponding emotional state, leveraging domain expertise and manual annotation to ensure accuracy and consistency.
In parallel, physiological signals were incorporated into the system using the Swell dataset, a publicly available resource focusing on stress detection. While the Swell dataset provided valuable physiological data, it required adaptation to suit the specific requirements of emotion recognition. Through careful preprocessing and thresholding techniques, the stress labels were transformed into categorical emotion labels, aligning them with the emotional categories targeted by the system. This adaptation process involved meticulous experimentation and validation to ensure the fidelity and relevance of the physiological data to the overall emotion recognition task.
By leveraging both bespoke audio datasets and adapted physiological data, the proposed system benefits from a rich and multidimensional dataset that captures a wide range of emotional expressions in the Hindi language. This dataset not only facilitates the training and evaluation of the emotion recognition models but also ensures that the system is tailored to the linguistic and cultural nuances of its target audience. Moreover, the meticulous labeling and preprocessing efforts undertaken during dataset curation enhance the reliability and effectiveness of the system, positioning it as a robust solution for emotion recognition in real-world applications.

Chapter 6: Testing of the Proposed System

6.1 . Introduction to testing

Software testing is the process of evaluating a software application or system to identify any defects, errors or malfunctions, and to ensure that it meets the specified requirements and performs as intended. Testing is an essential part of the software development lifecycle, as it helps to identify any problems before the software is released to users, reducing the risk of user dissatisfaction and costly recalls.

The testing process typically involves creating test cases or scenarios that simulate different user interactions with the software. These test cases are then executed to identify any defects or issues that prevent the software from working as intended. The test results are recorded and analysed, and any defects found are reported to the development team to be fixed.

6.2. Types of tests Considered

1. Black box Testing:

Black box testing is a testing method where the software is tested without knowledge of its internal workings or code. The tester focuses on the inputs and outputs of the system, without knowing how the system processes those inputs. The objective of black box testing is to identify defects in the software's functionality and to ensure that the software meets the specified requirements.

2. Cross-validation:

This involves partitioning the available data into several subsets, then training the model on one subset and testing it on another. This process is repeated for each subset, and the results are averaged to give an overall measure of performance.

3. Holdout testing:

This involves splitting the data into a training set and a testing set. The model is trained on the training set, and then tested on the testing set to evaluate its performance.

6.3 Various test case scenarios considered

Sr. no.	Description	Expected Result	Actual Result	Pass/Fail
1	Testing the noise reduction functionality of the system using spectral	The system should effectively reduce	The system successfully reduced the background	Pass

	gating and low pass filter.	background noise in the speech signal.	noise, resulting in a clearer and more easily recognizable speech signal.	
2	The Multilayer Perceptron (MLP) is used for training and testing the dataset to predict emotions as output	The MLP should accurately predict the emotions in the given hindi dataset.	The MLP accurately predicted the emotions with an accuracy of 60% for hindi dataset	Pass
3	Testing the API creation using Flask of the MLP model for speech emotion recognition system.	The API should successfully receive the input speech signal and return the predicted emotion label as output	The API successfully received the input speech signal and returned the predicted emotion label as output.	Pass
4	Verify that the user interface is intuitive and user-friendly.	Users can navigate through the system easily and find the necessary functionalities without any confusion.	Users can navigate through the system easily and find the necessary functionalities without any confusion.	Pass
5	Uploading the file involves selecting the audio file and submitting it to the system for emotion recognition.	The selected audio file should be successfully uploaded to the system	The selected audio file is successfully uploaded to the system for emotion recognition.	Pass
6	Testing the display functionality of the uploaded audio file, where the uploaded audio file is played in the display.	The uploaded audio file should be displayed on the screen, and on clicking the play button, it should start playing the	The uploaded audio file should be displayed on the screen, and on clicking the play button, it should start playing the	Pass

		audio file.	audio file.	
7	Testing the outcome of speech emotion recognition system by displaying the emotions in a bar graph along with their respective percentages for an uploaded audio file.	The bar graph should accurately display the emotions detected in the audio file and their respective percentages.	The bar graph displays the emotions and their percentages accurately, matching the emotions expressed in the uploaded audio file.	Pass
8	One predicted emotion is displayed in text form below the graph in speech emotion recognition.	The text displayed below the graph should match with the corresponding emotion predicted in the graph.	During testing, the text displayed below the graph matched with the corresponding emotion predicted in the graph.	Pass

6.4. Inference drawn from the test cases

From the above test cases, we infer that our system has certain requirements that must be met to ensure that it functions as intended. These requirements include:

- Our model should successfully reduce the background noise from the input audios.
- The system should take audio files as an input.
- System should be able to display the output once the predict button is clicked.
- Our model should display the percentage graph as an output.
- System should accurately predict the emotion.

The above tests ensure that the project functions as expected, meets the user requirements, and is secure and stable. The successful completion of all test cases indicates that the project is reliable and can be used effectively for emotion recognition through speech.

Chapter 7: Results and Discussion

7.1. Screenshots of User Interface (UI) for the respective module

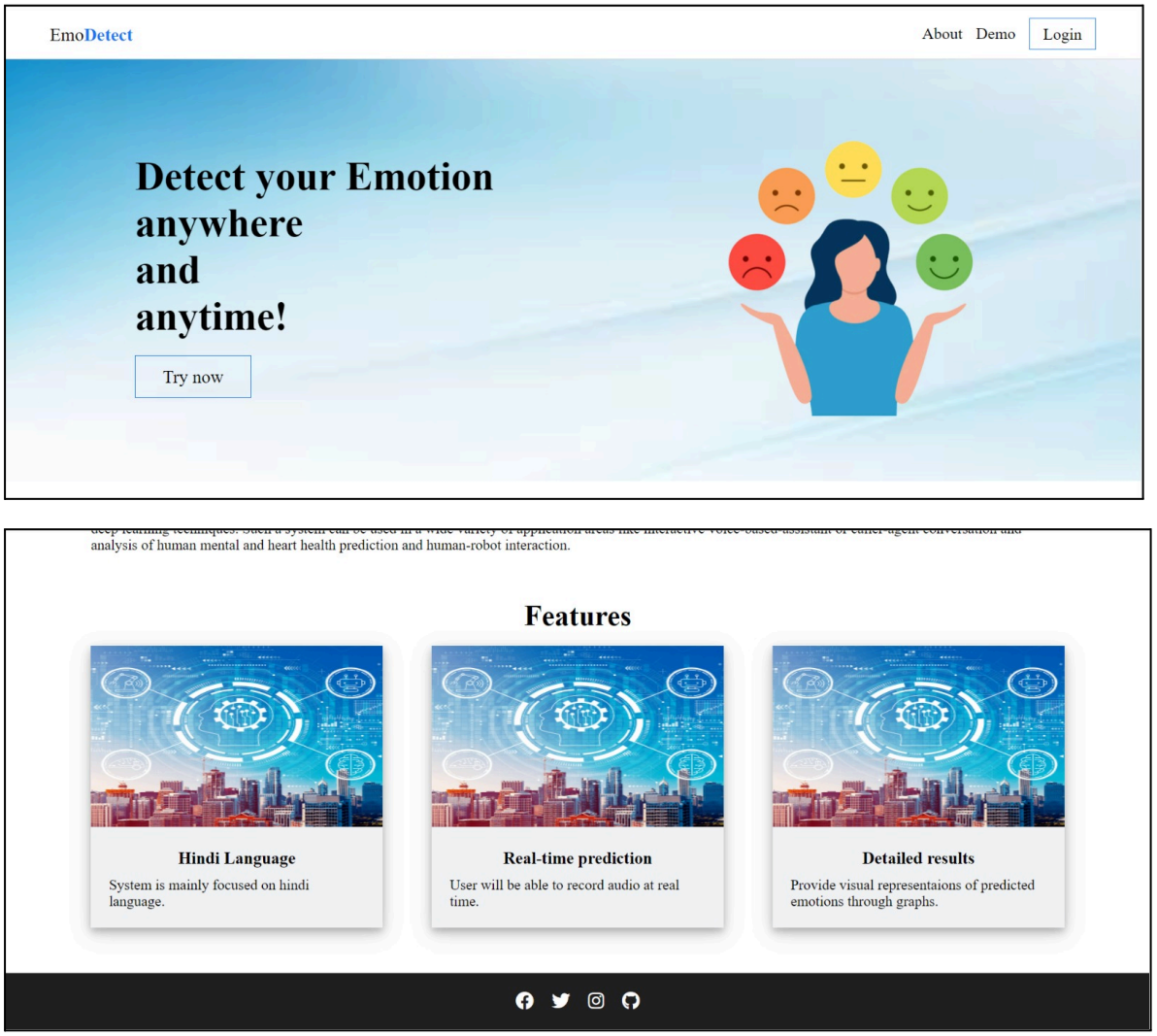

Fig. 15 Home Page UI

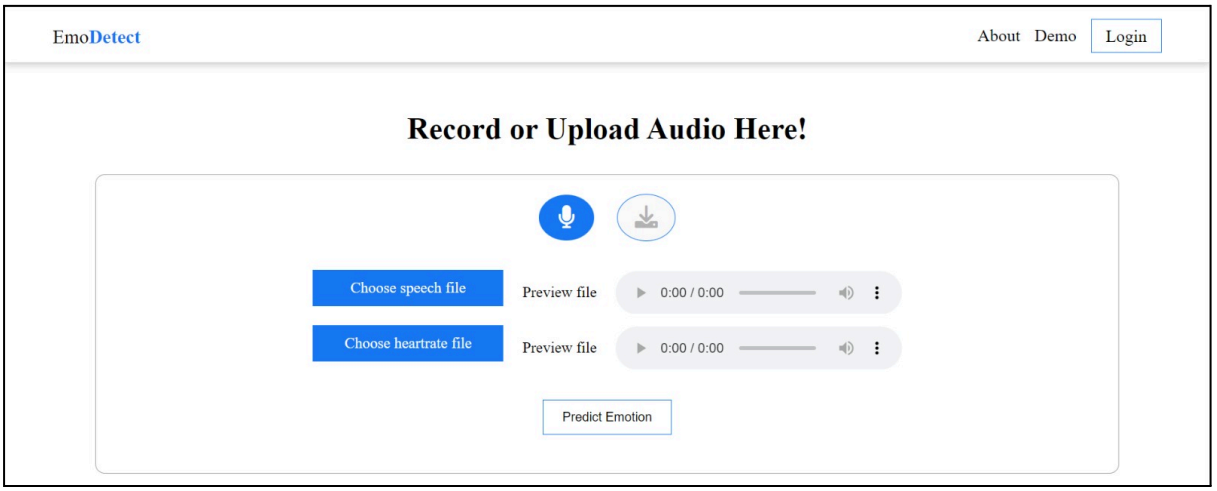

Fig. 16 Prediction Page

7.2. Performance Evaluation measures

Speech Emotion Confusion Matrix

Multilayer Perceptron (MLP):

	precision	recall	f1-score	support
angry	0.84	0.69	0.76	39
happy	0.86	0.92	0.89	59
neutral	0.85	0.92	0.88	65
sad	0.88	0.84	0.86	55
accuracy			0.86	218
macro avg	0.86	0.84	0.85	218
weighted avg	0.86	0.86	0.86	218

Fig. 17 Classification report of MLP model

Heart Emotion Confusion Matrix:

K-nearest neighbour(KNN)

	precision	recall	f1-score	support
angry	0.98	0.98	0.98	7875
happy	0.99	0.99	0.99	25240
neutral	0.96	0.96	0.96	7918
accuracy			0.98	41033
macro avg	0.98	0.98	0.98	41033
weighted avg	0.98	0.98	0.98	41033

Fig. 18 Classification report of KNN model

7.3. Input Parameters / Features considered

For Heart Rate features considered:

1. `hr`: Heart rate.
2. `sample_rate`: Sampling rate of the heart rate data.
3. `meanRR`: Mean of RR intervals (beat-to-beat intervals).
4. `medianRR`: Median of RR intervals.
5. `stdev_rr`: Standard deviation of RR intervals.

6. `rmssd`: Root means square of successive differences between RR intervals.
7. `nn25_count`: Number of pairs of successive RR intervals differing by more than 25 milliseconds.
8. `pnn25`: Percentage of successive RR interval pairs differing by more than 25 milliseconds.
9. `nn50_count`: Number of pairs of successive RR intervals differing by more than 50 milliseconds.
10. `pnn50`: Percentage of successive RR interval pairs differing by more than 50 Milliseconds.

For Speech Feature Considered:

Mel-frequency cepstral coefficients (MFCC):

Mel-frequency spectrum, Discrete cosine transform (DCT) coefficients, Logarithmic compression, Delta coefficients (first-order derivatives), Delta-delta coefficients (second-order derivatives), Energy, Pitch, Formants (for speech-related tasks), Zero-crossing rate, Spectral centroid

Chroma features:

Chroma energy, Chroma vector, Chroma deviation, Chroma-based pitch class profiles, Chroma-based autocorrelation, Chroma-based cross-correlation, Chroma-based cross similarity, Chroma-based onset detection, Chroma-based modulation spectrum, Harmonic ratio

7.4. Graphical and statistical output

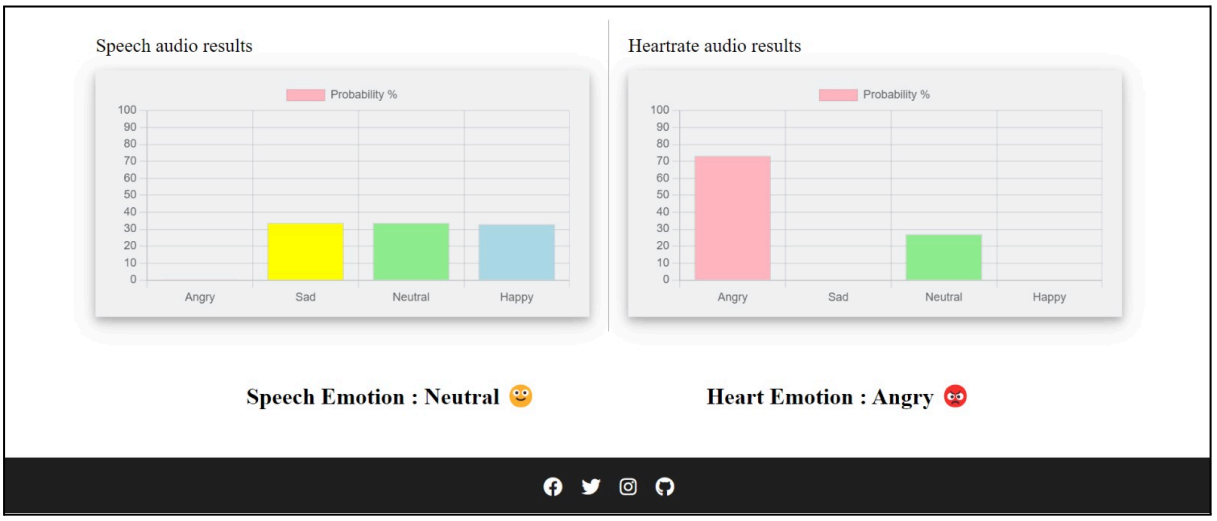

Fig. 19 Graphical output on UI

7.5. Comparison of results with existing systems

1. **Modality Integration Advantages:** The proposed system offers a distinct advantage over existing emotion recognition systems by integrating both physiological signals and speech analysis. While many existing systems rely solely on speech or facial expressions for emotion recognition, the inclusion of physiological data, such as heart rate, provides additional insights into the speaker's emotional state. By considering both speech patterns and physiological responses, the proposed system offers a more comprehensive understanding of emotions, potentially leading to improved accuracy and objectivity compared to traditional single-modality approaches. This unique integration sets the proposed system apart from existing systems and represents a significant advancement in the field of emotion recognition technology.

2. **Language and Cultural Specificity:** Another key point of comparison is the proposed system's specialisation in emotion recognition for the Hindi language and its curated dataset from Hindi entertainment channels. Existing emotion recognition systems often target English or other widely spoken languages, neglecting linguistic and cultural diversity. In contrast, the proposed system's focus on Hindi addresses this gap in existing literature, ensuring better alignment with the linguistic and cultural nuances of Hindi speakers. This language-specific approach enhances the relevance and applicability of the proposed system, particularly for Hindi-speaking populations, compared to existing systems that may lack such specificity.

7.6. Inference drawn

The website showcases an innovative approach to emotion analysis by seamlessly integrating simultaneous capture of both speech and heart sounds, providing users with a comprehensive understanding of their emotional state. Through a user-friendly interface featuring a predict button and intuitive bar graph visualisation, users can easily initiate the analysis process and interpret the results without technical complexity. Real-time or near real-time feedback offers timely insights into emotional states, while probability information for each emotion category enhances the reliability of the analysis. With reported accuracy levels of 80-85% for speech emotion analysis and 90-95% for heart emotion analysis, users can trust the robustness of the underlying classification models. The website's potential applications span mental health monitoring, personalised user experiences, and healthcare assessments, leveraging the versatility of Raspberry Pi integration for deployment in diverse environments.

Chapter 8: Conclusion

8.1 Limitations

Limitation 1: Electric Interference: The presence of electric interference in physiological signal acquisition could lead to crackling or artefacts, affecting the accuracy of emotion recognition and potentially introducing erroneous interpretations.

Limitation 2: Limited Emotions: Restricting the recognition to a smaller set of basic emotions (e.g., sadness, anger, happiness, neutral) may overlook the nuances of human emotional expression, limiting the system's applicability in capturing a broader range of emotional states and potentially hindering its effectiveness in certain contexts.

8.2 Conclusion

In this project, we proposed a novel approach to Emotion Recognition (ER) using a combination of speech audio and heart rate signals. By integrating physiological signals like heart rate, we aimed to improve the accuracy and objectivity of emotion detection in digital communication. Through rigorous data preprocessing, feature extraction, and the implementation of deep learning models, we have demonstrated the potential for enhancing emotion recognition capabilities.

The integration of heart rate data with audio analysis offers a more comprehensive understanding of the speaker's emotional state. Our system's performance, as evaluated through various metrics, indicates promising results in accurately identifying basic emotions such as sadness, anger, surprise, happiness, and neutrality.

The proposed system's potential applications extend beyond traditional human-machine interaction, encompassing areas such as mental health prediction, interactive voice assistants, and human-robot interaction. By considering both speech and physiological signals, we contribute to a more holistic and reliable approach to emotion recognition.

However, further research and development are required to refine the system, optimise computational efficiency, and extend its capabilities to recognize a wider range of emotions. Despite these challenges, the successful integration of audio and physiological signals offers a promising direction for advancing emotion recognition technology and improving human-computer interaction in the digital age.

8.3 Future Scope

Electric Interference Mitigation: Implement advanced signal processing techniques to mitigate the effects of electric interference on physiological signal acquisition. This may involve employing more advanced noise reduction algorithms, improving sensor design, or utilising alternative signal acquisition methods to ensure cleaner data.

Expansion of Emotional Categories: Extend the emotion recognition system to encompass a wider spectrum of emotions beyond the basic categories. This could involve incorporating more nuanced emotional states, such as disgust, fear, surprise, and contempt, through deeper analysis of both speech and physiological signals. Additionally, leveraging advanced machine learning models capable of capturing subtle emotional nuances may enhance the system's ability to recognize a broader range of emotions accurately.

Enhanced Data Collection and Annotation: Collect larger and more diverse datasets with comprehensive annotations to capture the variability of emotional expressions across different demographics, cultures, and contexts. Employing crowdsourcing techniques and leveraging naturalistic data collection methods can facilitate the acquisition of high-quality annotated datasets for training and evaluation purposes.

References

[1] A. Zhu and Q. Luo, “Study on Speech Emotion Recognition System in E-Learning”, presented at International Conference on Human-Computer Interaction, vol 4552, Springer, Berlin, Heidelberg, 2007, https://doi.org/10.1007/978-3-540-73110-8_59. [Online]. Available: https://link.springer.com/content/pdf/10.1007/978-3-540-73110-8_59.pdf?pdf=inline%20link

[2] M. Suzukia and J. Qib, “Improvement of multilingual emotion recognition method based on normalised acoustic features using CRNN”, Procedia Computer Science, vol. 207, pp. 684-691, 2022, https://doi.org/10.1016/j.procs.2022.09.123. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922010043

[3] A.V.Tsaregorodtseva et al.,”The architecture of the emotion recognition program by speech segments”, Procedia Computer Science, vol. 213, pp. 338-345, 2022, https://doi.org/10.1016/j.procs.2022.09.123. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050922017677

[4] F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," in IEEE Access, vol. 10, pp. 36018-36027, 2022, doi: 10.1109/ACCESS.2022.3163856. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9745599

[5] Yoon, Seunghyun & Byun, Seokhyun & Jung, Kyomin. (2018). “Multimodal Speech Emotion Recognition Using Audio and Text”. [Online]. Available: https://www.researchgate.net/publication/328213059_Multimodal_Speech_Emotion_Recognition_Using_Audio_and_Text

[6] A. A. Varghese, J. P. Cherian and J. J. Kizhakkethottam, "Overview on emotion recognition system," 2015 International Conference on Soft-Computing and Networks Security (ICSNS), 2015, pp. 1-5, doi: 10.1109/ICSNS.2015.7292443. [Online]. Available: https://ieeexplore.ieee.org/document/7292443

[7] Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh and Promod Yenigalla,” Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions” .(2020) . [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1906/1906.05681.pdf

[8] A. Africa, A. Tabalan and M. Tan, “Speech Emotion Recognition using Support Vector Machines”, International Journal of Emerging Trends in Engineering Research, Volume 8. No. 4, 2020, https://doi.org/10.30534/ijeter/2020/43842020. [Online]. Available: http://www.warse.org/IJETER/static/pdf/file/ijeter43842020.pdf

[9] Babak Basharirad and Mohammadreza Moradhaseli , "Speech emotion recognition methods: A literature review", AIP Conference Proceedings 1891, 020105 (2017) https://doi.org/10.1063/1.5005438. [Online]. Available: https://aip.scitation.org/doi/pdf/10.1063/1.5005438

[10] Luis Alberto Barradas Chacon, Artem Fedoskin, Ekaterina Shcheglakova, Sutthida Neamsup and Ahmed Rashed, "Emotion Analysis using Heart Rate Data", Information Systems and Machine Learning Lab, University of Hildesheim, Germany.
Available:
https://www.researchgate.net/publication/335232482_Emotion_Analysis_Using_Heart_Rate_Data

[11] Cheng Xiefeng, Yue Wang, Shicheng Dai, Pengjun Zhao & Qifa Liu Scientific Reports volume 9, Article number: 6486 (2019) "Heart sound signals can be used for emotion recognition"
Available: https://www.nature.com/articles/s41598-019-42826-2

[12] S. Leng, R. Tan, K. Chai et al., "The electronic stethoscope", BioMed Eng OnLine 14, 66, 2015, doi: 10.1186/s12938-015-0056-y. [Online].
Available: https://pubmed.ncbi.nlm.nih.gov/26159433/

[13] Wang L, Hao J, Zhou TH. ECG Multi-Emotion Recognition Based on Heart Rate Variability Signal Features Mining. Sensors (Basel). 2023 Oct 22;23(20):8636. doi: 10.3390/s23208636. PMID: 37896729; PMCID: PMC10610830.
Available:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10610830/#B26-sensors-23-08636

Paper

Paper title : Speech Emotion Recognition

Paper id : 10

Conference : 3rd International Conference of UG Students ICUS 2024

Certifications

3rd International Conference of UG Students

ICUS 2024

Organized By

Soft Computing Research Society

March 01, 2024

Certificate of Presentation

This is to certify that **Prabha Pamula** has presented the paper titled **Speech Emotion Recognition** authored by **Mannat Doultani, Ritika Hotwani, Vivek Balani, Yash Pahlani, Prabha Pamula** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma

(Organizing Chair, ICUS 2024)

https://www.scrs.in/conference/icus2024/

3rd International Conference of UG Students

ICUS 2024

Organized By

Soft Computing Research Society

March 01, 2024

Certificate of Presentation

This is to certify that **Yash Pahlani** has presented the paper titled **Speech Emotion Recognition** authored by **Mannat Doultani, Ritika Hotwani, Vivek Balani, Yash Pahlani, Prabha Pamula** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma

(Organizing Chair, ICUS 2024)

https://www.scrs.in/conference/icus2024/

3rd International Conference of UG Students

ICUS 2024

Organized By

Soft Computing Research Society

March 01, 2024

Certificate of Presentation

This is to certify that **Vivek Balani** has presented the paper titled **Speech Emotion Recognition** authored by **Mannat Doultani, Ritika Hotwani, Vivek Balani, Yash Pahlani, Prabha Pamula** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma

(Organizing Chair, ICUS 2024)

https://www.scrs.in/conference/icus2024/

3rd International Conference of UG Students

ICUS 2024

Organized By

Soft Computing Research Society

March 01, 2024

Certificate of Presentation

This is to certify that **Ritika Hotwani** has presented the paper titled **Speech Emotion Recognition** authored by **Mannat Doultani, Ritika Hotwani, Vivek Balani, Yash Pahlani, Prabha Pamula** in the **3rd International Conference of UG Students** held during March 01, 2024.

Dr. Harish Sharma

(Organizing Chair, ICUS 2024)

https://www.scrs.in/conference/icus2024/

Plagiarism Report

Speech Emotion Recognition

ORIGINALITY REPORT

30%	20%	23%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.warse.org Internet Source	4%
2	Lecture Notes in Computer Science, 2007. Publication	4%
3	www.researchgate.net Internet Source	3%
4	Submitted to University of Mumbai Student Paper	3%
5	Motoyuki Suzuki, Jinhai Qi. "Improvement of multilingual emotion recognition method based on normalized acoustic features using CRNN", Procedia Computer Science, 2022 Publication	2%
6	Samuel Kakuba, Alwin Poulose, Dong Seog Han. "Attention-Based Multi-Learning Approach for Speech Emotion Recognition with Dilated Convolution", IEEE Access, 2022 Publication	2%
7	Submitted to RMIT University Student Paper	2%

Award certificate for project competition

Patent Publication

Office of the Controller General of Patents, Designs & Trade Marks
Department of Industrial Policy & Promotion,
Ministry of Commerce & Industry,
Government of India

INTELLECTUAL PROPERTY INDIA
PATENTS | DESIGNS | TRADE MARKS
GEOGRAPHICAL INDICATIONS

Application Details	
APPLICATION NUMBER	202321003900
APPLICATION TYPE	ORDINARY APPLICATION
DATE OF FILING	20/01/2023
APPLICANT NAME	1 . Mrs. Mannat Amit Doultani 2 . Mrs. Priya R L 3 . Mr. Vivek Balani 4 . Mr. Yash Pahlani 5 . Ms. Ritika Hotwani 6 . Ms. Prabha Sudarshan Pamula
TITLE OF INVENTION	AUDIO EMOTION RECOGNITION USING DIGITAL STETHOSCOPE AND DEEP LEARNING TECHNIQUES
FIELD OF INVENTION	BIO-MEDICAL ENGINEERING
E-MAIL (As Per Record)	priya.rl@ves.ac.in
ADDITIONAL-EMAIL (As Per Record)	priya.rl@ves.ac.in
E-MAIL (UPDATED Online)	
PRIORITY DATE	
REQUEST FOR EXAMINATION DATE	20/01/2023
PUBLICATION DATE (U/S 11A)	03/02/2023

Vivekanand Education Society's Institute Of Technology, Mumbai
An Autonomous Institute affiliated to University of Mumbai, approved by AICTE
In association with VESIT - IQAC & VESIT - IIC
Department Of Computer Engineering

CERTIFICATE OF EXCELLENCE

This to certify that Vivek Balani
of D17B has won Consolation prize at Department
Technology Day, **Pradarshini '24**
conducted on 22nd-23rd March 2024.

DR. MRS. NUPUR GIRI, HOD, COMPUTER DEPARTMENT
DR. MRS. GRESHA BHATIA, DHOD, COMPUTER DEPARTMENT
PROJECT COORDINATOR, COMPUTER DEPARTMENT

Vivekanand Education Society's Institute Of Technology, Mumbai
An Autonomous Institute affiliated to University of Mumbai, approved by AICTE
In association with VESIT - IQAC & VESIT - IIC
Department Of Computer Engineering

CERTIFICATE OF EXCELLENCE

This to certify that Ritika Hotwani
of D17B has won Consolation prize at Department
Technology Day, **Pradarshini '24**
conducted on 22nd-23rd March 2024.

DR. MRS. NUPUR GIRI, HOD, COMPUTER DEPARTMENT
DR. MRS. GRESHA BHATIA, DHOD, COMPUTER DEPARTMENT
PROJECT COORDINATOR, COMPUTER DEPARTMENT

Vivekanand Education Society's Institute Of Technology, Mumbai

An Autonomous Institute affiliated to University of Mumbai, approved by AICTE

In association with VESIT - IQAC & VESIT - IIC

Department Of Computer Engineering

CERTIFICATE OF EXCELLENCE

This to certify that Yash Pahlani of D17B has won Consolation prize at Department Technology Day, **Pradarshini '24** conducted on 22nd-23rd March 2024.

DR. MRS. NUPUR GIRI
HOD, COMPUTER DEPARTMENT

DR. MRS. GRESHA BHATIA
DHOD, COMPUTER DEPARTMENT

PROJECT COORDINATOR
COMPUTER DEPARTMENT

Vivekanand Education Society's Institute Of Technology, Mumbai

An Autonomous Institute affiliated to University of Mumbai, approved by AICTE

In association with VESIT - IQAC & VESIT - IIC

Department Of Computer Engineering

CERTIFICATE OF EXCELLENCE

This to certify that Prabha Pamula of D17B has won Consolation prize at Department Technology Day, **Pradarshini '24** conducted on 22nd-23rd March 2024.

DR. MRS. NUPUR GIRI
HOD, COMPUTER DEPARTMENT

DR. MRS. GRESHA BHATIA
DHOD, COMPUTER DEPARTMENT

PROJECT COORDINATOR
COMPUTER DEPARTMENT

Progress review sheets

Sheet 1:

Inhouse/ Industry _Innovation/Research: Research

Sustainable Goal:

Project Evaluation Sheet 2023 - 24

Class: D17 A/B/C

Group No.: 35

Title of Project: Audio Emotion Recognition using Digital Stethoscope and Deep Learning Techniques

Group Members: Vivek Balani (5), Ritika Hotwani (27), Yash Pahlani (49), Prabha Pamula (51)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	5	4	3	4	2	2	2	2	2	3	3	3	3	4	47

Comments: Good Work.

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	5	4	3	4	2	2	2	2	2	2	3	3	3	4	46

Comments: good project

Date: 10th february, 2024

Name & Signature Reviewer 2

Sheet 2:

Inhouse/ Industry _Innovation/Research:

Sustainable Goal:

Project Evaluation Sheet 2023 - 24

Class: D17 A/B/C

Group No.: 35

Title of Project: Audio Emotion Recognition using Digital Stethoscope and Deep learning Techniques

Group Members: Vivek Balani (05), Ritika Hotwani (27), Yash Pahlani (49), Prabha Pamula (51)

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	5	5	3	5	2	2	2	2	2	3	3	3	3	3	

Comments:

Name & Signature Reviewer1

Engineering Concepts & Knowledge (5)	Interpretation of Problem & Analysis (5)	Design / Prototype (5)	Interpretation of Data & Dataset (3)	Modern Tool Usage (5)	Societal Benefit, Safety Consideration (2)	Environment Friendly (2)	Ethics (2)	Team work (2)	Presentation Skills (2)	Applied Engg&Mgmt principles (3)	Life - long learning (3)	Professional Skills (3)	Innovative Approach (3)	Research Paper (5)	Total Marks (50)
5	4	5	3	5	2	2	2	2	2	3	3	3	3	3	47

Comments:

Date: 9th March, 2024

Name & Signature Reviewer 2