SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I2P115

Volume 13 Issue 2, 195-207, February 2026
© 2026 Seventh Sense Research Group®

Original Article

A Hybrid Deep Learning Framework for Marathi Speech-Based Stress Detection Using GRU and Handcrafted Audio Features

Smita S. Patil[1], Meena Chavan[2]

[1]*Department of Electronics Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra, India.*
[2]*Department of E & TC Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra, India.*

[2]*Corresponding Author : meenachavan1180@gmail.com*

Received: 21 November 2025 Revised: 23 December 2025 Accepted: 24 January 2026 Published: 20 February 2026

Abstract *- Stress is one of the major factors affecting mental and physical health and requires the development of new non-invasive, early, and accessible detection methods. Prosodic and acoustic features indicative of emotional states can be employed for voice-based stress detection, which provides a non-invasive method. However, research on regional languages, such as Marathi, is scarce. This paper presented a hybrid deep learning-based framework for stress finding in Marathi speech that combines hand-tuned audio features with deep temporal representation using a Gated Recurrent Unit (GRU)-based network. The system applies pre-processing steps such as removing silent portions and reducing background noise in order to improve overall robustness. System extracted 19 handcrafted features, including MFCCs, Chroma, spectral contrast, zero-crossing rate, and spectral rolloff from each audio clip. Simultaneously, sequences of MFCCs are input to a GRU to model the temporal information. The outputs of the two feature branches are then concatenated and input into the fully connected layers to perform classification. The developed method demonstrated a 92% accuracy with F1-score of roughly 0.92 when examined in the regional Marathi language. This system has also been tested on CNN and RNN with accuracies of 84% and 78%, and the results show that the integration of statistical and temporal features with improved pre-processing leads to an improvement in stress detection performance, using which a scalable solution can be embedded for monitoring mental health in the context of a own-resource language.*

Keywords *- Stress, MFCC, Marathi, CNN, Deep learning.*

1. Introduction

Stress-related diseases are now acknowledged as a significant public health issue, with more than 264 million people affected worldwide. Long-term or poorly managed stress can result in major mental disorders, such as anxiety or depression, substance abuse, and, in very serious cases, suicide. Therefore, the early and accurate diagnosis of stress is currently an urgent requirement in both clinical and non-clinical environments. The Life Events and Challenges Timetable, the Perceived Stress Scale, and other tools based on standardized self-report measures or clinician interviews were employed to quantify stress. Despite their popularity, these tools are not the best. While self-report surveys may be subject to cognitive or social desirability bias and lack impartiality and item specificity, clinical interviews are labor-intensive and costly. Biological metrics such as cortisol, heart rate, and immune-related biomarkers are more objective, but they are generally more invasive (requiring blood or salivary samples) and are not practical for real-time or human-machine interaction use. On the other hand, speech appears to be a promising and non-invasive modality for stress identification. Speech is rich in emotional and physiological information that can be both prosodic (e.g., F0, intensity, and duration), spectral (e.g., mel-frequency), and temporal (e.g., temporal descriptors of speech) cues, which can reflect the psychological state of a speaker. In the domains of affective computing and human-computer interaction, automatic speech stress recognition is a hot topic due to the growing trend of human-machine and human-to-human communication. Although there has been significant progress in stress detection for English and other high-resource languages, little work has been carried out on regional and low-resource languages, such as Marathi. This is an important void because cultural and linguistic substrates can contribute to the profile of speech and emotional expression. This research provides a novel hybrid deep-learning strategy for

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

automatic stress identification in Marathi speech in order to address this challenge.

The proposed system uses both statistical handcrafted features (MFCCs, Chroma, spectral contrast, zero-crossing rate, and spectral roll off) and frame-level features from a Gated Recurrent Unit (GRU) neural network. To further improve system performance, superior audio initial processing techniques, such as silent removal and spectral noise reduction, are used. Such a hybrid model can exploit global statistical and temporal filter bank modulations in speech owing to stress.

This approach was tested on a balanced dataset of Marathi speech samples annotated processing + multistream representation, and as "stress" and "unstress." The findings show that the improved pre-deep temporal model's fused features significantly boost classification performance. Therefore, this study presents a scalable, non-intrusive, and culturally contextual method for stress detection in underrepresented languages, useful for applications such as mental health monitoring, workplace wellness systems, and intelligent human-computer interfaces.

Contribution and Organization of the Innovative Study This novel study introduces a hybrid framework for Marathi speech-based stress detection by combining GRU-based temporal modeling and manually created acoustic features.

Speech-based stress detection has seen much interest in recent years; however, the work done so far primarily focuses on languages such as English and other high-resource languages, where there is a lack of resources for regional and low-resource Indian languages like Marathi. Pronunciation rules across different groups and cultures contribute significantly to determining the manifestation of stress in speech, but unfortunately, such variability is rarely accommodated in existing works, making those systems less generalizable across diverse populations.

In addition, many existing methods depend on either rule-based linguistic features or deep learning based temporal modeling alone, and fail to integrate both of them effectively to model the entire range of stress-driven characteristics in speech. These constraints point towards the necessity of a sound, non-invasive, and culturally aware framework addressing Marathi speech explicitly. To fill this gap, a hybrid deep-learning framework that combines handcrafted statistical features with GRU temporal representations is supported by improved pre-processing for noise and silence removal.

The following is a summary of this study's main contributions.

Contribution 1: A curated and annotated database of 249 emphasized and 247 emphasized Marathi speech samples was collected from student participants, addressing the lack of region-specific emotional speech corpora.

Contribution 2: Implementation of a hybrid GRU + handcrafted model that merges statistical and sequential information from speech signals, outperforming traditional CNN and RNN architectures.

Contribution 3: Comparative analysis using accuracy, F1-score, specificity, Matthews's correlation coefficient, AUC-ROC, and Cohen's kappa demonstrated that the proposed model achieves 92% accuracy and improved robustness across all metrics.

Contribution 4: A thorough evaluation of feature engineering and selection strategies justifies the utilization of a reduced 19-feature subset to balance effectiveness and performance.

The framework of the paper is as follows: Section 2 covers past research and contextual information on speech-based stress identification. Section 3 delves into the datasets used, pre-processing processes, and feature extraction methodologies. Section 4 introduces a suggested hybrid model, which combines gated recurrent units and convolutional layers, as well as the baseline topologies for convolutional and recurrent neural networks.

Section 5 discusses the experimental setup and evaluation methods. Section 6 highlights results using graphs, tables, and an explanation of the findings. Finally, Section 7 finishes the paper by providing an overview of the work, its significance, and a potential direction for future research.

2. Literature Survey

In the past few decades, mental health has become an important issue, in which long-term stress is a major factor in a variety of health problems, such as anxiety, depression, cardiovascular disease, and lower quality of life. However, conventional methods for measuring stress, including questionnaires, interviews, and biological measures, are often cumbersome, invasive, and require skilled professional assistance.

Speech is a readily available source of rich information that is instantly accessible to others, and that shows changes in prosody, pitch, energy, and other acoustic features corresponding to physiological and psychological changes. However, the automated detection of stress in regional languages such as Marathi has not received sufficient attention.

Most previous studies have been conducted on resource-rich languages, such as English, which are not directly applicable to the Marathi language due to its phonetic, prosodic, and linguistic variation. Table 1 presents a literature survey conducted based on speech on different datasets.

Table 1. Literature survey table

Ref	Dataset Used	Methodology	Key Features
[1]	RECOLA and SEWA	Multi-resolution Modulation-filtered Cochleagram (MMCG)	LSTM & MMCG features
[2]	RAVDESS	MFCCs, Pitch/ Sample Rate	CNN
[3]	DAIC-WOZ	CNN	Spectrogram Texture
[4]	AVEC2013, AVEC2014	DCNN	Median Robust Extended Local Binary Patterns (MRELBP) from spectrograms
[5]	interviewed with questions	analyzed using PRAAT software	Spectrogram
[6]	wireless channels	digital signal processing algorithm	stress monitoring algorithm
[7]	IEMOCAP	several Deep Neural Network (DNN) architectures	MFCC-Text Convolutional Neural Network (CNN) model
[8]	Smartphone sensors	filter-based methods and standard binary-code chromosome Genetic Algorithm	C4.5, kNN, and Bayesian Network
[9]	KeioESD, EmoDB	Mean energy Mean intensity	ANN, SVM,GMM
[10]	Audio signals	Empirical Mode Decomposition	Hilbert, TEO, AM-FM
[11]	DAIC-WOZ	Deep Learning Pipeline	Spectrogram, Pre-processing

During this research work, a broad-spectrum review of previous work on stress/emotion prediction from speech was conducted, and it was found that several works for speech-based emotion/stress analysis have been done in multiple languages, but such research has been very scarce for Marathi. Initially, the work focused on handcrafted acoustic features for emotion recognition, such as MFCC, LPC, DWT, and FFT. These methods reported relatively low performing results, with most of them achieving class-wise accuracies between 60 – 75% for anger, sadness, and happiness, but were inherently developed for emotional prosody classification rather than stress detection.

Advanced systems such as deep neural networks and hybrid acoustic–prosodic models brought better learning of features but were still committed to emotion-level differences without accounting for physiological stress cues. More Recent approaches to emotion recognition in Marathi — e.g., mahaBERT, GPT-4, Llama3— have undertaken work on text-driven emotion recognition, achieving accuracies of 83–86%.

However, these methods are based exclusively on the text and therefore cannot capture acoustic features reflected in stress signatures (pitch variation, changes in spectral energy, or altered temporal behavior). Indeed, to the best of our understanding, no study both gives a Marathi-spelling stress-specific speech corpus and an explicit binary (stress vs. unstress) framework. These voids indicate that targeted work in detecting stress from low-resource Indian languages has not been undertaken, and a dedicated dataset, an appropriate emotion-to-stress conversion model, and an efficient hybrid network capable of capturing both statistical and temporal properties of speech are much needed. The current paper fills this void by building the first Marathi speech stress dataset and introducing a new architecture of fused GRU + handcrafted feature for non-intrusive stress detection.

3. System Methodology

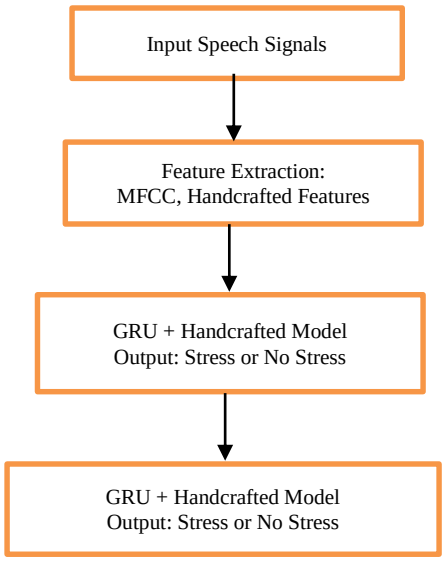


Fig. 1 System architecture

To improve the performance and generalizability of stress detection in Marathi speech, propose a hybrid deep learning architecture shown in Figure 1. In which manually designed acoustic features are combined with temporal features learned from recurrent neural networks. The hybrid solution combines two complementary sources of information: handcrafted features that provide statistical summaries of the speech signal, and deep temporal features

that capture the temporal dynamics of speech-related characteristics. A compact set of handcrafted features was developed, forming a 19-dimensional vector representation of the voice stream. This vector was produced by computing the top five mean values from major acoustic descriptors, namely Mel-Frequency Cepstral Coefficients (MFCCs), Chroma, Spectral Contrast, Spectral Rolloff, and Zero-Crossing Rate (ZCR). These features are good global acoustic descriptors that are both computationally fast and interpretable.

Simultaneously, 130×13 was extracted to maintain the dynamicity of the speech. These sequences are then passed through a Gated Recurrent Unit (GRU) network, which is particularly suitable for time–space data modeling and can preserve long-term dependencies without vanishing gradients. The GRU branch output (64-D) is further concatenated to the 19-D handcrafted feature vector as an 83-D fused representation. This combined feature vector was then processed by two fully connected dense layers (64 and 32 neurons) to capture high-order interactions and patterns. Finally, a sigmoid-activated output neuron performs binary classification to discriminate between stress and unstressed speech samples.

The hybrid model architecture takes advantage of the complementary modeling capabilities of deep learning, which uses sequential modeling and handcrafted features with domain knowledge, leading to higher classification accuracy and interpretability. The grouping of statistical and temporal representations enables the model to make more holistic decisions and effectively capture both short-term acoustic changes and long-term speech characteristics of stress.

3.1. Dataset Description

This study employed a dataset of 496 Marathi speech samples recorded by student volunteers in a controlled environment. The class distributions are presented in Figure 2. The samples were evenly distributed as follows:

Stress: 249 audio samples
Unstress: 247 audio samples

All of the samples were short Marathi-language sentences that were captured in a noise-free environment using standard microphones at an audio sampling rate of 16 kHz. To reduce the stress conveyed through semantic clues, the speech's content was emotionally neutral.

Participants were asked to read sentences in a relaxed state and under moderate stress (e.g., time pressure or copying/memory tests) to create a natural variation in voice stress indicators. It was manually tagged using visual context and checked for quality and balance. Real-life data were used to train and assess the proposed speech-based stress-detection model.

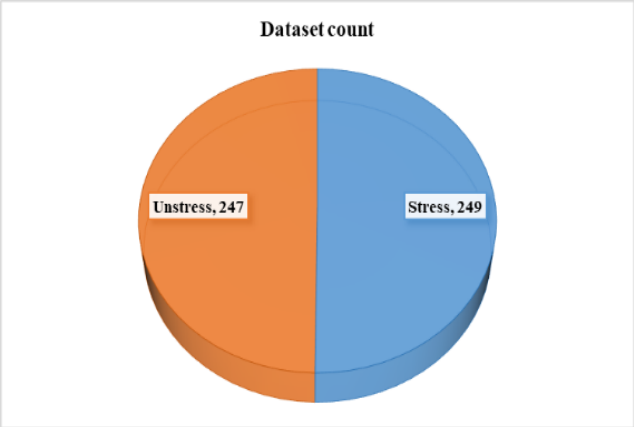


Fig. 2 Dataset count

3.2. Feature Extraction

In this study, both handcrafted acoustic features and temporal MFCC sequences are used to obtain a holistic representation of stress-related phenomena in Marathi speech. Nineteen handcrafted features were selected to represent spectral, prosodic, and temporal variabilities, which are typically affected by stress. The first five Mel-Frequency Cepstral Coefficients (MFCCs) are the metrics that make up these measures. The MFCCs are computed using the log power spectrum of the sound and serve as an approximation for the form of the short-term power spectrum. The STFT (Short-Time Fourier Transform) yielded the chroma features, which show the energy distribution of 12 spectral pitch classes. The system used the mean value of the first five Chroma coefficients to capture pitch-related changes.

A measure characterizing the contrast between spectral peaks and valleys within and between frequency bands is spectral contrast, parameterized by seven contrast values. The ZCR, which counts the number of audio signal change signs, was also utilized to record the frequency variation. Moreover, the spectral roll-offll was computed, which stands for the frequency at which 85% of the signal energy is contained. These features were averaged over time to obtain the fixed-length statistical vectors for each sample.

To capture temporal information along these handcrafted features, extracted the entire MFCC sequence from a 3-second time window, providing another matrix with dimensions (130-timesteps × 13 MFCCs). These ordered characteristics are fed into a GRU network capable of learning high-level temporal dependencies and speech dynamics with stress stimuli. This mixture of human-derived descriptors and GRU-based temporal modelling allows the system to effectively detect stressed and unstressed speeches.

3.3. Mel-Frequency Cepstral Coefficients (MFCC)

The spectral envelope of audio is captured by MFCCs in a manner that mimics the human auditory system. They are obtained by mapping a signal's log power spectrum onto the

Mel scale. To transform a signal into the frequency domain, use the Short-Time Fourier Transform.

Pass the power spectrum through a Mel filter bank:

$$M(m) = \sum_{k=1}^{N} |X(k)|^2 \cdot H_m(k)$$

where $H_m(k)$ is the m-th Mel filter, and $|X(k)|^2$ is the power at frequency bin k.

Take the logarithm of each Mel-filtered energy:

$$\log M(m)$$

Apply Discrete Cosine Transform (DCT):

$$\text{MFCC}(n) = \sum_{m=1}^{M} \log\big(M(m)\big) \cdot \cos\left[\frac{\pi n(m-0.5)}{M}\right]$$

3.4. Chroma Features

Chroma features (also called chromagrams) represent energy in each of the 12 pitch classes (C, C#, B), regardless of octave.

Formula: Based on energy per pitch class:

$$\text{Chroma}(c) = \sum_{f \in F_c} |X(f)|^2$$

Where:F_c: Set of frequencies mapped

• $|X(f)|^2$: Energy

Spectral Contrast: Spectral contrast measures the difference between peaks and valleys of the spectrum in each frequency sub-band. Formula:

$$\text{Contrast}_b = \text{Mean Peak}_b - \text{Mean Valley}_b$$

or in ratio form:

$$\text{Contrast}_b = \frac{\max(S_b)}{\min(S_b)}$$

Where S_b represents spectral magnitudes in sub-b and b.

Zero Crossing Rate (ZCR): High emotional stress can cause faster speech and irregular intonation, increasing ZCR

$$\text{ZCR} = \frac{1}{T-1}\sum_{t=1}^{T-1} 1\,[x_t \cdot x_{t-1} < 0]$$

Where: - x_t: Sample at time t - 1 is an indicator function that counts the sign change. The spectral roll-off is the frequency below which a specific percentage (often 85%) of the total spectral energy is contained. The equation

$$\sum_{f=0}^{f_{\text{rolloff}}} |X(f)|^2 = 0.85 \cdot \sum_{f=0}^{f_{\max}} |X(f)|^2$$

Where: - f_{rolloff}: Roll-off frequency - $|X(f)|^2$: Power spectrum

3.5. Temporal Features (MFCC Sequences)

Whereas the handcrafted features aggregate the global properties of a signal, temporal features maintain frame-wise variations over the course of time, which is important when recognizing stress-induced speech variations, such as tremor, jitter, or prosodic instability. These features represented these temporal dependencies via MFCC sequences (of 3-second audio) as inputs to a Gated Recurrent Unit (GRU). Table 2 shows the detailed summary of the extracted features with their counts.

Unlike mean MFCCs, which destroy the time-varying structure in speech, which is a property of speech that is affected by stress. This is advantageous because the model can learn the patterns of hesitation, speaking rate, and voice modulation that are lost in mean-based representations.

This creates a matrix of shape:

$$\text{MFCC}_{\text{sequence}} \in \mathbb{R}^{T \times D}$$

Where: - T = number of time frames (e.g., 130 for 3 seconds) - D = number of MFCCs

Table 2. Summary of extracted features

Feature Group	Feature Name	Count	Purpose
MFCC (mean)	MFCC 1–5	5	Speech timbre, vocal energy
Chroma	Chroma 1–5	5	Pitch class energy
Spectral Contrast	Contrast 1–7	7	Harmonics vs. noise
Spectral Rolloff	Single value	1	Energy distribution cutoff
Zero Crossing Rate	Single value	1	Signal frequency variation
Total (Handcrafted)		19	
MFCC (sequence)	13 MFCCs over 130 time steps	130×13	For GRU sequence learning
Total Features		149	

4. Feature Importance Analysis using ANOVA F-Test

A Univariate ANOVA F-test was used to assess the discriminative power of individual prosodic features for stress and unstressed speech frame classification. This quantitative test assesses how significant the variation in a feature's mean values is between the stress and unstress classes. Figure 3 shows the representation of important features generated by the ANOVA F-test.

The F score of a feature is obtained by the formula:

$$F = \frac{MSB}{MSW} = \frac{\sum_{k=1}^{K} n_k \, (\bar{x}_k - \bar{x})^2/(K-1)}{\sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 \, /(N-K)}$$

Where:

- K: Number of classes (2 in this case: Stress and Unstress)
- n_k: Number of observations in class k
- $\bar{x}_k$: Mean of feature x in class k
- $\bar{x}$: Overall mean of feature x
- N: Total number of samples
- MSB: Mean square between groups
- MSW: Mean square within groups

A higher F-score indicates greater discriminatory power of that feature.

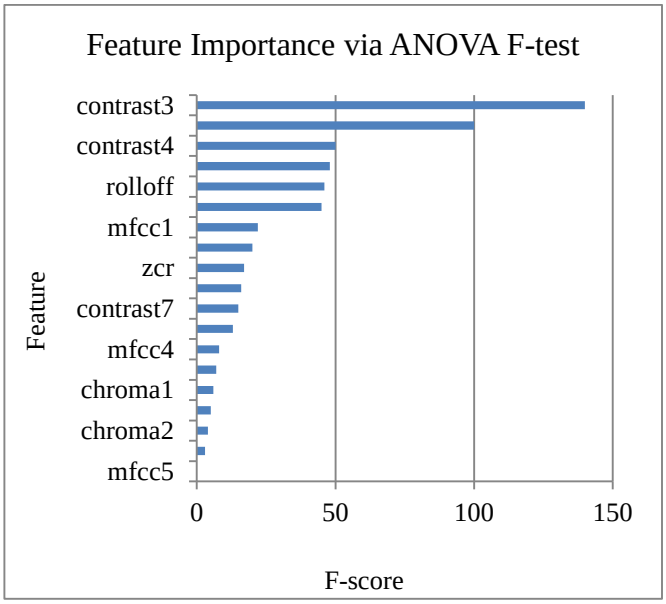


Fig. 3 Feature importance

While Figure 3 explicitly elucidates the preeminent elements informing stress classification, contrast3 stood out statistically, conveying a robust disparity between fraught and serene utterances. Mfcc3 has likewise emerged prevalently, capturing integral distributions pertinent to emotional and stressed speech deviations. Contrast4, contrast6, and rolloff likewise boasted substantive revealing power, reinforcing their importance in the categorization method.

On the other hand, features including mfcc5, chroma3, and chroma4 were assigned the lowest F-scores, implying their constrained relevance and feeble input to stress identification. These insights buttress judiciously opting for the most illuminating elements, which not only boost model precision but also aid in dimensionality reduction, particularly in hybrid models integrating both temporal and manually extracted elements.

5. Model Architecture

Figure 4 shows the proposed system, which begins with reading audio recordings categorized as stress and unstress. The signals are pre-processed using silence removal and noise reduction to enhance quality. Two parallel feature extraction pipelines are employed: handcrafted features (scaled with Standard Scalar) and MFCC sequences (padded to fixed length). Each feature set is divided into subgroups for testing and training.

The handcrafted features are passed through a Dense Neural Network, while MFCC sequences are learned via a GRU network for temporal dynamics. Outputs from both models are fused at the feature level, followed by additional dense layers and a sigmoid classifier. Finally, the system predicts the emotional state as either stressed or unstressed.

5.1. GRU

GRU variation intended for sequential data, like voice, serves as the foundational temporal modeling unit in this architecture. By keeping the dynamic memory of previous states when sweeping through time steps from left to right in a sequence, GRUs are able to record temporal dependencies in audio as they occur. GRUs employ a variety of gating techniques, including the update gate and reset gate, to regulate the flow of information, whereas normal RNNs struggle to retain information over extended timeframes due to the vanishing gradient problem.

Using the current input and the previous hidden state, the GRU calculates these gates analytically and fuses them to generate the new state. This enables the GRU to retain relevant information from the preceding portion and forget irrelevant information, making it particularly well-suited for modeling time-varying speech signals. In the framework of the report, the GRU network MFCC sequences, which are obtained from Marathi speech, eventually learn time-dependent stress-Induced vocal cues such as pitch variation, manner of sign, and prosody. Its output is a fixed-dimensional representation that characterizes the gist of such dynamic patterns, and is combined with handcrafted statistical features to enhance the robustness and context sensitivity of the final stress classification.

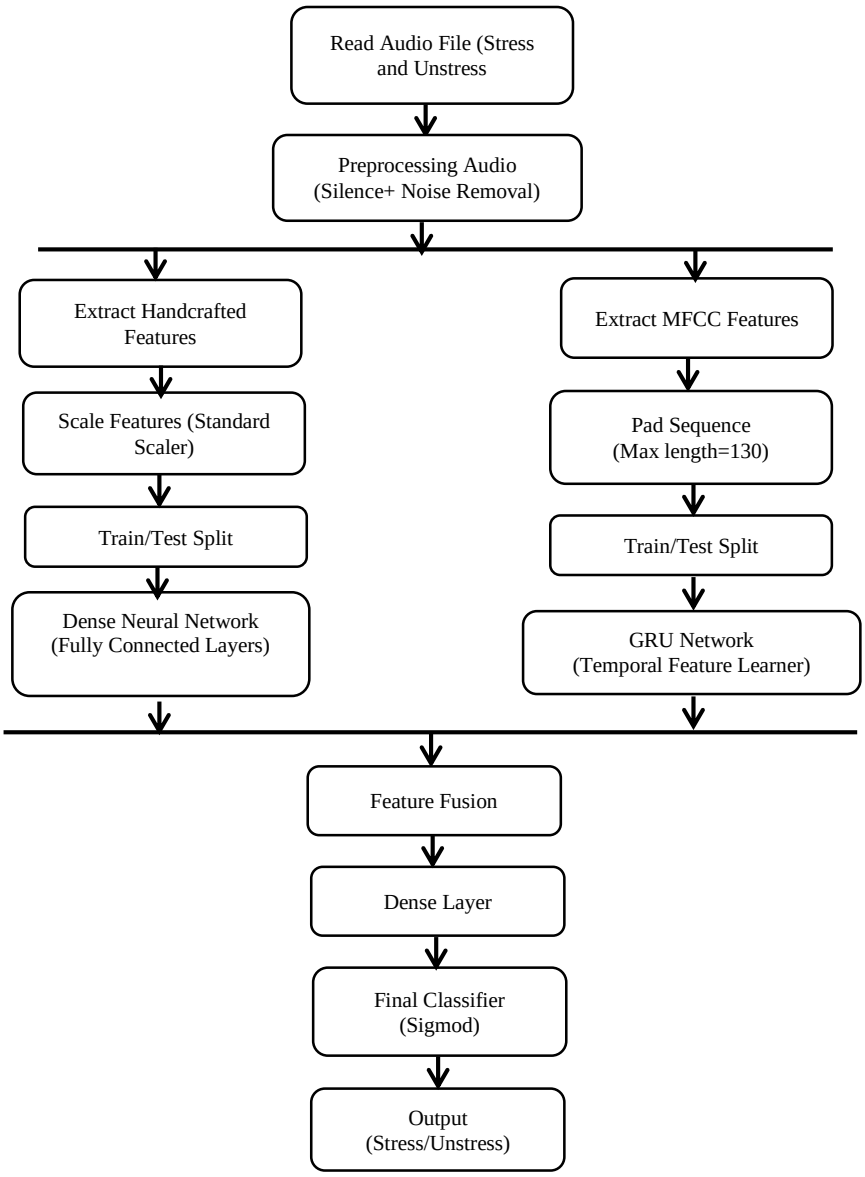


Fig. 4 System flow for hybrid

5.2. Dense Model

The dense layers in the proposed hybrid architecture are pivotal for learning and optimizing the higher-order representations in the combined feature space. A layer of neurons in which every neuron receives information from every other neuron in the layer above is known as a completely linked layer, or dense layer.

The output of a dense layer is calculated by adding a bias term to the weighted sum of the input from the preceding layer, after it has passed through a nonlinear activation function. Figure 5 shows the layers of the hybrid GRU model and the detailed layer-wise configuration mentioned in Table 3.

The model can capture the intricate nonlinear interactions between feature vectors thanks to a sigmoid or ReLU design. Upon concatenating the outputs of the GRU branch and the engineered feature branch, the concatenated vector is transferred through two dense layers. The initial dense layer (64 neurons with ReLU activation) learned complex patterns across both temporal and statistical domains. A dropout layer was applied between the dense layers to generalize the model. The representation was still abstracted by the second, a completely linked layer, which included 32 neurons.

Lastly, a single neuron with a sigmoid activation function builds the dense output layer to provide binary classification (whether or not a voice sample is stressed). A set of dense layers \(D\) composes the decision core of the model, where the model 'sees' the pooled multimodal features and comes up with a prediction it is confident about.

Table 3. Layer-wise configuration of the proposed hybrid deep learning model

Layer (Type)	Output Shape	# Parameters	Connected To
mfcc_seq (Input Layer)	(None, 130, 13)	0	–
not_equal_1 (Not Equal)	(None, 130, 13)	0	mfcc_seq
masking_1 (Masking)	(None, 130, 13)	0	mfcc_seq
any_1 (Any)	(None, 130)	0	not_equal_1
gru_1 (GRU)	(None, 64)	15,168	masking_1, any_1
dropout_6 (Dropout)	(None, 64)	0	gru_1
handcrafted (Input Layer)	(None, 19)	0	–
concatenate_1 (Concatenate)	(None, 83)	0	dropout_6, handcrafted
dense_9	(None, 64)	5,376	concatenate_1
Dropout	(None, 64)	0	dense_9
dense_10	(None, 32)	2,080	dropout_7
dense_11	(None, 1)	33	dense_10

This research introduces a novel architecture that utilizes a hybrid DL approach to apprehension temporal and statistical characteristics from Marathi speech signals in an effective manner for stress detection. The model consists of two parallel branches: one processes the time-sensitive features in speech with GRU, and the other takes handcrafted statistical features as input directly. Mel-Frequency Cepstral Coefficients (MFCCs), which are commonly employed to represent the short-term power spectrum of an audio source, are provided to the first branch in the form of a 130-step sequence with 13 coefficients each step. Before the inputs were sent to a 64-unit GRU layer, a mask layer was applied to allow for variable-length input and padded sequences. Sequential and intonation changes, which are significant sources of stress, can be encoded by this representation, since it is rich enough to prevent overfiting. A dropout layer that comes after the GRU randomly shuts off units during training.

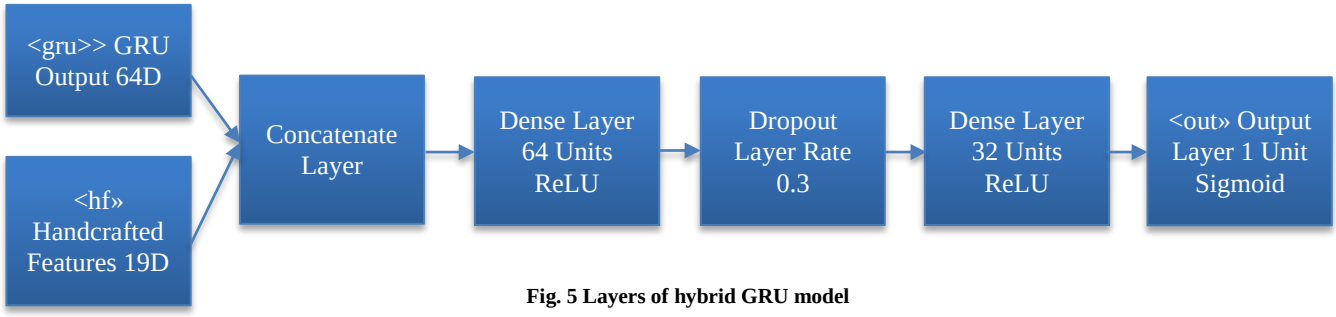


Fig. 5 Layers of hybrid GRU model

Simultaneously, the second branch uses a manually constructed 19-dimensional feature vector that includes zero-crossing rate, chroma characteristics, MFCC means, spectral contrast, and spectral rolloff. These features provide a summary of the spectral and temporal contents of the signal and are scaled using standard normalization.

In contrast to the GRU branch, the handcrafted feature branch was not further processed over time and was fed directly to the model. The GRU and handcrafted branches predict the results and are concatenated to form a unified feature vector that combines the temporal and statistical information.

Multiple dense layers were given this concatenated representation: a dense layer with 32 neurons and ReLU activation, a new dropout layer, and a fully linked layer with 64 neurons and ReLU activation. The final judgment, which returns the likelihood that the speech sample is stressed, is made by a single neuron employing sigmoid activation.

This hybrid model architecture can simultaneously incorporate the time-evolving-influenced dynamic characteristics of the speech signal and the completely affective nature of acoustic features, thereby facilitating a strengthened and precise classification system for detecting stress in Marathi speech.

5.3. Feature Fusion and Classification

After the extraction of the temporal and statistical features in parallel, the proposed method adopted a feature fusion scheme to fuse the output of the GRU and the handcrafted branch as an integrated feature representation. The GRU branch effectively models the sequential information and dynamics of speech with a 64-dimensional output vector, whereas the handcrafted branch yields a 19-dimensional vector that includes acoustic features such as MFCC mean, Chroma, Spectral Contrast, Spectral Rolloff, and Zero-Crossing Rate. These two results are concatenated by a concatenate layer to form an 83-D feature vector that

summarizes both the temporal evolution and global signal properties. This combination enables the model to have a wider and more informative feature space that can better differentiate between stressed and unstressed speeches. Although it is possible to extract a much larger feature set from speech acoustics, we deliberately restricted ourselves to 19 selected handcrafted features to ensure the best possible performance of the model and its generalization. Having too many features can result in redundancy and irrelevant information; hence, the curse of dimensionality — the feature space becomes very sparse, and the model can become overfit, thereby increasing the computational cost. By considering only the most discriminative features, which were selected by means of statistical testing (in this case, ANOVA F testing), the model could capture the crucial differences between stressed and unstressed speech while preserving an adequate compromise between performance, interpretability, and computational time. This alternative permits the method to be reliable even with smaller or noisier datasets and also facilitates pragmatic and clinical applications.

A concatenated vector is fed into dense layers that hierarchically perform feature abstraction and nonlinear transformations. The first dense layer (64 neurons) with ReLU activation of the nonlinearities between temporal and statistical cues. To avoid overfiting and improve generalization, a dropout layer with a dropout rate of 0.3 was used. The merged model is further abstracted by a second dense layer with 32 neurons and 3.5 Post-processing. Lastly, it links the RNN's output to a Dense layer with a single neuron (neurons=1) that outputs the likelihood that the voice sample is stressful or unstressed, using sigmoid activation. This architecture is intended to allow the better use of complementary feature sets for accurate and robust stress classification.

6. Experimental Setup

A supervised learning technique designed to categorize Marathi speech samples into "Stress" and "Unstress" groups was used to train the hybrid deep learning model. The training process involved several carefully selected hyperparameters and optimization strategies to ensure optimal model convergence and generalization.

6.1. Loss Function and Optimizer

6.1.1. Loss Function

Given that the task involves binary classification, the binary cross-entropy loss technique was used to assess the divergence between projected probability and actual class labels.

$$\text{LogLoss} = \frac{-1}{N}\sum_{i=1}^{N}\sum_{c\in\{0,1\}} y_{i,c}\ln(p_{i,c})$$

The function quantifies the gap between actual events and expected probability.

6.1.2. Optimizer

Adam Optimizer Adam (Adaptive Moment Estimation) was chosen for its robustness and adaptive learning rate capabilities. It combines the advantages of both AdaGrad and RMSProp.

6.2. Hyperparameters

Table 4 shows the hyper parameter details for the model which trained with a batch size of 16 for 20 epochs with the learning rate 0.001, which is the default setting of optimizer. It uses a validation split of 20% for robustness during training and sets the dropout to 0.3 in order to reduce overfiting and enhance the generalizability. In the recurrent architecture, the GRU layer with 64 units accurately learned temporal dependencies in MFCC sequences. The fully connected network consisted of 2 dense layers: Fully Connected 64 ReLU and Fully Connected 32 ReLU, hierarchically learning features for final classification.

Table 4. Hyper parameters

Parameter	Value
Batch Size	16
Number of Epochs	20
Learning Rate	0.001 (default)
Validation Split	0.2 (20%)
Dropout Rate	0.3
GRU Units	64
Dense Layer 1	64 neurons
Dense Layer 2	32 neurons

6.3. Evaluation Metrics

Several common assessment metrics were used to objectively handcraft a feature model. These measures not only evaluate the model's overall accuracy but also provide insight into its ability to correctly distinguish between stressed and unstressed speech samples, and evaluate the performance of the proposed hybrid GRU + handmade. The following metrics, shown in Table 5, were computed on the test set:

Table 5. Evaluation metrics

Performance Measure	Mathematical Representation
Accuracy	(TPR + TNR) / (TPR + TNR + FPR + FNR)
Precision (PPV)	TPR / (TPR + FPR)
Recall (Sensitivity)	TPR / (TPR + FNR)
Specificity (TNR)	TNR / (TNR + FPR)

False Positive Rate (FPR)	FPR / (FPR + TNR)
Negative Predictive Value (NPV)	TNR / (TNR + FNR)
F1 Score	2 × (Precision × Recall) / (Precision + Recall)
Matthews Correlation Coefficient (MCC)	(TPR×TNR – FPR×FNR) / √[(TPR+FPR)(TPR+FNR)(TNR+FPR)(TNR+FNR)]
Cohen's Kappa	$(P_o – P_e) / (1 – P_e)$
Logarithmic Loss	$–(1/N) \Sigma [y \log(\hat{y}) + (1–y) \log(1–\hat{y})]$

7. Discussion

7.1. GRU + Handcrafted Features Model

To improve the performance of the speech stress detection system, a hybrid model combining a Gated Recurrent Unit (GRU) network and a handcrafted feature branch was developed. This method merges the temporal dynamics modelled by the GRU from MFCC sequences (130 frames, 13 MFCC coefficients) and the statistical characteristics of handcrafted features (19 feature dimensions), which led to 83 input features.

The model GRU + Handcrafted Features consists of a GRU layer with 64 hidden units (Gated Recurrent Unit), dropout, and an input branch that takes the handcrafted feature vector as input. Before being produced as a sigmoid for classification, these branches were concatenated and processed through dense layers that were fully connected (64 and 32 units). The Binary Cross-Entropy loss function, Adam optimizer, and batch size of 16 were used to train the model over 50 epochs. An early stopping strategy was used with the validation loss to control overfiting and encourage the best generalization that could be obtained by the model. The model demonstrated steady convergence throughout training, with consistent decreases in training and validation losses. Validation accuracy reached a plateau around 35-40 epochs. The final results were further confirmed by demonstrating its excellent class discrimination ability compared with CNN and RNN.

The performance of the GRU + Handcrafted Features model in Table 6 and observe that it significantly outperformed the separate CNN (84%) and RNN (78%), with a validation accuracy of 92%. In the case of class-0 (unstressed), precision increased to 94%, recall was 90%, and the F1-score was 92%. However, precision and recall for Class 1 (stressed) were impressive, with a precision of 90% and recall of 94%, which gave an F1 score of 92%.

Additional metrics also confirmed the superiority of the joined method.

Specificity (TNR): 96%, indicating that it could efficiently detect negative cases.

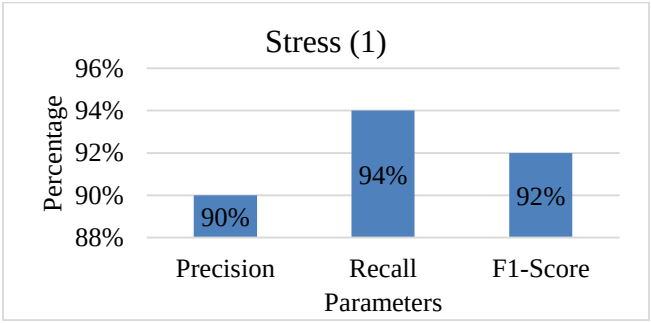
NPV of 89%, demonstrating the credibility of the unstressed cases being predicted.

The MCC and Cohen's Kappa Score were 84% (both), indicating good agreement and predictive reliability.

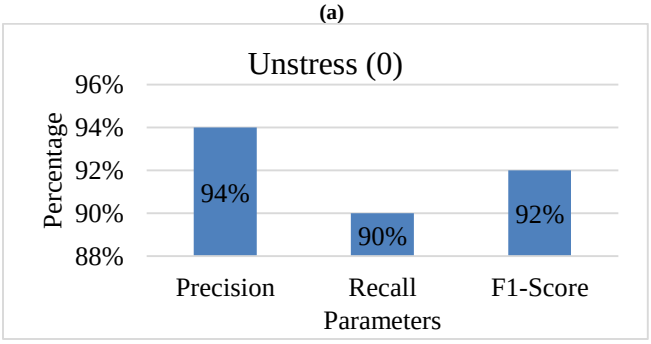
Log Loss: 26% which means that are relatively certain about the output probabilistic value.

Table 6. Results by using the hybrid model

Metric	GRU + Handcrafted
Precision (Negative Class)	0.94
Recall (Negative Class)	0.90
F1-Score (Negative Class)	0.92
Precision (Positive Class)	0.90
Recall (Positive Class)	0.94
F1-Score (Positive Class)	0.92
Accuracy	0.92
Specificity (TNR)	0.96
False Positive Rate	0.04
Negative Predictive Value (NPV)	0.89
MCC	0.84
Cohen's Kappa Score	0.84
Log Loss	0.2599
AUC-ROC Score	0.9500



(a)



(b)

Fig. 6 Classification of results for a) Stress, and b) Unstress.

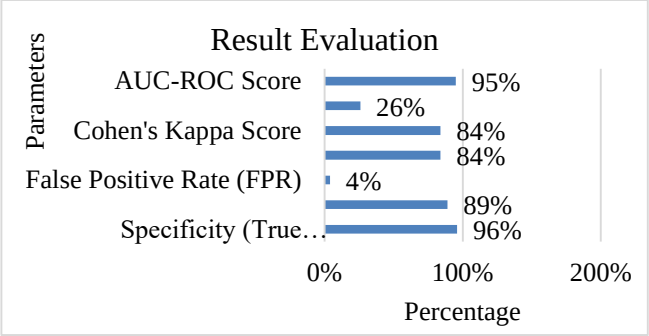


Fig. 7 Complete analysis of the result

Results Figures 6 and 7 depict the results of testing the proposed stress detection system. For both the Stress (1) class, 90% precision was attained by the model. recall of 94%, and F1-score of 92%, indicating strong sensitivity of detection. Moreover, the Unstress (0) class would have about 94% precision, 90% recall, and provide the same F1-score of 91% so that balanced performance on both classes was ensured.

The general performance measures (as depicted in Figure 7) also reinforce the strength of the model, with AUC-ROC of 95%, log loss of 26% and Cohen's Kappa and MCC values at 84% which denote substantial agreement and balanced classification. Furthermore, the system achieved a low false positive rate of 4%, a negative predictive value of 89%, and a specificity of 96%, indicating its ability to achieve accuracy for discrimination of stress and unstress.

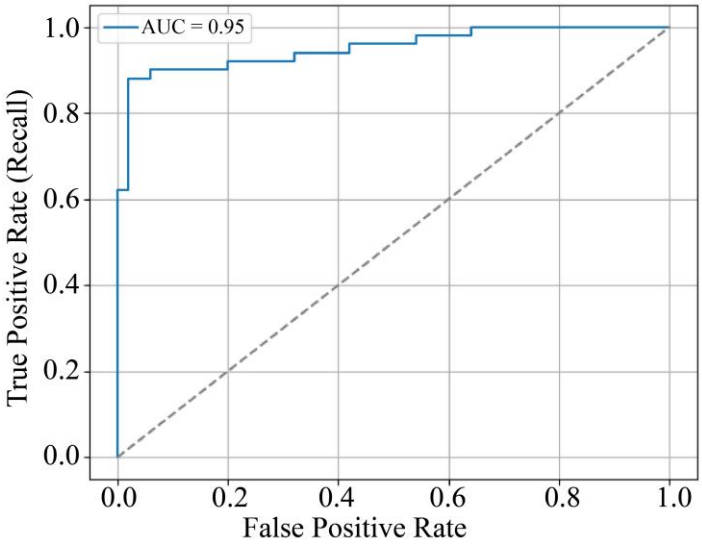


Fig. 8 ROC curve

The Receiver Operating Characteristics (ROC) curve in Figure 8 represents the classification performance of the proposed model at different thresholds. The AUC value of 0.95 suggests a very good discriminative performance, indicating that the model can differentiate between classes reasonably well. A larger AUC value tending to 1.0 indicates a stronger and more stable model, and thus, it is suitable for speech classification such as stress recognition.

A comparison of previous Marathi speech studies reveals that existing work has concentrated almost entirely on emotion recognition, not physiological stress detection. Methods based on MFCC, LPC, DWT, and FFT primarily classify isolated emotions such as anger, sadness, happiness, fear, or boredom. Even though these emotions can be mapped to stress categories (e.g., Stress = Anger + Sad, Unstress = Happy), none of these studies were designed for binary stress detection, nor did they include targeted pre-processing for stress cues such as spectral noise reduction or silence removal. Their performance also varies significantly: MFCC-based studies typically achieve 62–75%, FFT-based recognition reaches 87–100% but only for specific isolated emotions, and deep-learning attempts show high precision for some classes but extremely low recall for others (e.g., Fear: 100% precision but 33% recall), indicating instability across emotional categories.

Transformer-based works such as L3Cube-MahaEmotions achieve 83–86% accuracy but operate purely on text, meaning they do not capture acoustic stress markers like pitch instability or spectral shifts. None of the existing studies provides a stress-specific corpus, nor a binary stress–unstress framework, which is essential for mental health or human–machine interaction scenarios.

In contrast, the proposed Hybrid GRU + Handcrafted model is the first to explicitly formulate Marathi stress detection by mapping emotional states (Angry + Sad → Stress, Happy → Unstress) into a binary.

Table 7. Comparison analysis

Study / Method	Paper Name	Corpus Description	Mapping to Stress / Unstress	Speakers	Recognition Accuracy / Results	Model / Method
Emotional Prosody Speech Corpus (MFCC vs LPC) [12]	Emotional Speech Recognition for Marathi Language	Continuous Marathi sentences	Stress = Anger + Sadness Unstress = Happiness	24 speakers (14 Male, 10 Female)	MFCC Results: Anger (67.40%), Happy (70.76%), Sad (62.14%)	MFCC and LPC
L3Cube-MahaEmotions (GPT-4 vs CoTR Models) [13]	L3Cube-MahaEmotions: Synthetic Annotations using CoTR prompting & LLMs	Marathi Emotion Recognition Dataset	Stress = Anger + Sadness Unstress = Happiness	–	GPT-4 (83%), GPT-4 CoTR (86%)	MahaBERT, MuRIL, GPT-4, Llama3
MFCC + DWT Emotion Recognition [14]	Recognition of Emotion from Marathi Speech Using MFCC and DWT Algorithms	Marathi continuous sentences	Stress = Anger + Sadness Unstress = Happiness	10 speakers (5M, 5F)	Anger (75%), Boredom (84%), Sadness (89%), Neutral (64%)	MFCC, DWT
FFT-based Emotion Recognition [15]	Emotion Recognition in Marathi using FFT	Marathi emotional words database	Stress = Anger + Sad Unstress = – (No "Happy" class)	–	Surprise (100%), Fear & Anger (87.5%)	FFT
Deep Learning Emotion Recognition [16]	Marathi Speech Emotion Recognition using Deep Learning	180-phrase Marathi database	Stress = Anger + Sadness Unstress = Happiness	–	Anger: 79% Precision, 92% Recall; Fear: 100% Precision, 33% Recall	Deep Neural Networks
Hybrid Stress Detection (Proposed)	Recognition of Stress in Marathi Speech using a Hybrid Model	Marathi stress–unstress speech samples	Stress = Angry + Sad Unstress = Happy	–	Accuracy = 92%, Precision (94%), Recall (94%), AUC = 0.95	GRU + Handcrafted Acoustic Features

Classification task and training on a newly created speech corpus. The proposed system integrates global acoustic features (MFCCs, Chroma, Spectral Contrast, ZCR, Rolloff) with GRU-based temporal modelling, supported by advanced preprocessing. As a result, it achieves 92% accuracy, 94% precision/recall, and an AUC of 0.95, significantly outperforming all earlier approaches in consistency, binary-class discrimination, and suitability for real-world stress monitoring. This establishes the first benchmark for Marathi speech stress detection and fills a major research gap in low-resource languages.

8. Conclusion

This paper presents a hybrid deep learning model for stress detection from Marathi speech, which is an amalgamation of GRU-based temporal modeling of MFCC sequences with statistical handcrafted acoustic features such as MFCC means, Chroma, Spectral Contrast, Rolloff, and Zero-Crossing Rate. The algorithm is capable of capturing the temporal dynamics as well as the statistical information contained in the speech signal, and it can be considered a robust framework for automatic stress detection. The test accuracy reached 92%, which is higher than that of previous deep learning techniques, such as the RNN-based model (77%) and the CNN-based model (85%), demonstrating good performance in recognizing stress patterns. The evaluation of precision, recall, F1-score, specificity, AUC-ROC, and MCC demonstrated the usefulness and credibility of the approach. These findings show that hybrid deep learning strategies could be useful in addressing the challenges of stress based on speech, particularly in low-resource languages, such as Marathi. Future work could consider the extension of this

architecture to multiple languages and the application of more complex techniques, such as attention mechanisms or self-supervised embeddings, for a more robust performance.

Data Availability Statement

All of the data used was collected from the authors' software and tools' simulation reports. With appropriate rights, authors are attempting to accomplish the same with real-world data.

Consent to Participate

Verbal informed consent was obtained from all individual subjects included in the study.

Contributions of the Author

Author 1: Participated in the approach, formulation, gathering data, and developing the paper.
Author 2: Covered the writing, editing, and analysis of the broad concept.

References

[1] Zhichao Peng "Multi-Resolution Modulation-Filtered Cochleagram Feature for LSTM-based Dimensional Emotion Recognition from Speech," *Neural Networks*, vol. 140, pp. 261-273, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] S. Vaikole et al., "Stress Detection through Speech Analysis using Machine Learning," *International Journal of Creative Research Thoughts*, vol. 8, no. 5, pp. 1-6, 2020. [Google Scholar] [Publisher Link]

[3] Himani Negi et al., "A Novel Approach for Depression Detection using Audio Sentiment Analysis," *Proceedings 4th International Conference Computers & Management (ICCM)*, pp. 43-46, 2018. [Google Scholar]

[4] Lang He, and Cui Cao, "Automated Depression Analysis Using Convolutional Neural Networks from Speech," *Journal of Biomedical Informatics*, vol. 83, pp. 103-111, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] Savita Sondhi et al., "Vocal Indicators of Emotional Stress," *International Journal of Computer Applications*, vol. 122, no. 15, pp. 1-16, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[6] Fatma M. Talaat, and Rana Mohamed El-Balka, "Stress Monitoring using Wearable Sensors: IoT Techniques in Medical Field," *Neural Computing and Applications*, vol. 35, pp. 18571-18584, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Suraj Tripathi et al., "Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions," *arXiv preprint*, pp. 1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Ensar Arif Sağbaş, Serdar Korukoglu, and Serkan Ballı, "Real-time Stress Detection from Smartphone Sensor Data Using Genetic Algorithm-Based Feature Subset Optimization and K-Nearest Neighbor Algorithm," *Multimedia Tools and Applications*, vol. 83, pp. 1-32, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Kevin Tomba et al., "Stress Detection Through Speech Analysis," *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, Porto, Portugal, vol. 1, pp. 394-398, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[10] Rajib Sharma et al., "Empirical Mode Decomposition for Adaptive AM-FM Analysis of Speech," *Speech Communication*, vol. 88, pp. 39-64, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] Himanshu Churi et al., "A Deep Learning Approach for Depression Classification using Audio Features," *International Research Journal of Engineering and Technology*, vol. 8, no. 3, pp. 2930-2935, 2021. [Google Scholar] [Publisher Link]

[12] Bharati Borade, and R.R. Deshmukh, "Emotional Speech Recognition for Marathi Language," *Journal of Advanced Applied Scientific Research*, vol. 6, no. 3, pp. 85-105, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Nidhi Kowtal, and Raviraj Joshi, "L3Cube-MahaEmotions: A Marathi Emotion Recognition Dataset with Synthetic Annotations using CoTR prompting and Large Language Models," *arXiv preprint*, pp. 1-9, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[14] Dipti D. Joshi, and M.B. Zalte, "Recognition of Emotion from Marathi Speech using MFCC and DWT Algorithms," *International Journal of Advanced Computer Engineering and Communication Technology*, vol. 2, no. 2, pp. 59-63, 2013. [Google Scholar] [Publisher Link]

[15] R. Shinde Ashok et al., "Emotion Recognition in Marathi Language by using Fast Fourier Transform," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 10, pp. 43-47, 2019. [CrossRef] [Publisher Link]

[16] Akhilesh Ketkar et al., "Marathi Speech Emotion Recognition using Deep Learning Techniques," *Journal on Computer Hardware, Signal Processing, Embedded System and Networking*, vol. 5, no. 1, pp. 1-4, 2024. [CrossRef] [Google Scholar] [Publisher Link]