SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I7P109

Volume 12 Issue 7, 123-146, July 2025


*Original Article*

# An Analysis of Machine Learning Classifiers for Stress Detection Using Audio Features from TESS and RAVDESS Datasets

Smita Sagar Patil[1*], Meena Chavan[1]

[1]*Department of Electronics Engineering, Bharati Vidhyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra, India.*

**Corresponding Author : smitainstru5@gmail.com.*

Received: 06 May 2025 Revised: 08 June 2025 Accepted: 07 July 2025 Published: 31 July 2025

***Abstract*** *- This research addresses the critical need for accurate stress detection using speech signals, leveraging Machine Learning (ML) approaches applied to two distinct datasets: RAVDESS and TESS. Stress detection is pivotal in mental health monitoring and human-computer interaction; however, existing solutions often fail to generalize across diverse datasets due to the varying emotional complexities. The research gap lies in developing robust ML frameworks capable of handling nuanced emotional features, especially from datasets like RAVDESS, which exhibit significant overlap in stress-related signals. Comprehensive audio features, including Zero Crossing Rate (ZCR), Root Mean Square Energy (RMSE), Spectral Centroid, Spectral Bandwidth, Spectral Contrast, Spectral Rolloff, and Chroma features, are extracted to capture critical frequency and energy patterns. The study employs a suite of ML classifiers such as Random Forest (RF), Logistic Regression (LoR), Gradient Boosting (GB), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Support Vector Machines (SVM) with various kernels, along with an ensemble Voting Classifier. Among the models, SVM (linear) and Voting Classifier performed best, achieving 100% accuracy on TESS and up to 88.97% on RAVDESS. In contrast, NB showed lower performance, particularly on RAVDESS, with an accuracy of 72.06%. These findings reflect the sensitivity of model performance to dataset complexity and class separability. The significance of this study is in highlighting the impact of dataset characteristics on ML performance, providing a framework for feature extraction and model selection. Enhanced results confirm the necessity of tailored approaches for stress detection, paving the way for more sophisticated, dataset-aware methodologies to expand accuracy and reliability in real-world applications.*

***Keywords*** *- Machine Learning, Stress detection, RAVDESS, TESS, Voting Classifier, K-Nearest Neighbors, Gradient Boosting, Support Vector Machines.*

## 1. Introduction

Emotions are inner reactions that people experience in response to an event or scenario. Emotions are classified as happy, angry, sad, excited, peaceful, or neutral [1]. Emotions are an essential topic in communication. People can express their emotions through facial expressions and conversation. Since body language is often difficult to see or interpret, understanding emotions through sound becomes more crucial [2]. Humans communicate mostly through speech. Voice recognition, like speaker identification, is based on the analysis of audio signal data, as well as the construction of sound models, and, finally, when appropriate, language models. [3]. Understanding emotions from conversation is a sub-branch of voice detection. It can be challenging to interpret in human interactions, making determining human emotions using machines complex. Recent research has focused on ML and Deep Learning (DL), language, geography, age, cultural differences, and gender, which continue to have an impact on process performance [4]. Speech emotion recognition is most commonly utilized in medical care, educational recreation, and helping with driving, customer service, and online training. Essentially, this research falls into various categories that are hard to appraise. Communication signals vary in speed and properties, and they are uncontrolled [5]. Depression is one of the most prevalent diseases in everyday life. Stress has a major impact on people's lives. Anxiety can lead to a range of diseases, particularly heart problems, lung problems, breathing issues, and cancer. Global population growth has led to higher stress levels among individuals. Stress is a prevalent ailment among humans today. Stress can lead to serious, life-threatening issues for individuals worldwide. Stress can alter a person's behavior and disrupt daily routines. To precisely analyze their behavior, observe them using ML algorithms incorporated

into the scheme continuously. In India, 85% of individuals suffer from stress. Nowadays, people suffer from depression as a result of their stress [6]. Everyone experiences stress in various forms, making it difficult to measure mental stress [7]. Additionally, the accuracy of assessing and analyzing mental stress is influenced by the methodology used. Traditionally, stress has been assessed using subjective approaches. The results of the self-report questionnaires involve the estimated stress level. [8], are the most often utilized method [9]. Several studies have used questionnaires, self-reports, and interviews to evaluate mental stress levels. However, surveys are usually biased and require the user's full focus. People could not be knowledgeable of their real anxiety levels. Self-report surveys may not correctly reflect stress levels. Furthermore, they appear to provide less information than physiological assessments. Physiological indicators of stress include Heart Rate Variability (HRV), Electrodermal Activity (EDA), Electromyogram (EMG), blood vessel pressure, eyeball size, cytoplasmic hormone cortisol and salivary alpha amylase [10]. Other factors, such as mental stress, can have an impact on physiological signs. Circadian rhythm affects cortisol levels, which fluctuate throughout the day. Physical activity has an impact on salivary alpha amylase levels [11], while EDA is affected by skin illness and humidity [12].

To overcome the limitations of subjective methods and the variability of physiological signals, recent research has turned to speech-based stress detection. While previous studies have utilized individual classifiers or limited datasets, they often struggle with emotional overlap and lack the robustness needed for generalization across diverse data. Many existing models achieve high accuracy on specific datasets but fail to perform consistently in more complex emotional environments, such as those represented in the RAVDESS dataset. The innovation of this work lies in its integration of both fundamental and advanced spectral audio features, such as Mel-Frequency Cepstral Coefficients (MFCC), Chroma features, and Spectral Centroid, with a diverse range of ML algorithms.

Furthermore, the study explores how classifier performance is influenced by the clarity of emotional content in the datasets. It demonstrates that ensemble methods, particularly the Voting Classifier, offer better generalization capabilities, especially in datasets where emotional categories are more distinctly separated. This dataset-aware approach provides a scalable and robust framework for speech-based stress detection, an area that has received limited attention in the literature. In speech emotion recognition, existing ML models for stress detection still face robustness and cross-dataset generalization challenges. Most perform well on structured datasets but falter when confronted with complex emotional data where stress features overlap with other emotions. Moreover, limited comparative analysis exists regarding how different classifiers perform across datasets with varying emotional clarity. Addressing this research gap, the present study systematically evaluates a broad range of ML classifiers including Random Forest (RF), Logistic Regression (LoR), Gradient Boosting (GB), K-Nearest Neighbors (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM) with multiple kernel functions alongside an ensemble Voting Classifier. These models are applied to both the TESS and RAVDESS datasets to assess how dataset characteristics influence classification performance. To enhance emotional representation, the study incorporates advanced audio features such as Zero Crossing Rate, Spectral Centroid, Chroma features, and MFCCs. Ultimately, this work presents a robust and adaptable system for speech-based stress detection, underscoring the critical role of feature engineering and classifier selection in handling a wide range of emotional complexities in audio data.

The key contribution of the study is stated as follows:

- The study employs advanced audio features such as ZCR, RMSE, Spectral Centroid, and Chroma features to capture detailed frequency and energy patterns, improving the robustness of stress detection models.
- Through the comparison of ML classifiers' performance on two datasets (TESS and RAVDESS), the study underlines the importance of dataset properties. TESS's well-defined emotional borders are opposed to RAVDESS's subtle data, illustrating the necessity of specific model approaches according to data complexity.
- The research discloses performance heterogeneity with regard to datasets, especially with richer emotional data (RAVDESS), revealing a disparity in the current models' generalizability. This finding warrants the need for more advanced, dataset-conscious methods in stress detection studies.

The present study indicates the positive impact of diverse ML classifiers in stress detection using audio data, where ensemble models such as the Voting Classifier perform better. The significance of dataset properties is also brought out, as the distinct emotional characteristics of TESS result in more accurate results than the subtle RAVDESS database. Detailed feature extraction makes the model more robust, and the necessity for custom strategies according to data complexity is highlighted. These findings open up the possibility of more advanced, responsive stress detection devices in real-world use.

## 2. Literature Survey

The survey of literature emphasizes progress and limitations in ML model-based audio-based stress detection. Different investigations have tried feature extraction methods like Zero Crossing Rate, Spectral Centroid, and Chroma features to understand emotional intonations in speech. The set of classifiers including RF, LoR, and SVM variants, has proven to be of differential efficacy based on dataset

complexity. Studies show that for less complex datasets, such as TESS, higher accuracy results from distinguishable features, whereas more complex datasets like RAVDESS are problematic in that overlapping emotional signals occur. Current models lack generalizability, highlighting the importance of advanced, dataset-specific strategies to enhance stress detection consistency.

Ismail et.al [13] improved the voice recognition process; the improvement of the proposed framework incorporates SVM with the Dynamic Time Warping (DTW) technique. The proposed technique is an ML-based scheme that can drive intelligent devices with an accuracy of 97% using voice commands. The findings enabled patients and older individuals to use and control IoT devices using speech recognition technology. The recommended voice detection system is scalable, versatile, and compatible with existing smart Internet of Things devices. It also protects privacy when monitoring patient equipment. The research shows a great method for connecting systems across medical facilities to help the elderly and sick. It may deteriorate in loud surroundings, and its integration with varied IoT platforms requires additional testing for universal compatibility.

Rejaibi et.al [14], the proposed technique outperforms current methods on the DAIC-WOZ dataset in diagnosing depression, using a total precision of 76.27% and a root mean square error of 0.4, as well as predicting anxiety levels, with an RMS error of 0.168. The proposed framework has numerous benefits (including speed, non-invasiveness, and non-intrusion) that make it perfect towards applications that work in real time. The efficiency of the suggested strategy is evaluated using multi-modal and multi-feature tests. MFCC characteristics provide important data on anxiety. Including optical action units and other sounds boosts precision in classification by 20% and 10%, reaching 95.6% and 86%, respectively. However, the framework may struggle with cultural variations in depression expression and noisy real-time data inputs.

Liapis et.al [15], the study examines the effectiveness of Wearable Stress And Emotion Detection (WESAD), a publicly available physiological resource, with respect to User Experience (UX) evaluations. Three common ML methods for categorizing and a simple feedback DL Artificial Neural Network (ANN) that incorporates constant variables with entity embedding were trained using electrodermal activity (EDA) and Skin Temperature (ST) inputs from WESAD. Two training methods (DL and ML) attain an accuracy of 97.4%. Excellent results were obtained for the developed approaches' pressure detection capability in a variety of contexts, including UX assessment. However, reduced accuracy when applied to diverse, real-world UX scenarios beyond the WESAD dataset. Yildirim et al. [16] tested the proposed algorithm against two metaheuristic search methods: a Non-Dominated Sorting Genetic Algorithm-II (NSGA-II) and Cuckoo Search, for emotion detection. The presented approach for feature selection accurately classifies emotions from speech while considerably reducing the amount of characteristics needed. In speaker-dependent experiments, the EMO-DB dataset attained 87.66% and 87.20% accuracy. Meanwhile, the IEMOCAP dataset recorded accuracies of 69.30% and 68.32%. The speaker-independent trials yielded similar results from both datasets. The recognition rates for EMO-DB were 76.80%, 76.82%, and 59.37% despite reduced performance in speaker-independent scenarios, highlighting challenges in generalizing across different speakers.

Vázquez-Romero et al. [17] investigated a speech-based approach to predictive mood identification. The technique uses learning ensembles for Convolutional Neural Networks (CNN) and is tested employing files from the 2016 Audio-Visual Emotion Challenge's Depression categorization Sub-Challenge. Speech signals are pre-processed into log-spectrograms, with balanced sampling to retain both useful and irrelevant data. Multiple 1D-CNN models are trained and their outputs are combined using Collective Averaging to generate the final forecast. However, it struggles with generalization across diverse datasets or varying speaker conditions, limiting its robustness in real-world applications.

Sardari et.al [18] suggested a framework that employs a CNN-based Autoencoder approach for extracting extremely important and exclusive qualities from raw consecutive audio information, which allows for better diagnosis of depression. Furthermore, to address the data mismatch issue, they employ a cluster-based choosing strategy, which significantly decreases the possibility of bias toward a majority class (not depressed). According to the data, the suggested approach outperforms earlier well-known audio-based ADD approaches by at least 7% in the F-measure for classifying depression. May struggle with generalization to diverse, real-world audio data, leading to reduced accuracy in uncontrolled environments.

Zhu et al. [19] investigate the effectiveness of an actual-time stress detection device using biological data derived from worn sensors. Smartwatches collect several biological signals, including EDA, ECG, and Photoplethysmograph (PPG), which are then analyzed for stress categorization. In the post-acquisition phase, six methods from ML are used on a computer for categorization. SVM, KNN, RF, NB, LoR, and Stacking Ensemble Learning. Training and testing are carried out using data from two publicly available datasets. We assess the accuracy of each modality separately and as a whole. The results of the test show that when SEL is utilized for categorization, EDA has the highest accuracy. Furthermore, in both datasets, EDA outperforms other ML techniques in terms of reliability. The wearable device's EDA offers great promise for use in a stress assessment system. It may be limited by individual variability in physiological signals, affecting its accuracy across different users.

Abd Al-Alim et al. [20] proposed an ML model for stress detection in free-living environments using the SWEET dataset collected from wearable sensors (ECG, ST, SC) of 238 subjects. The study aimed to address the gap in real-world stress detection by shifting from controlled settings to spontaneous monitoring using four ML models, with KNN achieving 98% accuracy. To overcome data imbalance, the authors applied class reduction and SMOTE, demonstrating improved model stability. However, the inherent class imbalance and reduced stress label granularity may limit the ability to generalize stress levels or unseen subjects despite high accuracy.

Abdelfattah et al. [21] investigated stress detection using multi-modal physiological signals (ACC, ECG, BVP, TEMP, RESP, EMG, EDA) from the WESAD dataset, comparing seven ML and three DL models across chest and wrist data. The study showed that RNN achieved an F1 score of 93% in cross-subject evaluation, while XGBoost and RF attained 99% F1 scores in intra-subject settings.

The RNN's ability to capture temporal dependencies benefited generalization across users, while tree-based ML models performed better but risked overfitting on subject-specific patterns. A key limitation is the increased computational cost and training time of DL models compared to traditional ML.

The reviewed literature highlights advancements in emotion and stress detection systems using ML, focusing on Biological signals including EDA, ECG, and speech data. Various classifiers, including SVM, RF, and CNN, have demonstrated high accuracy in diagnosing conditions like stress and depression, with some achieving up to 97% accuracy. Despite these promising results, challenges remain, such as difficulty in generalizing across different datasets, speaker conditions, and real-world environments, as well as individual variability in physiological signals. These models illustrate the potential for real-time applications in healthcare and user experience, yet further modification is essential to overcome these difficulties.

## 3. System Methodology

The suggested methodology for stress detection using audio signals is based on a stepwise procedure, from raw audio data acquisition to classification of levels of stress through different ML models. The suggested method of stress detection from audio signals adheres to a systematic approach, with data collection starting with preliminary processing, followed by feature acquisition for identification, as depicted in Figure 1.

The process begins with the collection of audio samples, which are often gathered from reliable databases, such as the one held by the Toronto Emotion Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). These data sets are selected based on their extensive emotional tags, which are needed to accurately determine stress levels. This process is crucial in standardizing audio data by filtering out noise.

Normalizing amplitude and removing silence provides consistency and quality in the dataset. This process is important as it gets the audio ready for accurate feature extraction, which in turn has a direct effect on the accuracy of the analysis that follows. Feature extraction is concerned with extracting both fundamental and higher-level audio features so that the subtleties of speech stress are accurately captured. Major features like ZCR, RMSE, Spectral Centroid, and MFCCs are extracted to analyze the frequency, amplitude, and timbre of the signal.

These attributes are essential for differentiating stressed speech from neutral or relaxed speech, thereby filling the gap in research that effectively identifies stress using subtle vocal cues. More sophisticated spectral features, such as Spectral Bandwidth, Spectral Contrast, and Chroma features, enhance the analysis further, presenting a more precise representation of the emotional state in speech.

The classification process utilizes a variety of ML models, such as LoR, RF, GB, K-NN, NB, and different SVM kernels. The incorporation of ensemble models such as RF and Voting Classifier solves the issue of model bias and variance, which enhances the strength and accuracy of stress identification.

The significance of employing multiple classifiers is due to their ability to identify unique patterns from sound data, leading to a comprehensive analysis. This strategy addresses the gap in research in optimizing classifier performance for detecting stress, as demonstrated by Figure 1, which is typically constrained by the variation of speech patterns among individuals.

The main benefit of this approach is that it is holistic, combining various audio features with an ensemble of classifiers, and hence, the detection accuracy is increased. The use of both spectral and temporal features makes the analysis more accurate, thereby making this approach particularly useful for real-world applications such as mental health monitoring and workplace stress measurement.

Through the use of ensemble classifiers, the method proposed improves the capability of the system to generalize to different sets of data, thus addressing limitations connected to overfitting and underfitting issues inherent in conventional models. The extensive flow not only fulfils current gaps in stress detection precision but also provides a scalable approach to the deployment of stress monitoring systems in real-world situations.

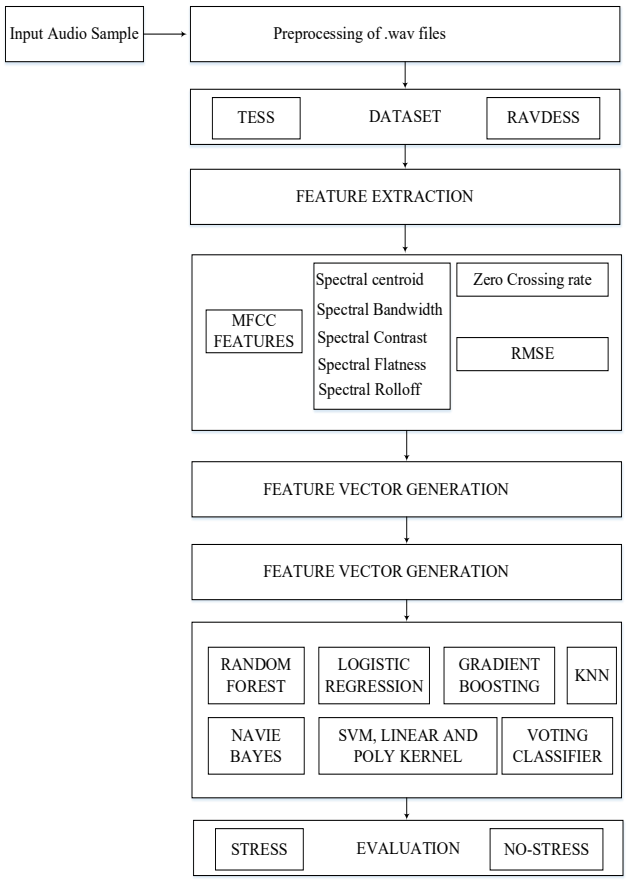



**Fig. 1 Architecture of the proposed ML-based stress detection framework**

### *3.1. Data Acquisition*

The process starts with gathering audio recordings, most often in .wav format, from popular datasets such as the TESS and the RAVDESS. These datasets are widely used in the domain of emotion detection due to the broad range of emotional states they cover, ranging from stress and anxiety to calm.

### *3.2. Pre-Processing*

Irrelevant sounds and background noise are removed by methods such as adaptive filtering. The process eliminates everything except the necessary parts of the speech signal. Noise reduction by adaptive filtering is a process of continuously reducing the background noise without degrading the necessary components of the speech. A noise profile is first estimated from low-speech or silence regions to create a baseline for the filtering process.

An adaptive filter using Least Mean Squares (LMS) is set up to modify its parameters according to the input signal. A reference noise signal from background sounds is subtracted from the main audio signal, and the filter is adjusted constantly to cancel this noise without altering the speech quality. Any remaining noise may further be suppressed with additional refinement, like spectral subtraction. This adaptive method leads to purer audio data, improves feature extraction accuracy, and improves the accuracy of ML methods in stress detection problems.

### *3.3. Feature Extraction*

Proper identification of stress from an audio signal is dependent to a large extent on the extraction of significant features that convey the affective content of speech. The approach utilizes a combination of primitive and higher-level audio features for constructing a comprehensive feature set. Primitive features like ZCR capture frequency change, while RMSE captures signal power, which is critical for detecting stress-induced amplitude variations. Spectral Flatness expresses the noisiness of the signal, and Spectral Centroid is related to the perceived brightness of sound. Spectral Bandwidth estimates frequency dispersion, providing information on speech complexity, while Spectral Contrast expresses amplitude differences between spectral valleys and peaks, indicating stressed speech versus neutral tones. Spectral Rolloff estimates the overall distribution of a wavelength and identifies the primary frequency range. Mel-Frequency Cepstral Coefficients are central to the representation of the timbral texture of speech, successfully encoding fine-grained emotional subtleties. Chroma Features examine energy distribution over pitch classes, mirroring

intonation and pitch variations associated with stress. These extracted features are incorporated into many ML models. The Voting Classifier combines predictions from various models to provide improved accuracy and reliability. This multi-dimensional feature extraction method ensures detailed analysis, allowing for trustworthy and accurate stress detection in speech.

*3.3.1. Zero Crossing Rate (ZCR)*

ZCR represents the frequency at which a signal changes sign, indicating variations in its frequency content. It measures the rate at which the signal shifts from positive to negative or vice versa, capturing rapid changes in energy. ZCR is a crucial property for recognizing short and loud sounds; it identifies minor variations in the amplitude of a signal. It is used to determine whether human speech is present in a speech sample or not. ZCR is employed in all speech processing applications, including synthesis, augmentation, and recognition. The negative crossing count represents the rate at which electricity gathers in the signal wavelength. [22]. ZCR is also used to evaluate the spectral features of speech signals, as shown by

$$\text{ZCR} - \frac{1}{N-1}\sum_{n-1}^{N-1} \mathbb{I}\{x[n]x[n-1] < 0\} \quad (1)$$

Where, $N$ specifies the total no of samples, $x[n]$ states the audio signal, and $\mathbb{I}\{\cdot\}$ is the indicator function that equals 1 if the argument is true and 0 otherwise.

*3.3.2. Root Mean Square Error (RMSE)*

Reflects the strength of the auditory indication, with amplitude fluctuations potentially indicating stress levels. (RMSE) Calculates an audio signal's power by taking the square root of its average squared amplitudes over a segment. It captures alterations in the signal's amplitude, reflecting deviations in intensity that can indicate stress. The process involves dividing the audio into short frames, squaring each amplitude value to eliminate negative effects, averaging these squares, and then returning the square root to the original amplitude scale. RMSE is particularly useful in stress detection because stressed speech often exhibits higher and more irregular energy patterns compared to calm speech, making it a reliable feature for identifying emotional states.

$$\text{RMSE} - \sqrt{\frac{1}{N}\sum_{n-1}^{N} x[n]^2} \quad (2)$$

Where:

- $x[n]$ is the audio signal.
- $N$ is the total number of samples

*3.3.3. Advanced Spectral Features*

Higher-order spectral features are characteristics generated by the amplitude domain of a sound signal that offer finer details about its spectral content. Such features, including Spectral Centroid, Spectral Bandwidth, and Spectral Contrast, give information regarding the quality of tone, texture, and frequency distribution variation, improving emotional state discrimination in stress detection.

*Spectral Centroid*

The position of the center of gravity of the wavelength is related to the brightness of the audio signal as observed. In order to calculate the Spectral Centroid, the audio input is first broken up into small structures, which are next converted to the frequency domain using a Fourier Transform (FT). The center of the spectrum is subsequently obtained by approximating the weighted average among the wavelengths in the spectrum. Each frequency is weighted by its magnitude. This means adding the product of each frequency bin and its value and dividing by the total of all values. This process helps to calculate the middle of the spectrum's weight, where the principal frequency region is. Large spectral centroid values correspond to bright sounds with a larger amount of high-frequency information, while small values correspond to darker sounds with low frequency, providing valuable information for stress classification in speech.

$$\text{Spectral Centroid} - \frac{\sum_{k-0}^{N-1} f[k]\cdot|x[k]|}{\sum_{k-0}^{N-1}|x[k]|} \quad (3)$$

Where:

- $f[k]$ is the frequency at bin $k$.
- $X[k]$ states that the magnitude of the FT is measured at bin k.

*Spectral Bandwidth*

The extent of frequencies in the audio signal gives information about the complexity of speech. The Spectral Bandwidth is a measure of the spread or extent of frequencies in an audio signal, showing how spread out the spectral energy is relative to the spectral centroid. It is calculated based on the normalized (RMS) deviation around frequency towards the spectral centroid, employing the magnitudes of each frequency bin as weights. Mathematically, it is a sum of the squared differences between every frequency f[k] and the spectral centroid, weighted by the magnitude |X[k]|, and a square root taken from it. Increasing spectral bandwidth indicates an increased spread of frequencies, which means a noisier or complex signal, whereas decreased bandwidth indicates a more tonal and centralized sound. Variations in spectral bandwidth in stress detection can mirror voice quality and intensity changes and help distinguish between stressed and neutral speech patterns.

$$\text{Spectral Bandwidth} - \sqrt{\frac{\sum_{k-0}^{N-1}|f[k]-\text{Spectral Centroid}\quad|^2\cdot|x[k]|}{\sum_{k-0}^{N-1}|x[k]|}} \quad (4)$$

Where:

- $f[k]$ And $X[k]$ are as defined above.

*Spectral Contrast*

Preserves amplitude variation between the peaks and troughs of each spectrum, enabling stressed speech to be differentiated from neutral speech. Spectral difference is a variation in magnitude between highs and lows in different spectrum bands of a stream of audio, representing changes in spectral structure. It is derived by combining the variation among each band's greatest (peak) and the lowest (valley) magnitudes. High spectral contrast indicates significant differences between frequency components, often found in stressed speech due to increased vocal tension and dynamic variation. In contrast, lower spectral contrast values suggest smoother, more uniform signals typical of calm or neutral speech. This property is essential for stress detection since it distinguishes states of mind according to the level of detail and texture of the spoken signals.

$$\text{Spectral Contrast} - \frac{1}{B}\sum_{b-1}^{B}\left(\text{Peak}_b - \text{Valley}_b\right) \quad (5)$$

Where:

- $B$ is the total number of bands of frequency.
- $\text{Peak}_b$ and $\text{Valley}_b$ are the peak and valley amplitudes in band $b$.

*Spectral Flatness*

Assesses how noise-like a signal is; stress-induced speech often shows different spectral flatness patterns compared to neutral speech. The spectrum flats are determined by multiplying the geometrical average of all the spectrum values by the mathematical mean. High spectrum flattening indicates that the spectrum characteristics are more evenly dispersed. Low spectral Flatness suggests that energy is concentrated in specific frequency components rather than being evenly spread. The spectral flatness values for spectra obtained using the SFF and ZTW methods are presented in Equation (6) [23].

$$\text{Spectral Flatness} - \frac{\left(\prod_{k-0}^{N-1}|X[k]|\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{k-0}^{N-1}|X[k]|} \quad (6)$$

*Spectral Roll off*

Describes the wavelength under which a particular percentage (e.g., 85%) of the overall spectral power lies, which aids in capturing the signal's spectral distribution. Spectral rolloff is the frequency at which a specific percentage of the overall spectral energy is concentrated, often 85%. It is calculated by summing the magnitudes of the spectral components and determining the frequency at which this cumulative sum reaches 85% of the total energy. This feature helps capture the distribution of energy across the spectrum, distinguishing between tonal and noisy sounds. Higher rolloff values indicate a greater concentration of energy in higher frequencies, often associated with stressed or excited speech, while lower values suggest energy is focused in lower frequencies, typical of neutral or calm speech. Spectral rolloff is crucial in stress detection, providing insights into vocal tension and energy distribution patterns.

$$\text{Spectral Rolloff} - f_T \text{ such that } \sum_{k-0}^{T}|X[k]| - 0.85 \times \sum_{k-0}^{N-1}|X[k]| \quad (7)$$

Where:

- $f_r$ is the rolloff frequency.
- $X[k]$ is the magnitude of the FT (Fourier Transform) at bin $k$.

*3.3.4. Mel-Frequency Cepstral Coefficients (MFCCs)*

MFCCs are commonly utilized in speech recognition because they accurately represent the timbral texture in audio sources. These coefficients are retrieved by applying a logarithmic filter bank on the signal's energy bandwidth, then performing a discrete cosine transform. Mel-Frequency Cepstral Coefficients are important characteristics in speech analysis because they capture the timbral texture of audio signals by simulating the human auditory system's sensitivity to various wavelengths. The extraction method begins with transforming the audio signal to its frequency range representation utilizing the Fourier Transform. The power spectrum is processed over a series of overlapping triangular filters spaced on the Mel scale. Logarithm of the filter bank energies SmS_mSm is then computed, emphasizing the perceptually important frequency components.

Discrete Cosine Transform is employed on logarithmic values to decorrelate the filter outputs and produce the MFCC coefficients. The mathematical formula for the n-th coefficient is given as:

$$\text{MFCC}_n - \sum_{m-1}^{M}\log S_m \cos\left[\frac{n(m-0.5)\pi}{M}\right] \quad (8)$$

Where:

- $M$ denotes the Mel filters.
- $S_m$ specifies log energy at filter $m$.
- $n$ demonstrates the MFCC coefficient index.

*3.3.5. Chroma Features*

These characteristics show the distribution of energy across distinct tone categories, which may suggest differences in intonation and pitch caused by stress. Chroma Features indicate the distribution of spectral energy over twelve unique pitch classes. (e.g., C, C#, D, etc.), capturing harmonic and melodic characteristics of an audio signal. These features are

taken from a signal's frequency domain representation, and each frequency bin is mapped to its associated pitch class, with magnitudes summed within each class. Mathematically, the Chroma value for a pitch class c is calculated as:

$$\text{Chroma Feature}_{c} - \sum_{k \in \text{pitch class } c} |X|k|| \quad (9)$$

Where:

- $c$ Represents a particular pitch class (e.g., C, C#, D, etc.).

### *3.4. Classification*

The process concludes by identifying the retrieved features to detect stress using different training algorithms. The research investigates the effectiveness of various classifiers, including:

#### *3.4.1. Random Forest*

RF is a supervised ML method frequently employed for classification and regression problems, such as stress detection from audio signals. It performs by creating multiple decision trees from data samples and then aggregating their outputs to get a final forecast via majority voting. This ensemble approach enhances accuracy and reduces the risk of overfitting compared to individual decision trees. In the context of audio feature analysis, RF starts by assigning all labelled data to a root node and randomly selecting a feature subset. Each feature is evaluated against a threshold to split the data into left and right subsets. These subsets are recursively split further, forming branches until a stopping criterion (e.g., a minimum sample size) is reached. At this point, leaf nodes are created, and everything is labelled according to the majority class of information it contains.

The method employs a technique known as "feature bagging," in which a random selection of features is chosen at each split. Such randomization increases variety across the structures, which improves the model's generalization performance. MFCCs, spectral properties, and Chroma values are evaluated over many trees to detect auditory stress. By leveraging the strength of diverse features and the ensemble voting mechanism, RF can accurately detect stress patterns in speech, even when dealing with complex and noisy audio data [24].

Algorithm: 1
Step 1: Initialize Inputs:

Given training data X (features) and y (class labels), the number of decision trees num_trees, and test data $X_{text}$

Step 2: Create an Empty Forest:

Initialize an empty list forest to store decision trees.

Step 3: Build Decision Trees:

For each tree (repeat num_trees times), generate a bootstrap sample by randomly selecting data points (with replacement) from X and y. Train a decision tree on the bootstrap sample. Add the trained tree to the forest.

Step 4: Make Predictions:

For each test data point in $X_{text}$

- Collect predictions from all decision trees in the forest.
- Aggregate the forecasts using majority voting to determine the final class label.

Step 5: Return Final Predictions:

Output the predicted class labels for all test data points based on majority voting across all trees.

#### *3.4.2. Logistic Regression*

LoR is an effective statistical technique for modeling and analyzing multivariate issues. LR analyzes the likelihood of a classification being associated with a set of explanatory elements (includes) and the link among factors and a variable that responds. In the case of M categories and N recorded earthquake qualities, the logistic model determines the probability for every category except the final one. LR helps estimate the likelihood of an outcome based on the input features, providing valuable insights into the data.

$$P(m|Z) = \frac{e^{z}}{1+\sum_{m=1}^{M-1} e^{z}} \quad (10)$$

The final group has a chance of

$$P(M) = 1 - (\sum_{m=1}^{M-1} P(m|Z)) = \frac{1}{1+\sum_{m=1}^{M-1} e^{z}} \quad (11)$$

While P (m|Z) determines the category responses to factors xi, which represents the probability of a given result, m, Z is a logistic model-computed amount of observable predictor elements xi to classification m. The logistical framework is a weighted average of a set of factors that explain, given as:

$$Z = \sum_{i=1}^{N} \beta i xi + \beta o, \quad (12)$$

Whereas β0 is a constant (capture) and βi are the predictor variables (regression coefficients) determined using the highest probability approach. In this scenario, the result factors are the occurrence depth categories (deep or shallow), and P(m|Z) represents the likelihood of occurrence as a very deep event, depending on the observed characteristics (xi).

The probability of a deep event is influenced by the term βixi. An event characterized by seismic attribute xi is classified as a deep event, P (deep|Z) is bigger than P (shallow|Z) (Amidan&Hagedom 1998).

Algorithm: 2
Step 1: Initialize Weights and Bias

Set initial values for weights w (usually small random values) and bias b (usually set to 0 or a small random value).
Step 2: For each epoch, Calculate Predictions

- Compute the predicted output using the sigmoid function

$$\hat{y} = Sigmoid\,(X.\omega + b) \quad (13)$$

- Let X be the input features, w be the weight vector, and b be the bias term.

Step 3: Compute the Loss

- Compute the binary cross-entropy loss between the actual labels y and the predicted labels. $\hat{Y}$

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}(\,y_i \log \hat{y}_i) + (1 - y_i)\log\,(1 - \hat{y}_i)) \quad (14)$$

- Let N be the number of samples, $y_i$ is the actual label, and $\hat{y}_i$ is the predicted label.

Step 4: Update Weights and Bias

- Compute the gradients of the loss function with respect to weights and bias.

$$\frac{\partial Loss}{\partial w} = \frac{1}{N}X^T(\hat{y}\text{-}y) \quad (15)$$

$$\frac{\partial Loss}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}(\,\hat{y}_i\ \text{-}y_i) \quad (16)$$

- Update weights and bias using gradient descent

$$w = w - \propto . \frac{\partial Loss}{\partial b} \quad (17)$$

$$b = b - \alpha . \frac{\partial Loss}{\partial b} \quad (18)$$

Where α\alphaα is the learning rate (step size).

Step 5: Return Weights and Bias

After training, return the optimized weight vector, w and bias b.

*3.4.3. Gradient Boosting:*

The enhancement of gradients is an example of team learning. Ensemble education, which combines weak learners to create strong learners, differs from traditional approaches. Unlike the bagging strategy, which creates models independently, the ensemble boosting technique creates models sequentially by repeatedly decreasing errors from previously learned models. The model predicts results by integrating M additive tree models (f0, f1, and f2… FM) Equation (19).

$$\text{f (x)} = \sum_{m=0}^{M} f_{m(x)} \quad (19)$$

To optimize the tree ensemble model, the anticipated generalization error (L) is reduced using Equation (20).

$$L = \sum_{i}^{n}(y_{i\_}\hat{y}_i)^2 \quad (20)$$

The loss function L calculates the delta loss between a data point's target ($y_i$) and prediction $\hat{y}_i$ [25].

Algorithm: 3
Step 1: Initialize Inputs

Given training data X (features), (target values), and the number of weak learners num_estimators.

Step 2: Initialize the Model

Start with an empty list model to store decision trees (weak learners).

Step 3: Iterative Boosting Process

For each estimator (repeat num_estimators times):

- Compute the residuals (errors) by subtracting the current predictions from the true target values y.
- Train a new Decision Tree on the residuals to model the errors.
- Add the trained decision tree to the model.

Step 4: Make Predictions

For each test data point in $X_{text}$

- Initialize the predicted value as zero.
- Sum the predictions from each tree in the model.

Step 5: Return Final Predictions:

Output the final aggregated predictions by summing the contributions from all weak learners.

*3.4.4. K-Nearest Neighbors (K-NN)*

The KNN divides information into groups according to the distance between its features. When the distance between data points is small, a group is produced. When the distance is large, multiple groups are formed. KNN is a popular classifier for categorizing EEG signals in research.

The KNN is a non-parametric classification approach that compares test and training data. The majority vote of each object's KNN determines its classification, placing the item in the most common class (where k is a positive number).

The optimal match among the training and testing data was found using a range of k values. The item is assigned to the KNN class if k = 1. The widely used Euclidean distance metric has the following definition:

Xi and Xj represent the starting point evaluation and training data. In this study, a value of k was selected to oscillate between two. This k value outperforms all others in terms of categorization performance.

$$d(X_i X_j) = \sqrt{\sum_i (X_i - X_j)} \quad (21)$$

Algorithm: 4
Step 1: Initialize Inputs
- Input training data $X_{train}$ (feature set) and $y_{train}$ (corresponding labels).
- Input test data $X_{train}$ (data points to classify).
- Define k, the number of nearest Neighbors to consider.

Step 2: For each test point in $X_{test}$ for Calculate Distance
- Determine the distance from each test point to each of the other points. In $X_{train}$
- Common distance metrics include Euclidean distance, which is used for the 3 equation.
- Where $X_i$ and $X_j$ are data points, and D is the number of features.

Step 3: Find Nearest Neighbors
- Identify the k training points with the shortest distance to the test point.
- Store these k nearest Neighbors and their corresponding labels.

Step 4: Perform Majority Vote
- Calculate the number of instances of every category among the k closest relatives.
- Determine the class label with the highest vote (majority class)

Step 5: Assign Predicted Class
- Assign the group with the most votes to take the test point.

*3.4.5. Naïve Bayes*

NB classifier assumes independent feature values for a given class variable. Dependence between qualities is generally determined by chance. The category having the highest posterior likelihood is allocated to the information point [26]. In order to classify several classes, each group's chance must be evaluated, and the class with the greatest probability is selected as the final forecast result. Given an N-dimensional vector of attributes $x$, $P(x|y)$), the probability of class $y$, may be calculated as

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \quad (22)$$

The vector of features x is assigned to the category having the highest posterior value (MAP), which may be computed by

$$y_{MAP} = arg \max_{y \epsilon Y} P(y|x) \quad (23)$$

$$y_{MAP} = arg \max_{y \epsilon Y} \frac{P(x|y)P(y)}{P(x)} \quad (24)$$

$$y_{MAP} = arg \max_{y \epsilon Y} P(x|y)P(y) \quad (25)$$

The probability P (x|y) is given by

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x-\mu_y)^2}{2\sigma_y^2}} \quad (26)$$

Where μy and σ_y^2 are the average and the variance of every characteristic related to class y, respectively, p (y) is computed by multiplying the probability of every category in the collection of data by the total number of vectors with features. NB's train and run time complexities are O (Q^N.Y) and O (N.Y), where Q denotes the feature vectors (data points) and Y specifies the total classes [27]

Algorithm: 5
Training Phase:
Step 1: Input X (features) and y (class labels).
Step 2: Calculate prior probabilities P(c) for each class.
Step 3: Calculate likelihoods $P(X_f|c)$ for each feature.
Step 4: Return prior_probabilities and likelihoods

Prediction Phase:
Step 1: Input $X_{test}$ , prior probabilities, and likelihoods.
Step 2: For each test point:
Step 3: Calculate posterior probabilities for all classes.
Step 4: Choose the group that has the highest posterior likelihood.
Step 5: Return a list of projected class labels.

*3.4.6. Voting Classifier*

In this study, an ensemble technique was employed by generating multiple weak classifiers and combining their outputs using a Voting Classifier [28]. Specifically, the hard voting approach was applied, where each individual ML model independently classifies each instance in the dataset.

Each model acts as a "vote" to select the projected class identification. The majority class label determines the ultimate forecast for a specific instance, which is the class that receives the votes from all approaches.

This ensemble technique enhances the overall precision and robustness of recognizing stress in audio signals by leveraging the strengths of various models, mitigating individual biases, and compensating for their weaknesses. This approach enhances reliability, especially when dealing with complex audio features such as MFCCs, spectral characteristics, and Chroma features, ensuring a more consistent and accurate detection of stress in speech.

Algorithm: 6
Step 1: Input:
- Models: A list of trained models.

- X_test: Test dataset for prediction.

Step 2: Initialize:

Create an empty list prediction to store each model's prediction.

Step 3: For each model in models:

- Use the model to predict the output for X_test.
- Append the predicted output to the predictions list.

Step 4: Aggregate Predictions:

- Perform a majority vote on the predictions.
- Choose the class label that has the most votes for each test instance.

Step 5: Return:

The final predicted class labels are based on majority voting.

*3.4.7. Support Vector Machines (SVM)*

SVM constitutes a nonlinear ML method [26]. Support vectors are features from altered categories with a small separation among each other in an N-dimensional feature space. These are used to create a hyperplane defined as

$$W^{\gamma} x + b = 0, \quad (27)$$

Two more hyperplanes that lie on the support vectors and split the two classes are given by

$$W^{\gamma} x + b = 1, and\ W^{\gamma} x + b = -1, \quad (28)$$

The two hyperplanes are created by addressing an optimization problem that maximizes the distance between them, with a difference of $2/\|W\|$. To effectively deal with nonlinear data, SVM may employ a kernel known as the Radial Basis Function (RBF). SVM can be used to perform binary classification and classification of multiple classes. Provides the complete institutionalization of SVM. SVM has training and run time complexities of O (Q.R) and O (Q.U), where U specifies support vectors. RBF-SVM has train and run time complexity of O(Q^2.N+Q^3) and O(U.N), accordingly

*Linear Kernel*

A Linear Kernel in SVM is applied to stress detection in audio signals when features exhibit linear separability. It computes the dot product between feature vectors, establishing a linear decision boundary. This Kernel effectively differentiates stressed and neutral speech by capturing key features like spectral energy and pitch. Its computational efficiency makes it ideal for simpler classification tasks, ensuring clear distinctions in emotional states within audio data.

*Polynomial Kernel*

A Polynomial Kernel in SVM is used to detect nonlinear patterns in stress identification from audio signals. It transforms input features, like spectral and temporal properties, into a higher space so that the model can identify complicated patterns in speech. It works best when indicators of stress are weak or non-linearly distributed. The analysis of interaction among features enhances classification accuracy in detecting subtle emotional changes in audio data.

*Radial Basis Function (RBF) Kernel*

The RBF Kernel in SVM was very effective in capturing complex, nonlinear patterns in stress detection in audio recordings. It transforms audio features, such as spectral and frequency features, into an infinite-dimensional space, allowing the model to identify subtle emotional variations. The Kernel is especially helpful when indicators of stress are subtle and non-linearly spread out. Its adaptability ensures accurate discrimination between stressed and neutral speech, resulting in overall classification efficiency.

## 4. Result and Discussion

Each classifier's performance is measured using metrics including precision, recall, F1-score, and a confusion matrix. After training and assessing numerous models, a comparison study is conducted to select a particularly effective classifier for stress detection. The ensemble models, particularly RF and Voting Classifier, are expected to show superior performance due to their ability to reduce variance and bias.

### *4.1. Dataset Description*

The analysis of two datasets, RAVDESS and TESS, reveals that RAVDESS captures nuanced emotional features, while TESS provides clearer, more distinguishable data for stress detection. For consistency in evaluation, the emotional labels from both datasets were mapped into two stress-related classes, forming a binary-class classification problem:

Class 0 - No Stress
Class 1 - Stress

This mapping ensures that each audio sample is categorized based on its likely psychological stress implication. This study uses these class labels (0 and 1) consistently across all figures, confusion matrices, and evaluations.

*4.1.1. RAVDESS*

The RAVDESS provides a validated, comprehensive dataset for emotional speech. The words, like the song, express tranquillity, happiness, sadness, rage, fear, astonishment, and disdain. Every movement is produced with a range of mental strength and an expression that is neutral. All conditions are presented in a voice-only format. The 7356 recordings were evaluated ten times for emotional validity, intensity, and genuineness. A further 72 participants provided

test-retest results. There were reports of strong emotion accuracy and reliability between tests. Researchers can choose stimuli based on corrected accuracy and composite "goodness" evaluations. Refer to Table 1 for comparison.

*4.1.2. TESS*

It provides extremely high-quality audio samples. A significant portion of the other sample consists of male speakers, resulting in a relatively lopsided sample. As a result, this collection of data would make a great training set for the emotion classifier in terms of standardization.

**Table 1. Class description**

| Files used | RAVDESS | TESS |
|---|---|---|
| Angry | 192 | 400 |
| Sad | 192 | 400 |
| Neutral | 96s | 400 |
| Happy | 192 | 400 |
| Total | 672 | 1600 |

A total of 200 objective sentences were spoken in the speaker's phrase, and each emotion was recorded (anger, disgust, fear, joy, pleasant surprise, sorrow, and neutral). There are 2800 data points (audio files) in total. The collection is organized so that every reaction is saved in its own folder. It contains all 200 target word audio files. See Table 1 for more details.

***4.2. Performance Evaluation of Various ML Classifiers***

The effectiveness of different ML classifiers, including RF, LoR, GB, KNN, NB, and SVMs with different kernels, along with a Voting Classifier. The TESS dataset showed consistently high accuracy, especially with KNN and Voting Classifier achieving perfect results.

In contrast, the RAVDESS dataset presented more challenges due to its nuanced features, with SVM and LoR performing better. NB exhibited the lowest accuracy, highlighting the importance of feature dependencies.

Overall, classifier performance varied significantly based on dataset complexity and model robustness.

*4.2.1. Confusion Matrix*

The confusion matrix offers a detailed view of each classifier's performance by illustrating true and false predictions across classes. It enables a clearer understanding of class-wise accuracy, especially in distinguishing between stress and no-stress states across both datasets.

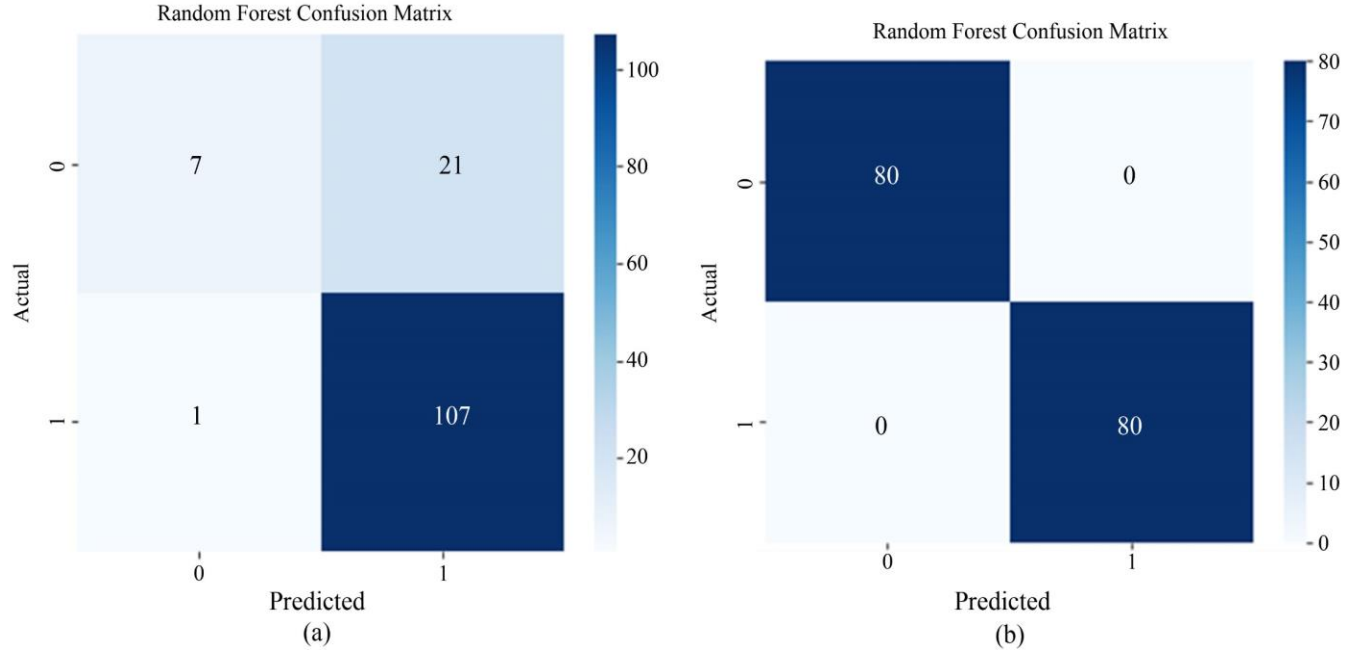



**Fig. 2 Confusion matrices obtained for RF using: (a) RAVDESS data set, and (b) TESS dataset.**

As shown in Figure 2, the RF model achieved flawless classification on the TESS dataset, correctly classifying all 80 examples of both stressed (Class 1) and unstressed (Class 0) conditions. It predicted 107 stress samples and 7 no-stress samples correctly on the RAVDESS dataset, but mispredicted 21 no-stress samples as stress and 1 stress sample as no stress. These findings indicate the model's excellent performance on clean data (TESS) and average confusion in Recognizing Subtle Stress Patterns In Complex Sounds (RAVDESS). As shown in Figure 3, the LoR was assessed employing a binary classification setup. The RAVDESS dataset correctly predicted 16 no-stress instances and 104 stress instances but misclassified 12 no-stress samples as stress and 4 stress samples as no stress, suggesting mild confusion due to

emotional overlap. On the TESS dataset, it achieved perfect classification, accurately labeling all 160 samples, demonstrating its strong generalization on clearer emotional patterns.

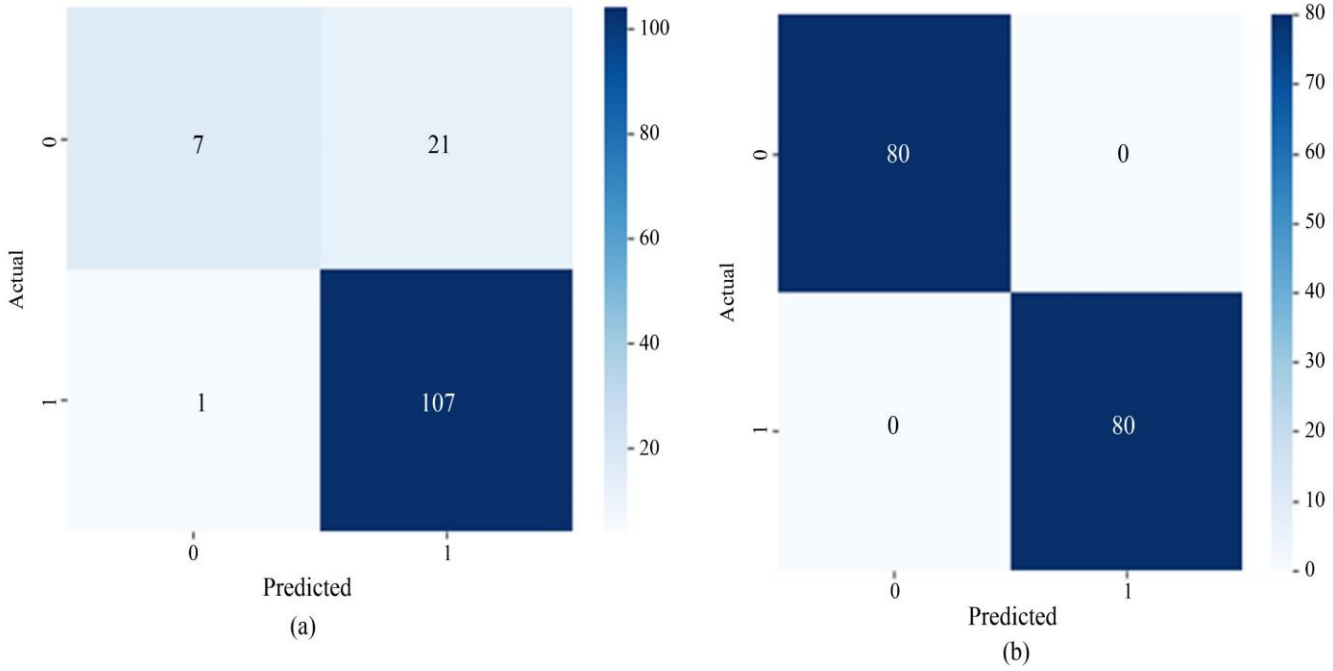



**Fig. 3 Confusion matrices obtained for logistic regression using: (a) RAVDESS data set, and (b) TESS dataset.**

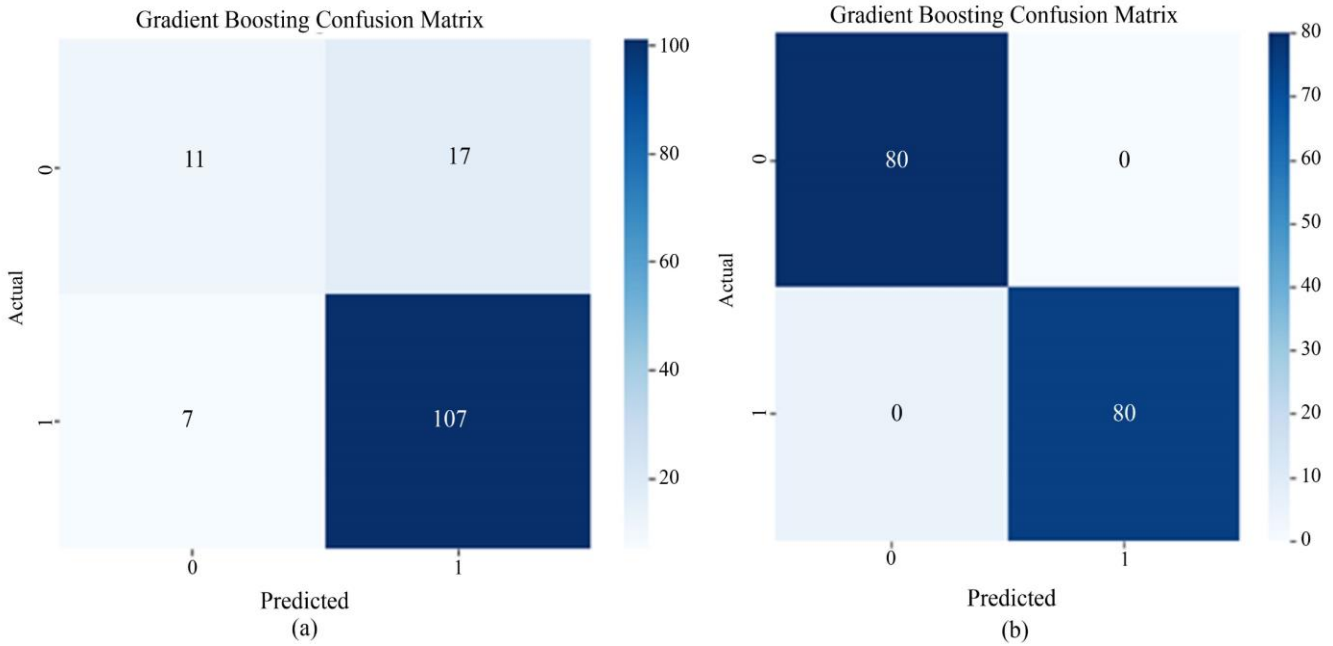



**Fig. 4 Confusion matrices obtained for gradient boosting using: (a) RAVDESS data set, and (b) TESS dataset.**

As shown in Figure 4, the GB model correctly predicted 101 stress instances in the RAVDESS dataset but struggled to distinguish between stress and no-stress samples, with some no-stress samples being misclassified.

The TESS dataset accurately predicted all 80 no-stress instances but misclassified 5 stress instances, indicating a minor limitation in detecting certain stress patterns. As shown in Figure 5, the KNN model on the RAVDESS dataset showed confusion with only 4 correctly predicted instances of "no stress" and 24 misclassified as "stress." However, it showed

strong performance in detecting stress, with 107 instances correctly predicted and 1 misclassified. On the TESS dataset, the model accurately classified all 80 instances as “no stress” and “stress” without misclassifications.

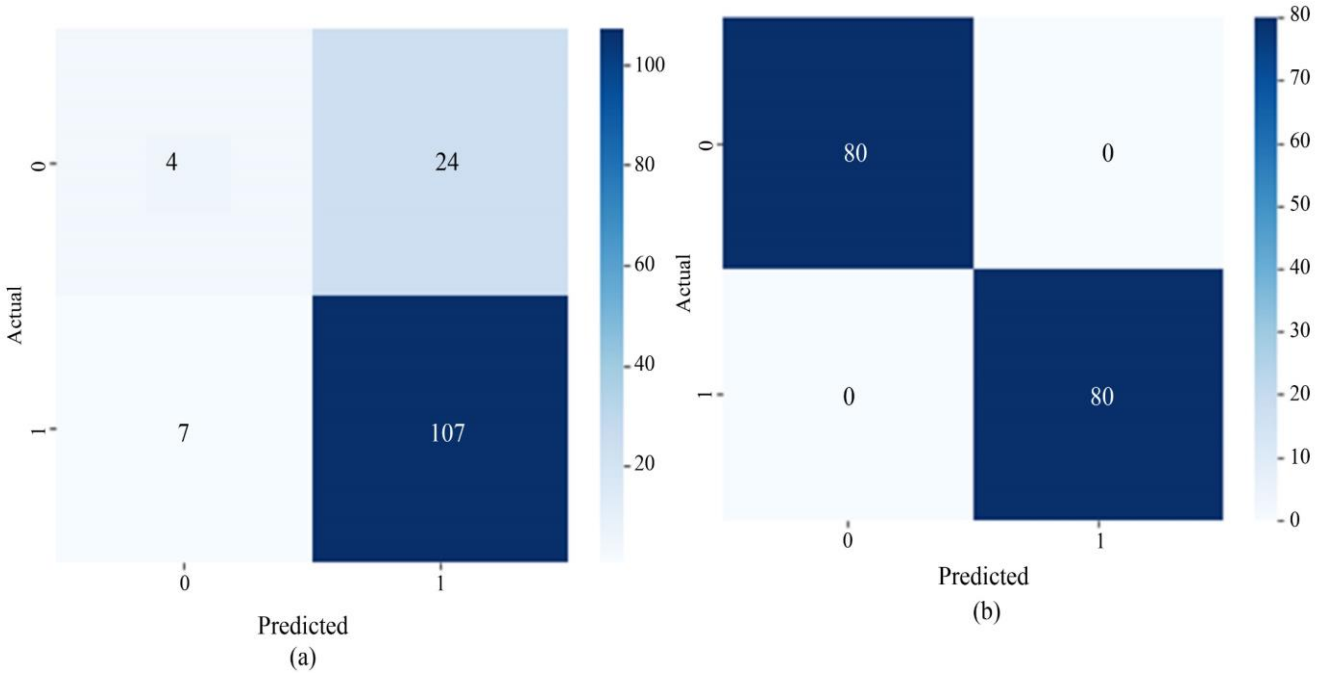



**Fig. 5 Confusion matrices obtained for K-Nearest Neighbors using: (a) RAVDESS data set, and (b) TESS dataset.**

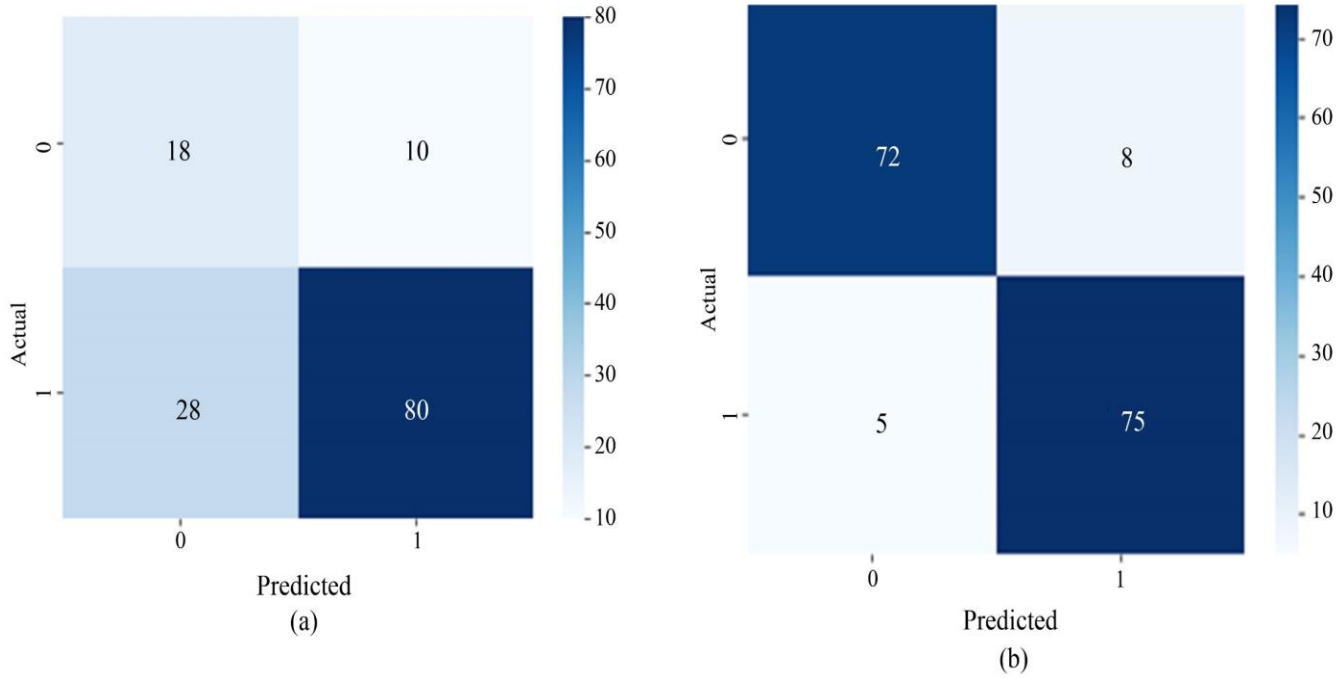



**Fig. 6 Confusion matrices obtained for NB using: (a) RAVDESS data set, and (b) TESS dataset.**

As shown in Figure 6, the NB model on the RAVDESS and TESS datasets struggled to distinguish between “no stresses” and “stress,” with 28 instances of false negatives, suggesting potential bias towards predicting “stress” due to overlapping features or data distribution limitations.

The model also showed misclassification in Class 0 and Class 1, indicating potential issues with distinguishing between stress and no-stress signals. As shown in Figure 7, the SVM with Linear Kernel model achieved 100% accuracy on the TESS dataset, correctly classifying 16 instances as “no

stress" and 105 instances as "stress," despite some misclassifications in the "no stress" class. However, some misclassifications in the "no stress" class suggest potential challenges in distinguishing between stress and no-stress signals.

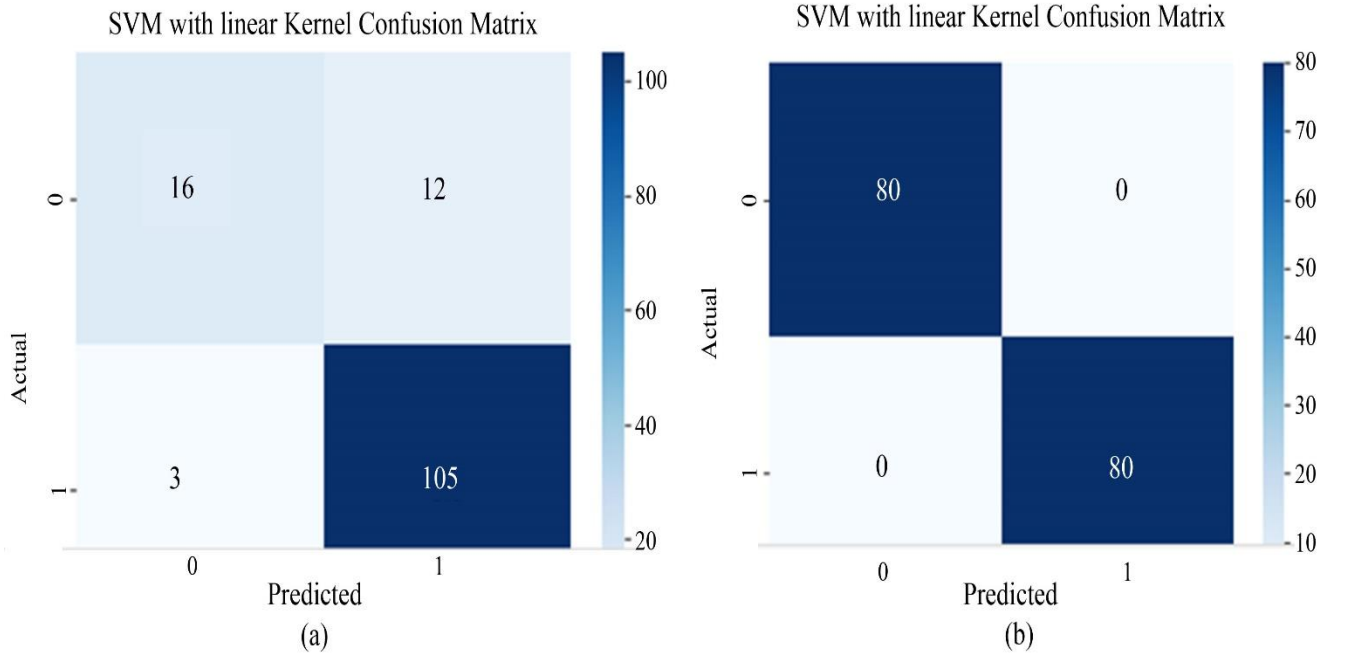



**Fig. 7 Confusion matrices obtained for SVM with linear Kernel using: (a) RAVDESS data set, and (b) TESS dataset.**

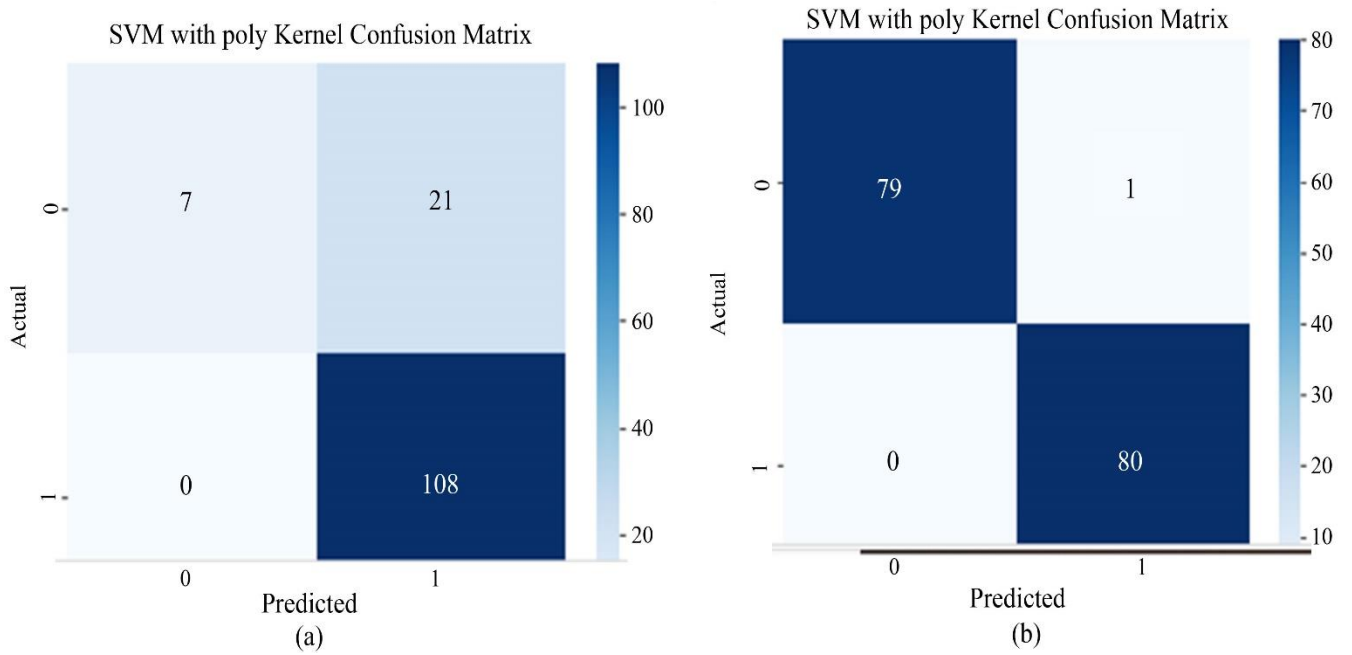



**Fig 8: Confusion matrices obtained for SVM with poly Kernel using: (a) RAVDESS data set, and (b) TESS dataset.**

As shown in Figure 8, the RAVDESS and TESS datasets show strong performance in detecting stress, with 7 instances correctly predicted and 108 correctly classified, respectively.

However, difficulty in distinguishing no-stress instances may be due to class imbalance or feature overlap. The confusion matrix for the RAVDESS dataset shows 79 instances correctly classified as no stress, while all 80 instances were correctly classified as stress. As shown in Figure 9, the SVM with RBF Kernel achieved perfect classification on the TESS dataset, correctly identifying all 80 samples of No Stress (Class 0) and all 80 samples of Stress

(Class 1) without any misclassifications. This result demonstrates the exceptional capability to distinguish between stress and no-stress emotional speech when the dataset is clean, well-balanced, and acoustically distinct. The RBF kernel's nonlinear decision boundary likely contributed to capturing the subtle emotional patterns effectively.

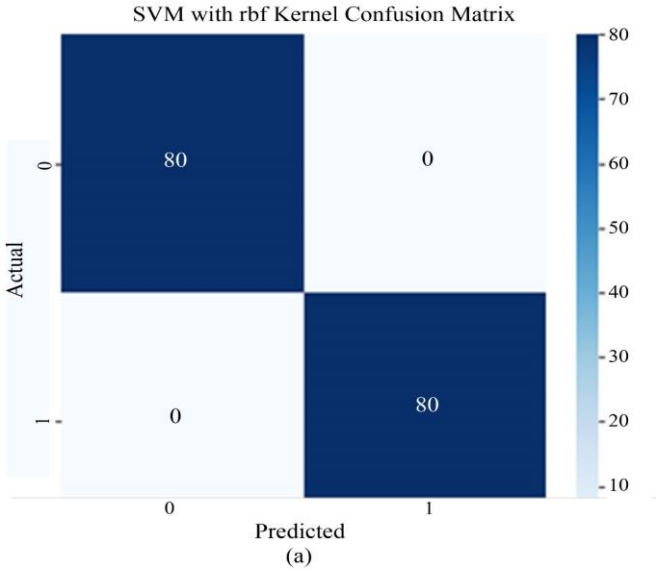



**Fig. 9 Confusion matrices obtained for SVM with poly Kernel using: (a) TESS dataset**

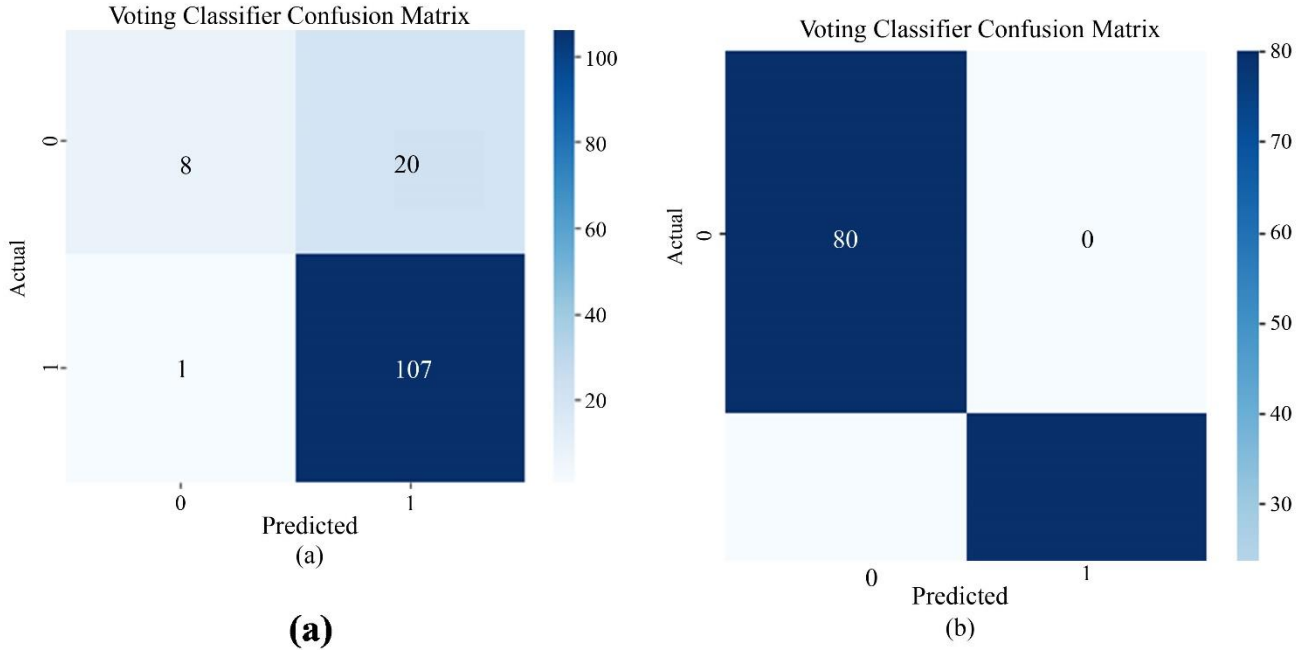



**Fig. 10 Confusion matrices obtained for voting classifier using (a) RAVDESS data set, and (b) TESS dataset.**

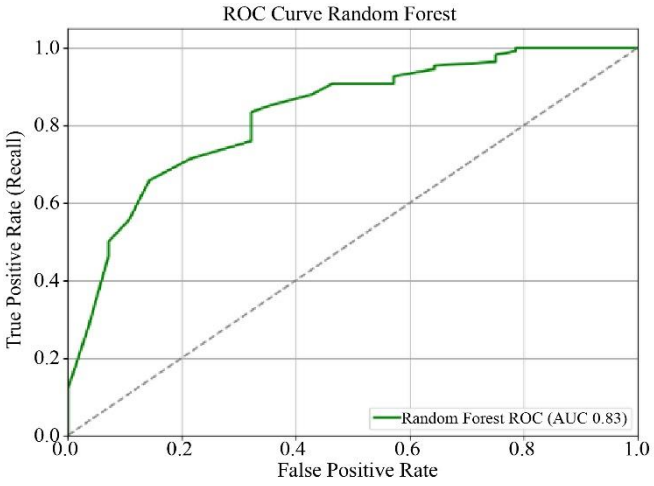
As shown in Figure 10, the model correctly classified 107 instances of Stress (Class 1) and 8 instances of No Stress (Class 0) on the RAVDESS dataset.

However, it misclassified 20 No Stress instances as stress and 1 Stress instance as No Stress, indicating difficulty in distinguishing between stress and non-stress patterns in more emotionally complex or overlapping data like RAVDESS. These results emphasize the Voting Classifier's robustness on well-structured datasets and highlight the challenges of generalizing across nuanced emotional datasets without misclassification.

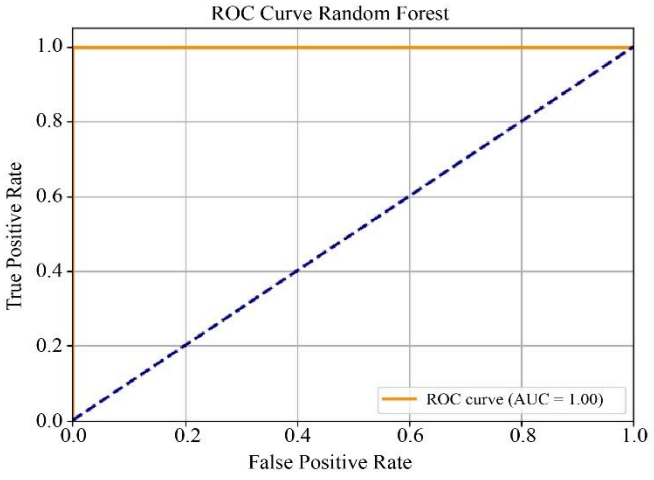
*4.2.2. ROC Curve*

The ROC curve is used to evaluate the trade-off between sensitivity (true positive rate) and specificity (false positive

rate) for each classifier. It provides a visual measure of classification performance, with the AUC indicating the model's ability to distinguish between stress and no-stress classes. As shown in Figure 11, the ROC curve of the Random Forest model on the RAVDESS database Figure 11(a) indicates an AUC of 0.83, demonstrating good but not ideal class discrimination as a result of overlapping emotions. On the other hand, the TESS database Figure 11(b) had a perfect AUC of 1.00, indicating error-free discrimination between stress and non-stress classes. This emphasizes that RF does extremely well with clean, well-segregated emotional data, but has moderate difficulties with intricate signals.

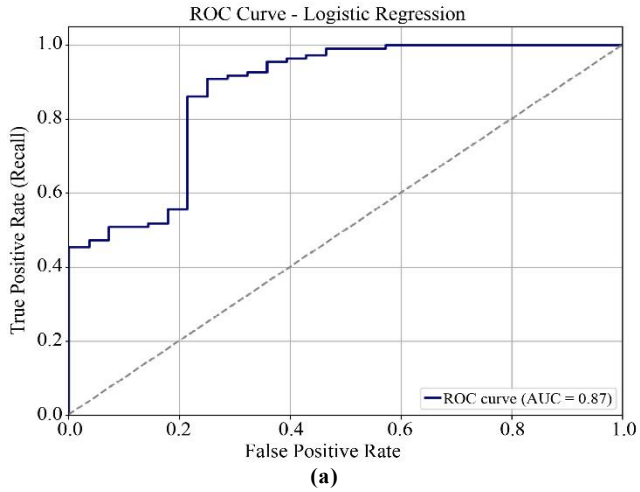



**(a)**

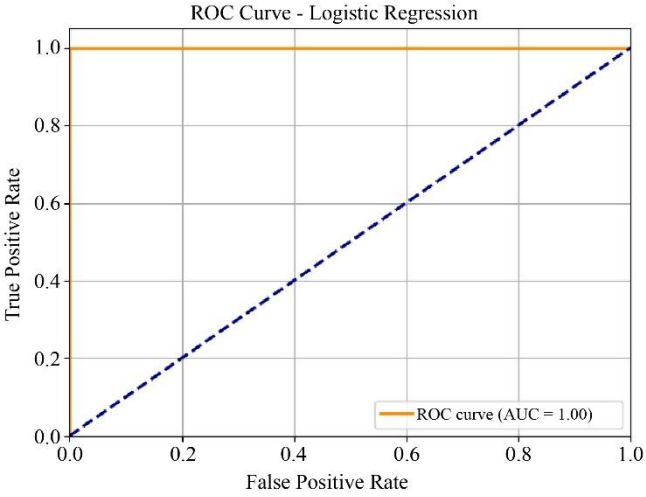



**(b)**

**Fig. 11 ROC curve obtained for Random forest using: (a) RAVDESS data set, and (b) TESS dataset.**




**(a)**




**(b)**

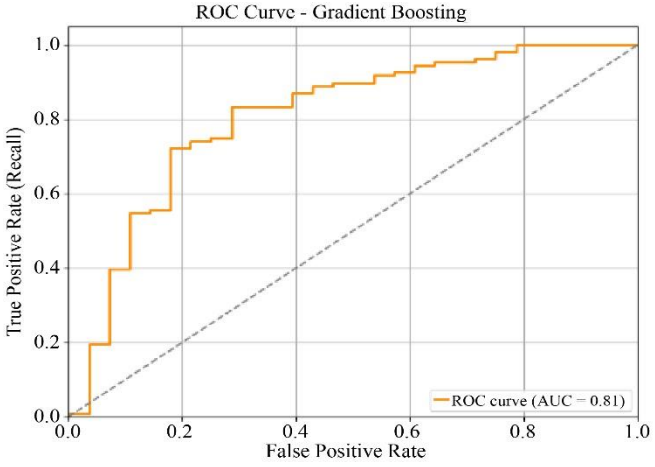
**Fig. 12 ROC curve obtained for Logistic Regression using: (a) RAVDESS data set, and (b) TESS dataset.**

As expressed in Figure 12, the ROC curve for LoR registers an AUC of 0.87 on the RAVDESS dataset, pointing to excellent but not ideal class separation. The TESS data yielded an AUC of 1.00, indicating ideal classification performance with zero false positives and false negatives.

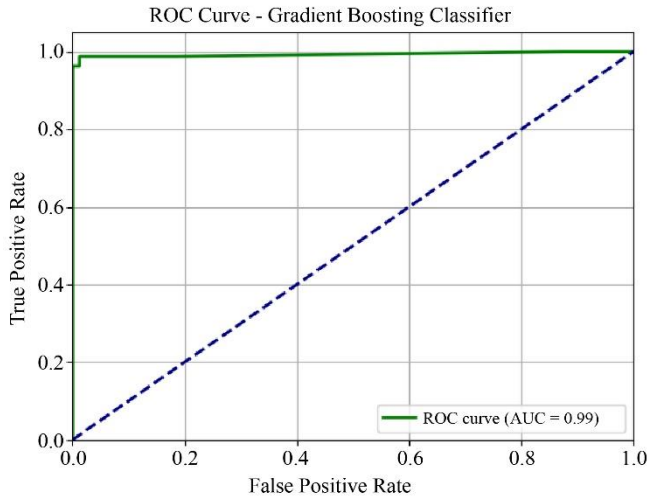
As shown in Figure 13, the GB classifier achieved an AUC of 0.81 on the RAVDESS dataset, indicating good classification with some class overlap.

On the TESS dataset, it was at 0.99, reflecting nearly perfect separation between stress and no-stress classes. As shown in Figure 14, the ROC curve for KNN on the RAVDESS dataset has an AUC of 0.81, which is a good performance with some class overlap.

On the TESS dataset, KNN has a perfect AUC of 1.00, which proves that it can perfectly differentiate stress and no-stress classes in well-defined emotional speech data.

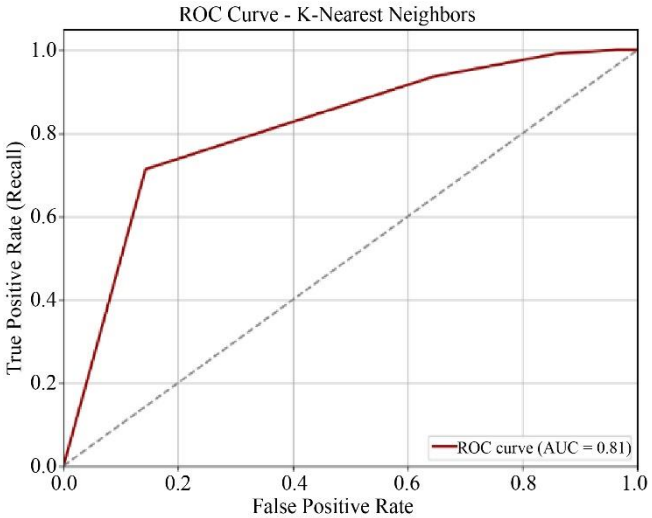



(a)

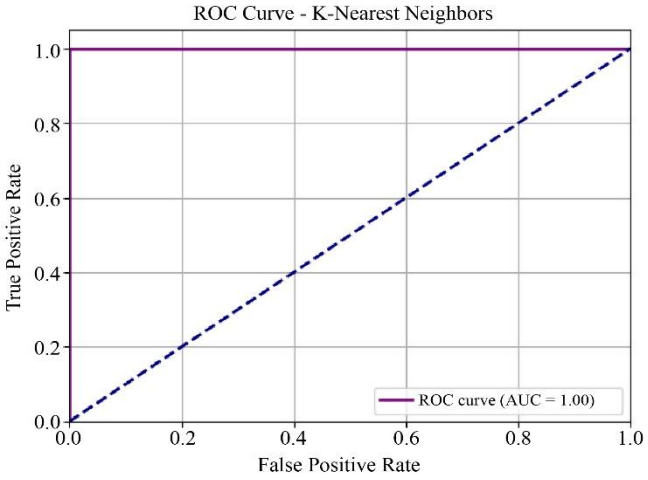



(b)

**Fig. 13 ROC curve obtained for Gradient Boosting using: (a) RAVDESS data set, and (b) TESS dataset.**

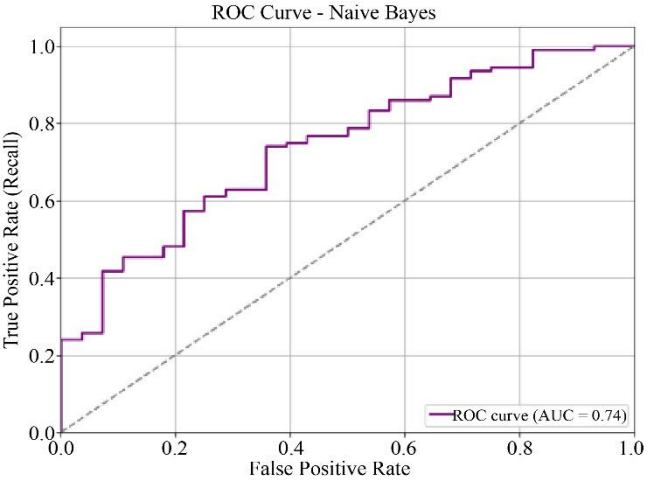



(a)

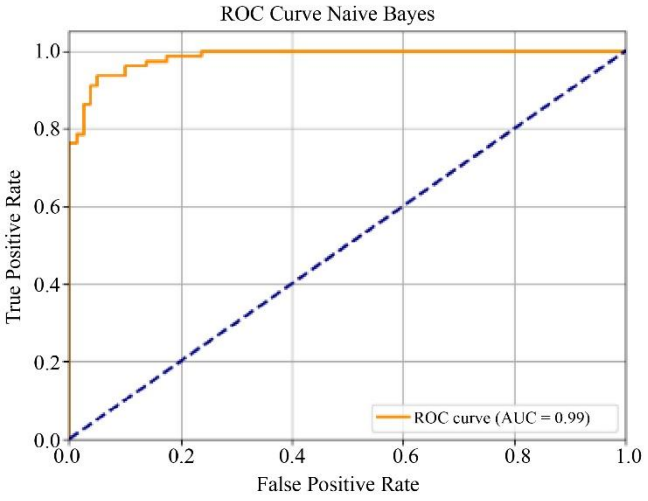



(b)

**Fig. 14 ROC curve obtained for KNN using: (a) RAVDESS data set, and (b) TESS dataset.**




(a)

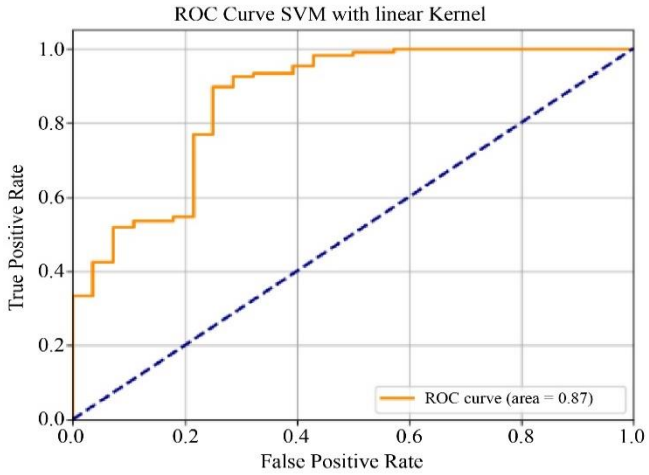



(b)

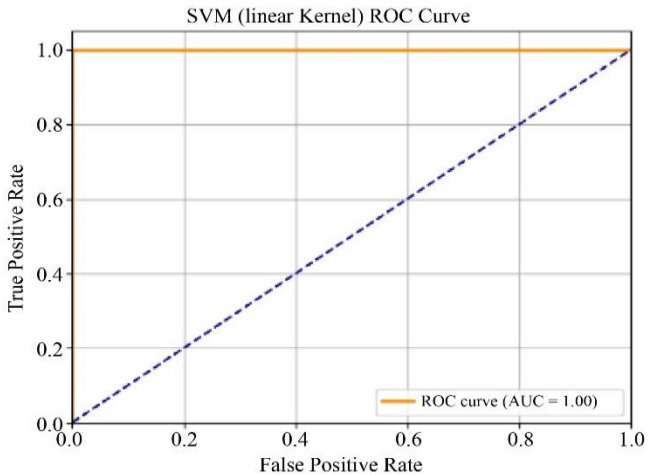
**Fig. 15 ROC curve obtained for Naïve Bayes using: (a) RAVDESS data set, and (b) TESS dataset.**

As shown in Figure 15, in the RAVDESS dataset, Naïve Bayes achieved a moderate AUC of 0.74, indicating a limited capability to distinguish between stress and no-stress conditions due to overlapping emotional cues. Conversely, the model did very well on the TESS dataset at an AUC of 0.99 with near-perfect classification in a more differentiated emotional setting.

As shown in Figure 16, the ROC curve of SVM with a linear kernel on the RAVDESS dataset has an AUC of 0.87, indicating good classification performance with some misclassifications due to overlapping features. On the TESS dataset, it had a spotless AUC of 1.00 and exhibited perfect discrimination between stress and no-stress cases. As shown in Figure 17, the SVM with a polynomial Kernel achieved an AUC of 0.85 on the RAVDESS dataset, indicating good classification despite some false positives. On the TESS dataset, it attained a perfect AUC of 1.00, reflecting excellent class separation between stress and no-stress samples.

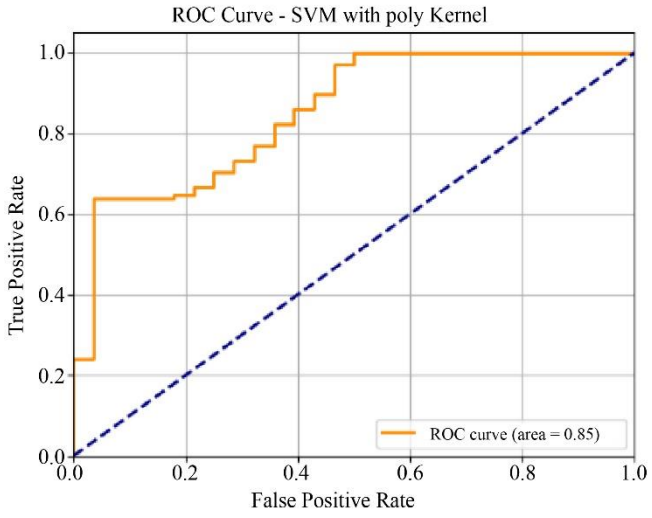



(a)

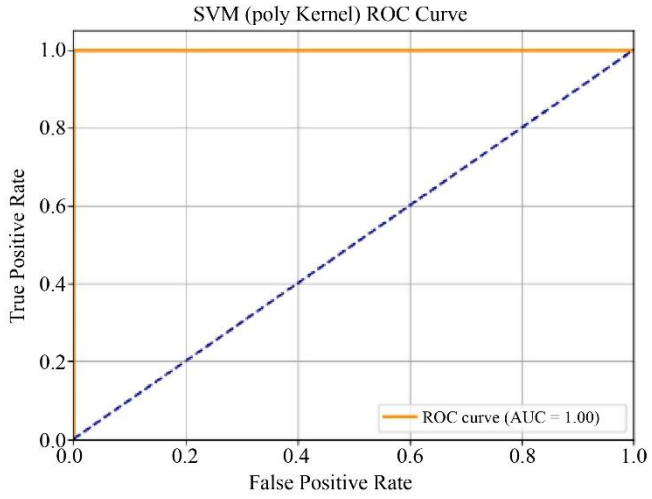



(b)

**Fig. 16 ROC curve obtained for SVM with linear Kernel using: (a) RAVDESS data set, and (b)TESS dataset.**




(a)




(b)

**Fig. 17 ROC curve obtained for SVM with poly Kernel using: (a) RAVDESS data set, and (b)TESS dataset.**

### *4.3. Comparison Metrics*

To measure the effectiveness of the proposed models, a comprehensive comparison was conducted across multiple classifiers on two benchmark datasets, such as TESS and RAVDESS.

**Table 2. Comparison with existing models**

| Reference | Dataset(s) Used | Techniques | Accuracy |
|---|---|---|---|
| [14] | DAIC-WOZ | RNN with MFCC | 95.6% |
| [15] | WESAD | DL + ANN | 97.4% |
| [16] | EMO-DB, IEMOCAP | NSGA-II, Cuckoo + SVM | 87.66% |
| Proposed | TESS | Voting Classifier | 100% |
| Proposed | RAVDESS | SVM (Linear) | 88.97% |

Table 2 offers a comparative evaluation of the proposed approach with notable recent studies in stress and emotion recognition.

While [16] employed a metaheuristic feature selection approach and achieved respectable accuracy, especially in speaker-dependent setups, their models demonstrated significant drops in speaker-independent settings.

In contrast, the proposed approach maintains robust performance across two datasets, TESS and RAVDESS, without relying on DL or physiological signals.

Notably, the ensemble model using only audio-based features achieved 100% accuracy on TESS and 88.97% on the more complex RAVDESS dataset, outperforming several prior works.

**Table 3. Comparison of algorithm**

| Models | Accuracy (%) | | Precision (%) | | Recall (%) | | F1-score | |
|---|---|---|---|---|---|---|---|---|
| | RAVDESS | TESS | RAVDESS | TESS | RAVDESS | TESS | RAVDESS | TESS |
| RF | 83.82 | 99.38 | 0.83 | 1 | 0.99 | 0.98 | 0.90 | 1 |
| Logistic Regression | 88.24 | 83.09 | 0.89 | 0.85 | 0.96 | 0.93 | 0.96 | 1 |
| Gradient Boosting | 82.35 | 96.88 | 0.85 | 1 | 0.93 | 0.93 | 0.89 | 0.96 |
| K-NN | 81.62 | 100 | 0.81 | 1 | 0.99 | 1 | 0.89 | 1 |
| Naïve Bayes | 72.06 | 91.88 | 0.88 | 0.90 | 0.74 | 0.93 | 0.80 | 0.92 |
| SVM with Linear Kernel | 88.97 | 100 | 0.89 | 1 | 0.97 | 1 | 0.93 | 1 |
| SVM with Poly Kernel | 84.56 | 99.38 | 0.83 | 0.98 | 1 | 1 | 0.91 | 1 |
| SVM with rbf Kernel | 83.09 | 100 | 0.82 | 1 | 1 | 1 | 0.91 | 0.99 |
| Voting Classifier | 84.56 | 100 | 0.84 | 1 | 0.99 | 1 | 0.91 | 1 |

The outcome of the evaluation of different ML algorithms utilizing the TESS and RAVDESS datasets gives useful information about their efficiency in detecting stress using audio features.

*4.3.1. Comparative Performance based on Accuracy*

As shown in Figure 18, the TESS dataset always reported higher accuracy values for all the classifiers than the RAVDESS dataset. Some models, like SVM with Linear Kernel, SVM with RBF Kernel, K-NN, and Voting Classifier, attained 100% accuracy on TESS. This indicates that TESS can possibly have more unique features or lower variability, which might make it easier for models to classify stress accurately. On the other hand, the RAVDESS dataset was more difficult to distinguish, with the best accuracy of 88.97% attained by SVM Linear Kernel. The comparatively lower accuracy on RAVDESS suggests that perhaps there's more complexity or noise in the dataset, and more advanced techniques or pre-processing might be necessary to attain similar performance.

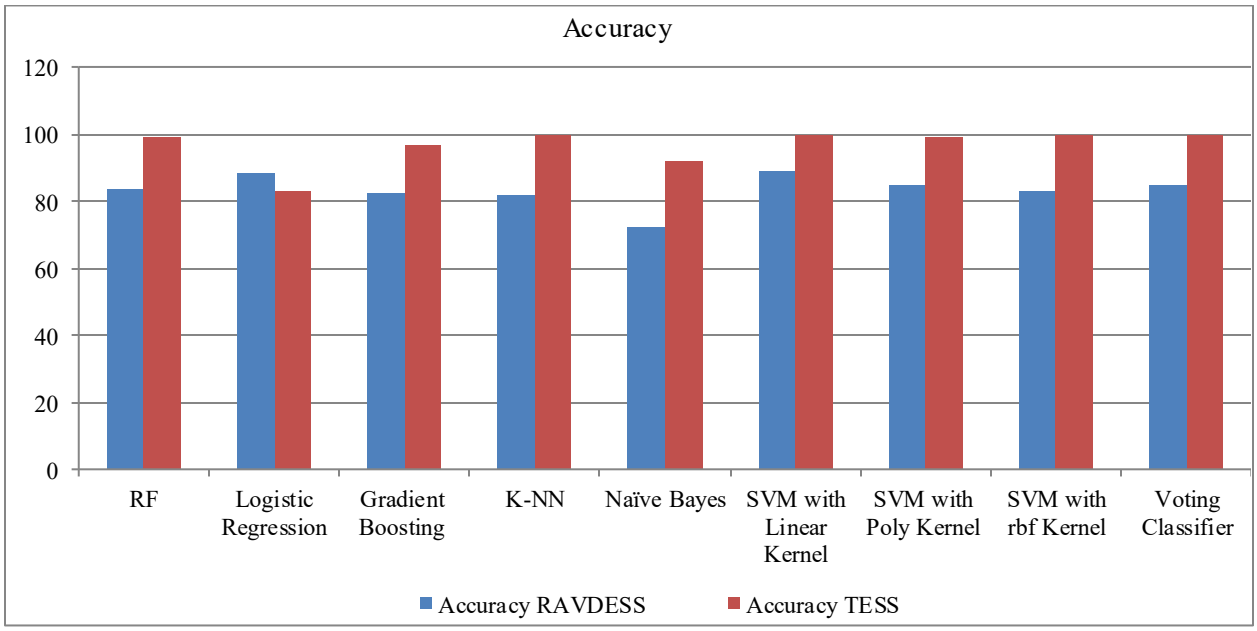



**Fig. 18 Comparison of accuracy for various classifiers**

*4.3.2. Comparative Performance based on Precision*

Comparison of the performance of classifiers between the TESS and RAVDESS datasets captures the considerable impact of data properties on the efficacy of a model. The TESS dataset had excellent performance for the majority of the classifiers, indicating that it has more separable features and sharper decision boundaries.

The RAVDESS dataset, on the other hand, had more subtle and intermingled features, making it more difficult to get such results, hence necessitating sophisticated modeling methods. This highlights the pivotal position played by dataset choice in stress detection studies, perhaps explaining why, to best optimize model performance, the underlying data characteristics must be understood.

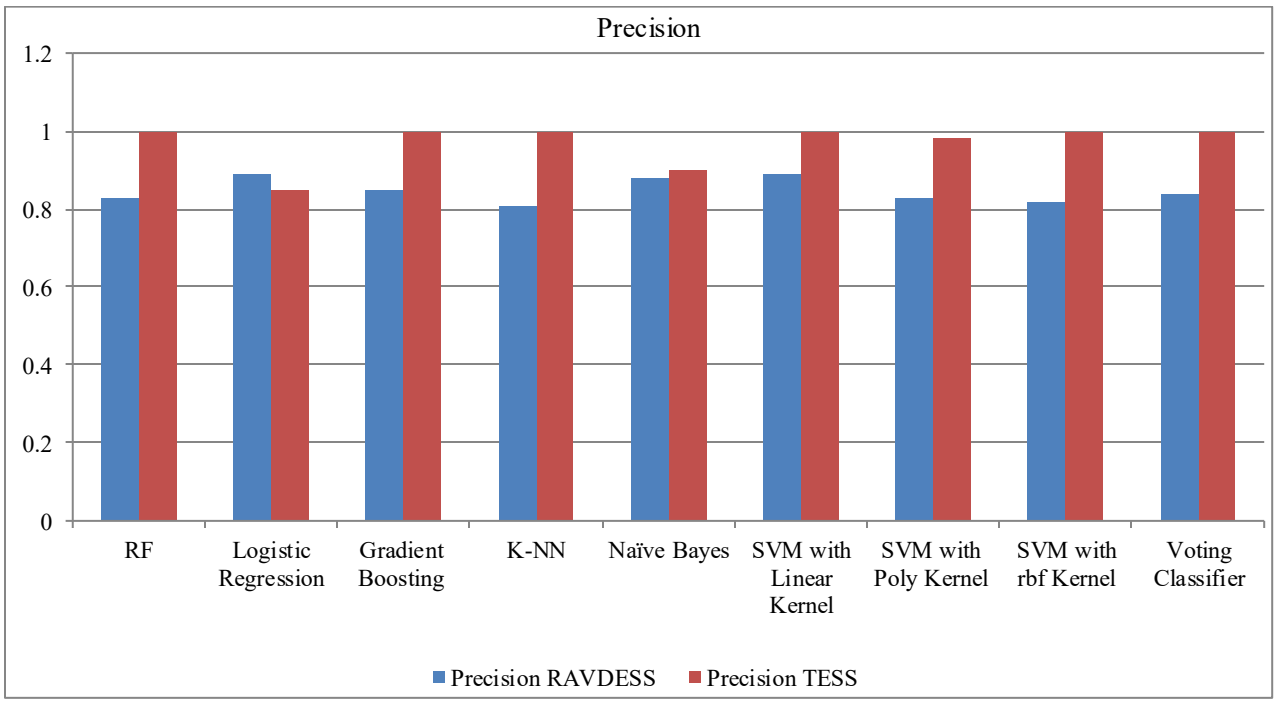



**Fig. 19 precision comparison of different models**

*4.3.3. Comparative Performance based on Recall*

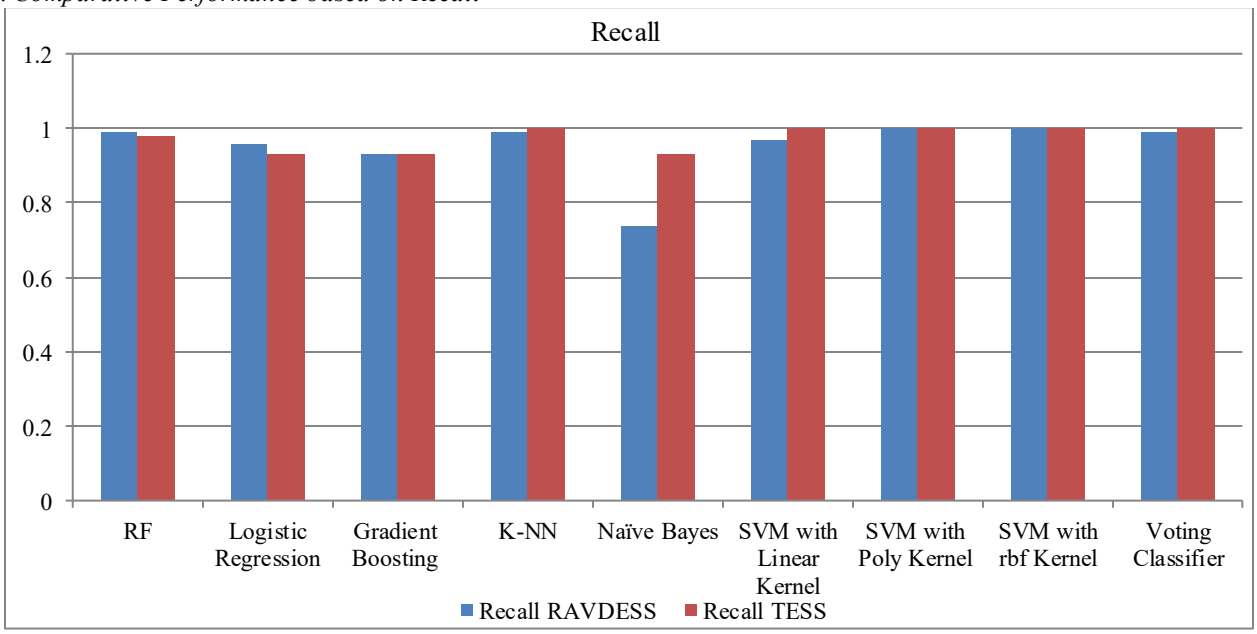



**Fig. 20 Recall of different ML models**

*4.3.4. Comparative Performance based on F1-Score*

Comparison of TESS and RAVDESS datasets illustrates the significance of data properties in model efficiency. The TESS dataset illustrated uniformly high accuracy across models, which indicates the existence of easily separable features that create clearer boundaries of decision. The RAVDESS dataset showed more complexity with subtle or overlapping features that defied some models and needed sophisticated methods for best performance. These results highlight the importance of choosing appropriate datasets and knowing their characteristics to improve the efficacy of anxiety detection devices.

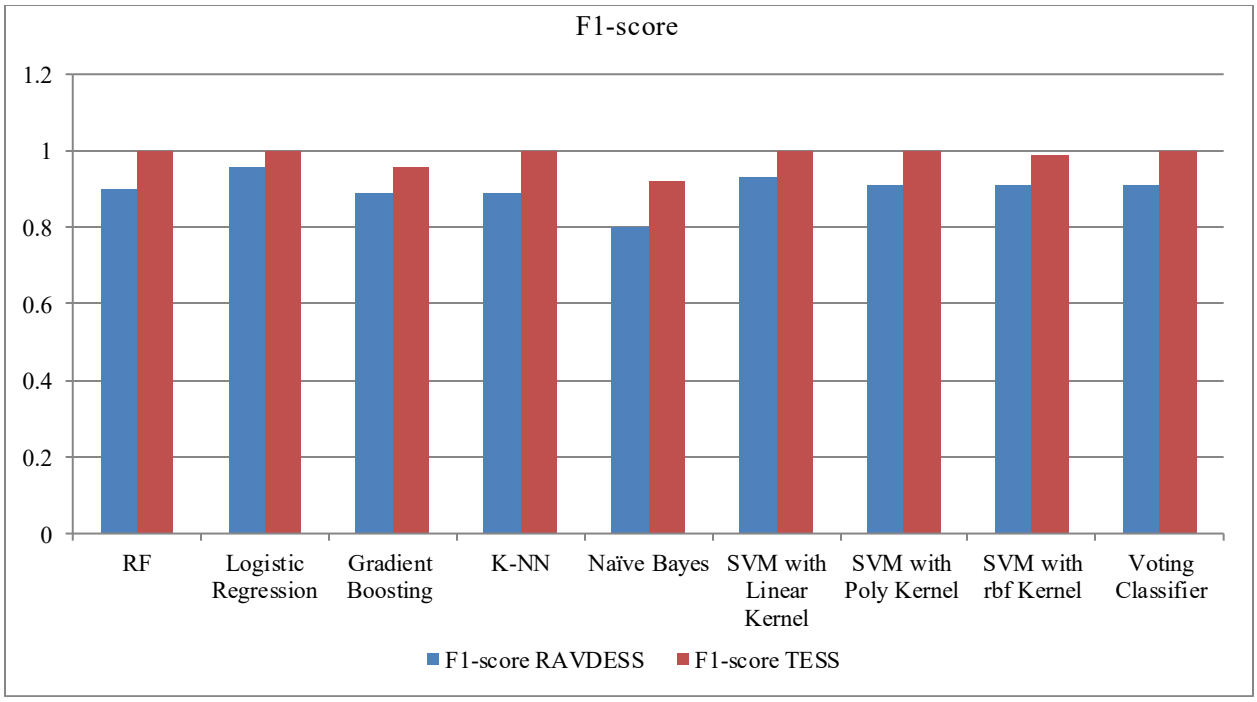



**Fig. 21 F1-score of different ML models**

***4.4. Discussion***

The performance evaluation using the TESS and RAVDESS datasets highlights the significant impact of dataset characteristics on classifier behaviour. Models such as SVM with linear and RBF kernels demonstrated stable and consistent performance across both datasets, confirming their robustness in identifying stress-related patterns. In contrast, K-Nearest Neighbors (KNN) achieved perfect accuracy on the TESS dataset but showed poor results on RAVDESS, likely due to its sensitivity to noise and overlapping emotional features.

Similarly, Naïve Bayes performed suboptimally on RAVDESS, reflecting the limitations of its feature independence assumption in more complex datasets. The Voting Classifier emerged as the best-performing model on TESS, benefiting from ensemble diversity and effectively combining individual learners' strengths. This underscores the potential of ensemble strategies in achieving higher accuracy in scenarios where emotional features are well-separated. Notably, all classifiers performed better on the TESS dataset, with Voting Classifier and KNN achieving 100% accuracy, making TESS a promising dataset for practical stress detection applications. On the more emotionally complex RAVDESS dataset, Logistic Regression and SVM (linear) achieved F1-scores exceeding 0.96, while other models showed diminished accuracy. These observations reinforce the importance of selecting classifiers based on the complexity and distribution of dataset features. Compared to prior studies such as Abd Al-Alim et al. (2024), who achieved 98% accuracy using wearable sensor data, and Abdelfattah et al. (2025), where RNN models reached an F1-score of 93% with multi-modal physiological data, the proposed approach achieved 100% accuracy on the TESS dataset and 88.97% on RAVDESS using only speech-based acoustic features. This superior performance is largely attributed to the use of carefully selected spectral and prosodic features. This binary classification framework reduces emotional overlap, and the ensemble Voting Classifier, which enhances predictive power by combining multiple base learners. In contrast to deep learning and sensor-based methods, the proposed approach is lightweight, non-invasive, and well-suited for scalable, real-time applications. This study exclusively used publicly available, anonymized datasets (TESS and RAVDESS), which are ethically approved for academic research. No personal or identifiable data was processed, and no human subjects were involved directly. However, as stress detection systems advance toward real-world deployment, it is vital to consider the societal implications, including privacy,

informed consent, and potential misuse in sensitive domains like insurance or employment. Ethical deployment must be guided by principles of transparency, data security, and responsible use. Overall, the findings not only validate the effectiveness of classical machine learning classifiers for audio-based stress detection but also emphasize the critical role of dataset structure, feature design, and model selection in achieving generalizable and ethical outcomes.

## 5. Conclusion

This study contributes greatly to stress detection with audio features by comparing different ML classifiers, such as RF, LoR, GB, KNN, NB, and SVM models, on two different datasets (RAVDESS and TESS). The research highlights the need for detailed feature extraction methods like Zero Crossing Rate, Spectral Centroid, and Chroma features to capture emotional patterns successfully. The Voting Classifier's superior performance with 100% accuracy on the TESS dataset shows the strength of ensemble techniques in stress detection tasks. The results also indicate the influence of dataset features on model efficiency and that more detailed datasets, such as RAVDESS, need tailored approaches for accurate detection. To further develop this research, further work will investigate the integration of DL methods more appropriately to manage temporal and nonlinear emotional patterns from speech. Cross-dataset testing and real-world deployment in embedded or mobile platforms will be examined for better generalizability and usability. In addition, hyperparameter tuning, feature selection methods, and automated model optimization will be utilized to increase accuracy at the expense of efficiency. Finally, the creation of morally accountable, privacy-protecting stress detection systems will take precedence to allow secure deployment in healthcare and real-world applications.

## Reference


[1] Serhat Hızlısoy, and Zekeriya Tüfekci, "Emotion Recognition from Turkish Music,"*European Journal of Science and Technology*, pp. 6-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Engin Demir, and Abdulkadir Tepecik "Analysis of Turkish Voice Recording Data with CountVectorizer and TF-IDFVectorizer Methods as BERT Models on Google Colab Platform and RapidMiner with Machine Learning Algorithms," *Firat University Journal of Science*, vol. 34, no. 1, PP. 19-29, 2022. [Google Scholar] [Publisher Link]

[3] Recep Sinan Arslan, and Necaattin Barişçi, "The Effect of Different Optimization Techniques on End-to-End Turkish Speech Recognition Systems that Use Connectionist Temporal Classification," *In IEEE 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, pp. 1-6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Kishor B. Bhangale, and Mohanaprasad Kothandaraman, "Speech Emotion Recognition Using the Novel PEmoNet (Parallel Emotion Network)," *Applied Acoustics*, vol. 212, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Emel Colakoglu, Serhat Hizlisoy, and Recep Sinan Arslan, *T-ser: An Efficient Speech Emotional Recognition Model for Turkish Language Based on Machine Learning Algorithms*, Innovations and Technologies in Engineering, Education Publishing House, İstanbul, pp.106-127, 2022. [Google Scholar] [Publisher Link]

[6] V.R. Archana, and B.M. Devaraju, "Stress Detection Using Machine Learning Algorithms," *International Journal of Research in Engineering, Science and Management*, vol. 3, no. 8, pp. 251-256, 2020. [Google Scholar] [Publisher Link]

[7] Xiyuan Hou et al., "EEG Based Stress Monitoring," *In 2015 IEEE International Conference on Systems, Man, and Cybernetics*, Hong Kong, China, pp. 3110-3115, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[8] Scott M. Monroe, "Modern Approaches to Conceptualizing and Measuring Human Life Stress," *Annual Review of Clinical Psychology*, vol. 4, no. 1, pp. 33-52, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[9] Thomas H. Holmes, and Richard H. Rahe, "The Social Readjustment Rating Scale," *Journal of Psychosomatic Research*, vol. 11, no. 2, pp. 213-218, 1967. [CrossRef] [Google Scholar] [Publisher Link]

[10] Giorgos Giannakakis et al., "Review on Psychological Stress Detection Using Biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440-460, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Eri Koibuchi, and Yoshio Suzuki, "Exercise Upregulates Salivary Amylase in Humans," *Experimental and Therapeutic Medicine*, vol. 7, no. 4, pp. 773-777, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[12] Matteo Zanetti et al., "Multilevel Assessment of Mental Stress Via Network Physiology Paradigm Using Consumer Wearable Devices," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 4, pp. 4409-4418, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Ahmed Ismail, Samir Abdlerazek, and Ibrahim M. El-Henawy, "Development of a Smart Healthcare System Based on Speech Recognition Using Support Vector Machine and Dynamic Time Warping," *Sustainability*, vol. 12, no. 6, pp. 1-15, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Emna Rejaibi et al., "MFCC-Based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech," *Biomedical Signal Processing and Control*, vol. 71, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Alexandros Liapis et al., "Advancing Stress Detection Methodology With Deep Learning Techniques Targeting UX Evaluation in AAL Scenarios: Applying Embeddings for Categorical Variables," *Electronics*, vol. 10, no. 13, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Serdar Yildirim, Yasin Kaya, and Fatih Kılıç, "A Modified Feature Selection Method Based on Metaheuristic Algorithms for Speech Emotion Recognition," *Applied Acoustics*, vol. 173, pp. 1-13, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Adrián Vázquez-Romero, and Ascensión Gallardo-Antolín, "Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks," *Entropy*, vol. 22, no. 6, pp. 1-17, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Sara Sardari et al., "Audio-Based Depression Detection Using Convolutional Autoencoder," *Expert Systems with Applications*, vol. 189, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Lili Zhu, Petros Spachos, and Stefano Gregori, "Multi-Modal Physiological Signals and Machine Learning for Stress Detection by Wearable Devices," *In 2022 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Messina, Italy, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Mohamed Abd Al-Alim et al., "A Machine-Learning Approach for Stress Detection Using Wearable Sensors in Free-Living Environments," *Computers in Biology and Medicine*, vol. 179, pp. 1- 37, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Eman Abdelfattah, Shreehar Joshi, and Shreekar Tiwari, "Machine and Deep Learning Models for Stress Detection Using Multi-Modal Physiological Data," *IEEE Access*, vol. 13, pp. 4597-4608, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[22] Anusha Koduru, Hima Bindu Valiveti, and Anil Kumar Budati, "Feature Extraction Algorithms to Improve the Speech Emotion Recognition Rate," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 45-55, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[23] Sudarsana Reddy Kadiri, RaviShankar Prasad, and Bayya Yegnanarayana, "Detection of Glottal Closure Instant and Glottal Open Region from Speech Signals Using Spectral Flatness Measure," *Speech Communication*, vol. 116, pp. 30-43, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Anthony M Dart, Xiao-Jun Du, and Bronwyn A Kingwell, "Gender, Sex Hormones and Autonomic Nervous Control of the Cardiovascular System,"*Cardiovascular Research*, vol. 53, no. 3, pp. 678-687, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[25] Eman Elsaeed et al., "Detecting Fake News in Social Media Using a Voting Classifier," *IEEE Access*, vol. 9, pp. 161909-161925, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Apit Hemakom, Danita Atiwiwat, and Pasin Israsena, "ECG and EEG-Based Detection and Multilevel Classification of Stress Using Machine Learning for Specified Genders: A Preliminary Study," *PLOS One*, vol. 18, no. 9, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Tatiur Rahman et al., "Mental Stress Recognition Using K-Nearest Neighbor (KNN) Classifier on EEG Signals," *International Conference on Materials, Electronics & Information Engineering, ICMEIE-2015*, Faculty of Engineering, University of Rajshahi, Bangladesh, pp. 1-4, 2015. [Google Scholar]

[28] Fatimah Alzamzami, Mohamad Hoda, and Abdulmotaleb El Saddik, "Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation," *IEEE Access*, vol. 8, pp. 101840-101858, 2020. [CrossRef] [Google Scholar] [Publisher Link]