I.J. Intelligent Systems and Applications, **2025, 6, 44-57**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2025.06.04

Determining Emotion Intensities from Audio Data Using Ensemble Models: A Late Fusion Approach

Simon Kipyatich Kiptoo*
Faculty of Pure and Applied Science, Department of Computing, Jomo Kenyatta University of Agriculture and Technology Nairobi, Kenya
E-mail: Simon.kiptoo@students.jkuat.ac.ke
*Corresponding author

Kennedy Ogada
Faculty of Pure and Applied Science, Department of Computing, Jomo Kenyatta University of Agriculture and Technology Nairobi, Kenya
E-mail: kodhiambo@scit.jkuat.ac.ke
ORCID iD: https://orcid.org/0000-0002-5671-0066

Tobias Mwalili
Faculty of Pure and Applied Science, Department of Computing, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
E-mail: tmwalili@jkuat.ac.ke

Received: 05 July 2025; Revised: 10 September 2025; Accepted: 16 October 2025; Published: 08 December 2025

Abstract: This paper presents an ensemble model in the determination of manifestation of emotion intensities from audio-dataset. An emotion denotes the mental state of the human mind or/and thought processes that represents a recognizable pattern of an entity like emotion arousal having a good similarity with its manifestation of vocal, facial or/and bodily signals. In this paper, we propose a stacking, late fusion approach where the best experimental outcome from two base models build from Random Forests and Extreme Gradient Boost are combined using simple majority voting. RAVDESS audio datasets, a public gender balanced dataset built by Ryerson University of Canada for the purpose of emotion study was used. 80% of the dataset was used for training while 20% was used for testing. Two features, MFCC and Chroma were introduced to the base models in a series of experimental setups and the outcome evaluated using confusion matrix, precision, recall and F1-Score. It was then compared to two state-of-the-art works done on KBES and RAVDESS datasets. This approach yielded an overall classification accuracy of 93%.

Index Terms: Emotion, Emotion Intensity, Multi-modal, Late Fusion, MFCC, Chroma.

1. Introduction

The manifestation of feelings by human beings occur through dialogue, facial expressions and written communications [1]. During dialogue, the pitch at which the sound is produced by the communicating persons, differ with every change of emotional state of the communicating person at the time. The pitch can be intense or mild depending on the topic of discussion. Emotion states of a human being relay the feelings expressed by the person every time of an occurrence and it helps to communicate the feelings and the kind of assistance required by the person at the time of expressing those emotions. These emotions are expressed via speech, facial expressions, gestures or other non-verbal signs [2]. The intensity of the emotion varies with the kind of occurrence the person is experiencing at a particular time.

Existing studies on emotion detection, avoid encoding the intensity of the observed emotions due to challenges posed by the approaches employed in emotion detection. These challenges range from complex models and algorithms to the choice of features. For example, Dimensional models map emotions into a continuous spectrum as compared to Categorical models, which map emotions into hard categories. Selecting a model for annotation therefore, proves to be challenging. More so, emotions change as conversations progress and coupled with sarcasm in the speaker's dialogue, the dynamicity of emotions becomes challenging in detecting and mapping emotions correctly. Fear and terror might be taken as one emotional state, but one is more intense than another. Also anger and rage is one emotional state, but one is a graduate of another. The most difficult part in differentiating these states is knowing the boundaries [3]. Adding to the

complexity of dynamicity is the modality factor. Modalities in speech signals are categorized into audio, text and visual. Emotions occur because of one or a combination of these modalities [4]. In that case, the ability to categorize feelings to belong to a certain emotional category requires a combination of modalities for classification as features, which indicate diverse emotional states may overlap, and may present multiple ways of expressing the same emotional states [5]. However, most researchers only detect, recognize and predict emotions, leaving out the intensities, which are vital in categorizing the emotions depicted by a person.

This study seeks to address the above shortcoming by using a late fusion approach, a more robust kind of ensemble models to map out the intensities in a discrete format. Ensemble models are a subfield of artificial intelligence that combines information obtained from diverse modalities like audio, text, video and text with the aim of building a more comprehensive and precise understanding of the underlying data. It seeks to process and scrutinize data from several modalities. It is the area of machine learning involved with combining dissimilar data sources (models) with the aim of capitalizing on the exceptional and complementary data in an algorithmic framework [6]. This concept is mostly applicable in the field of autonomous car, speech and emotion recognition. The concept employed in ensemble learning is that of model combination. It uses multiple model combination to advance the performance of a single model by using the strengths of one model to improve on the weaknesses of another model. The outcome is a robust and more accurate prediction. There are three major methods used to achieve the concept of multiple model combination. These are stacking, bagging and ensemble models and the research employed late fusion-based stacking which has an advantage over the others in terms of performance.

The structure of the paper is such that section 1 introduces the background of the study, section 2 delves into the works related to the field of study, section 3 presents the approach followed in the study giving the analysis of the architecture of the proposed model, section 4 explains the experimental setups, section 5 gives an analysis and discussion of the experimental outcome and finally section 5 gives the conclusion of whole study.

2. Related Works

In this section, the review of the state-of-the art models presented by various researchers, reveal the challenges posed during the determination of emotion intensities. A number of research experiments on determination of emotion intensities have been carried out using diverse methods. However, it is noted that while the studies have been successful, they mainly revolve around databases with facial image datasets and data from texts. Determining emotion intensities from audio datasets has proved elusive and challenging due to the dynamicity of conversations, which generate the audio data.

[7], proposed a deep learning-based implementation for speech emotion recognition, which combines DCNN and a BLSTM network with a time-distributed, flatten (TDF) layer. They evaluated the model by employing a cross-corpus/multi-corpus training and transfer learning for the Bangla and English languages using the SUST Bangla Emotional Speech Corpus (SUBESCO) and RAVDESS datasets for cross-lingual experiments. The model achieved an overall weighted accuracy of 86.9% on SUBESCO and 82.7% on RAVDESS.

[7-9], separately investigated a deep learning model for classifying emotions and measuring intensity. The Single and Cascaded Deep Learning frame models were employed on three integrated speech signal transformation methods, MFCC, STFT, and Chroma STFT. They evaluated the proposed model using a self-developed corpus; KUET Bangla Emotional Speech (KBES) dataset. On the Single DL frame, they achieved 61.11%, 58.89% and 57.22% accuracy respectively. The Cascaded DL frame achieved 71.67%, 69.44% and 67.78% accuracy respectively.

[7-11], each proposed a novel scheme for Recognition of Emotion with Intensity from Speech. They used Single and Cascaded Deep Learning frame models by integrating three speech signal transformation methods MFCC, STFT, and Chroma STFT. The models were evaluated on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) benchmark dataset. On the Single DL frame, they achieved 84.21%, 78.65%, 76.02%, 67.25% and 75.43% accuracy respectively. The Cascaded DL frame achieved 88.30%, 87.71%, 81.28%, 72.51%, and 77.48% accuracy respectively.

3. Proposed Approach

In the study, experimental approach was adopted in which feature extraction from the audio data was done where MFCC features were extracted first followed by the Chroma features. Feature selection was done after separating acoustic features from the lexical features [12]. The resulting features selected were planned to be used in facilitating the learning process of the model as well as the generalization in order to understand the human interpretation of the model. The next step in the approach was feature classification and this was done based on the domain which each feature belongs. There are two main feature domains in audio dataset; Time domain and Frequency domain. The main domain features obtained for our experiments were Mel Frequency Cepstral Coefficients (MFCC) shown in Table 1, and pitch. Feature coding had already been done on the dataset, since RAVDESS dataset is a database which had already been prepared and readily available online.

Table.1. A mel spectrogram

Row No.	Mel Frequency Cepstral Coefficient Feature
0	[-76.38477, -76.38477, -76.38477, -76.38477, -….
1	[-75.33552, -75.44532, -75.55403, -75.20395, -….
2	[-75.15071, -75.15071, -75.15071, -75.15071, -….
3	[-75.26845, -75.26845, -75.26845, -75.26845, -….
4	[-80.14738, -80.14738, -80.14738, -80.14738, -….

The features were then divided into two sets, one for testing and the other for training. The training dataset was placed at 80% of the whole dataset while the testing data set took the remaining 20%. For the purpose of maintaining same class distribution in both training and testing sets, stratified splits are introduced to ensure unbiased evaluation by use of train_test_split with stratify parameter set to target labels. The features were then deployed on a model and a pre-trained classifier shown in figure1.

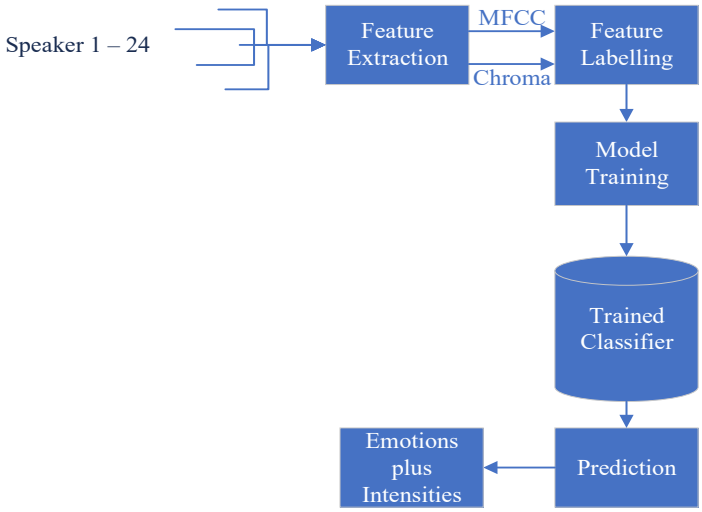


Fig.1. System design pipeline

To ensure equal contribution to the model training by all features in order to improve numerical stability, a feature scaling technique known as StandardScaler is used to scale features to zero mean and unit variance. Recursive feature elimination (RFE) plays an important role in reducing the complexity of the model, avoiding overfitting and enhance performance by choosing the most useful features. This is done by deploying RFE with a RandonForestClassifier which removes less impactful features and chooses the top n-features recursively. Adaptive Synthetic (ADASYN) Sampling is used to create synthetic samples based on density distribution for minority classes. This solve class imbalance which can result in biased models which favor majority class. A block diagram representation of the proposed model is shown in figure 2.

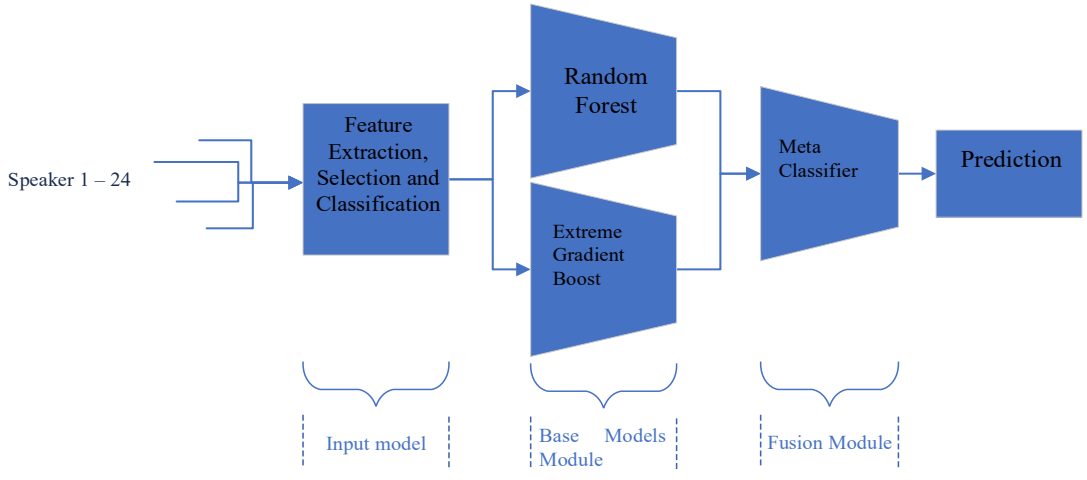


Fig.2. Block diagram of the proposed model

4. Experiments

Our experimental process started with feature extraction from both MFCC and Chroma following the recommended feature extraction procedure.

4.1. MFCC Feature Extraction

MFCC features were extracted following a step by step process as illustrated in figure 3. Conversion of analog audio data to digital signals enables the extraction of acoustic features where the compressed energy resulting from the process is produced at a high energy level at the pre-emphasis stage. This is necessary for the provision of balance in the voiced sound spectrum having a high roll off in the high frequency region. To acquire the short term spectral measurements, the individual speech temporal sound characteristics are tracked at the windowing stage. This is done by passing them through a 20ms window at an interval of 10ms and then apply a Discrete Fourier Transform to convert the sound signal to frequency domain from time domain. The process changes the signal into a magnitude spectrum when a mathematical function (1) is applied.

$$X(k) = \sum_{n=0}^{n-1} x\,(n) e^{\frac{-j2\pi nk}{N}}; o \leq k \leq N-1 \quad (1)$$

where; x(n) = long analysis window (hamming) of N samples and K = the length of DFT.

The calculation of Mel spectrum is obtained when the Fourier transformed signal is passed through group of band pass filters known as Mel-filter bank. This exhibits the short-term energy of the actual logarithm on the Mel frequency scale as a discrete cosine transform. The filter bank assists the grounding of MFCC on a signal disintegration that can be transformed from an actual frequency (f) to Mel frequency Mel (f).

$$\text{MeI}(f) = 2595\text{x}\log_{10}(1 + \text{f}/_{700}) \quad (2)$$

where: *Mel(f)* is the Perceived frequency in Mels, *f* is the Physical frequency in Hertz (Hz)

From the Mel filter banks, we get the feature energies logarithms which are due to a compression operation scaling down the features to closely imitate the real human mechanism and permit the use of cepstral mean subtraction channel normalization technique. The next step is to compute the DCT log filter banks energies in order to eliminate the correlation of the filter banks resulting from their overlapping characteristic. The Discrete Cosine Transform yields 26 coefficients out of which only 12 are used in the study. Because of their fast-changing characteristics, the Coefficients, which lie in the higher bands of the DCT, are eliminated to avoid the degradation of the performance of automatic speech recognition. Mel Frequency Cepstrum (MFC), which is a short-term sound spectrum representation where used in the study because of the extensive usage in the classification of audio and due to its good performance [13]. Mel Cepstral principles acquire the short-time spectral shape, which is a sound characteristic feature, which contains important data representing the voice quality and production effect. By calculating the energy short-term spectrum cosine transform of the actual logarithm, we acquire the coefficients.

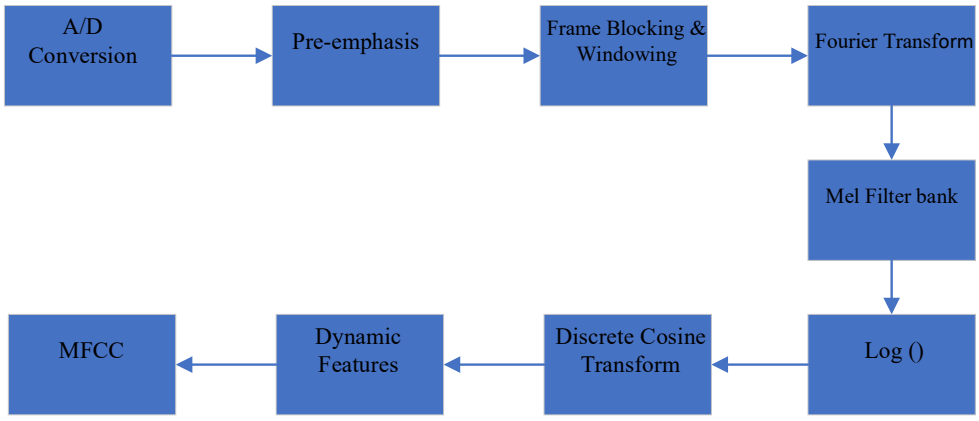


Fig.3. MFCC Feature extraction process

4.2. Chroma Feature Extraction

The 12 diverse pitch categories in music is sometimes referred to as chroma feature of chromagram due to its close relation. These features are very prominent tools for music analysis when categorizing pitches. Their main property is the ability to capture the two important music characteristics, harmony and melody, while being robust to instrumental and timbre dynamics [14]. The main objective of chroma features is to represent harmonic content of a short time audio window. They are obtained as feature vector through extraction from the magnitude spectrum using STFT (Short Time

Fourier Transform), CENS (Chroma Energy Normalized) and CGT (Constant Q Transform) among others. The basic thought is that the perception regarding two musical pitches by humans is related in color if they vary by an octave. According to this perception then pitch can be divided into two components known as chroma and tone height as shown in figure 4.

The extraction of chroma features follows an eight-step procedure; signal input, spectral analysis, Fourier transform, frequency filtering, peak detection, frequency computation referencing, mapping of pitch class and feature normalization as labeled in figure 5. Music signal is the main input in chroma feature extraction. The purpose of spectral analysis is to extract the frequency components of the signal before converting them into a spectrogram using Fourier transform which is a time frequency analysis. Filtering of the frequency is done using a frequency range of 100 Hertz to 5000 Hertz and for peak detection consideration, the spectrum maximum local values are taken. A 440 Hertz frequency is applied when doing frequency computation referencing procedure in order to estimate deviation. The estimated reference frequency is then used to map the pitch class by use of a cosine function weighting scheme. This scheme considers a total of eight harmonic frequencies for every frequency and for the purpose of mapping values on a third of a semitone, pitch class vector distribution size must be equal to thirty-six. Finally, frame by frame normalization of the feature is done to remove global loudness dependency by dividing through the maximum value.

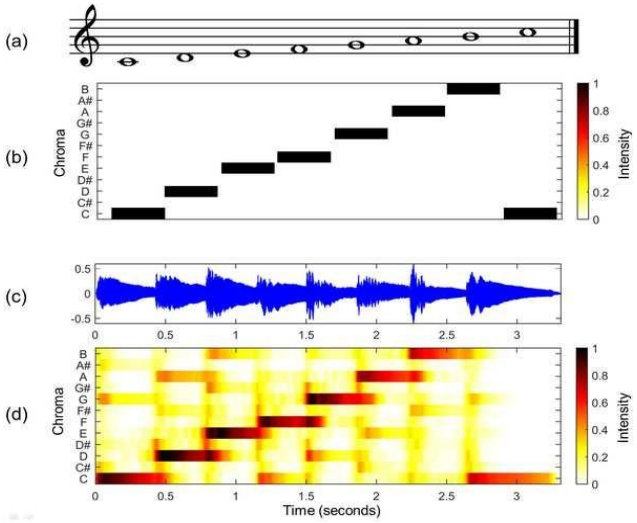
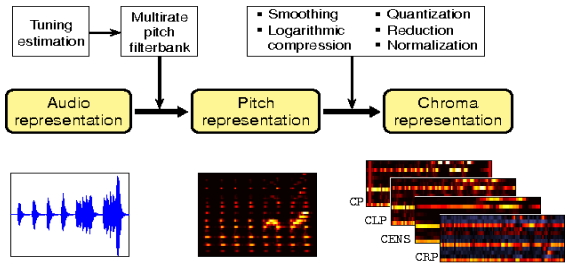


Fig.4. (a) C-major scale musical score. (b) Chroma gram of the score. (C) Piano played Audio recording of the C-major scale. (d) Chroma gram of the audio recording



(Source: M. Müller& S. Ewert; 2011)

Fig.5. Chroma feature extraction pipeline

Performing Fourier Transform, Constant-Q Transform and Log Frequency Spectrogram

In Chroma feature extraction, the application of Fourier transform process exposes the spectrum of frequency components composing the initial signal by mapping a time-dependent signal to a frequency dependent function. While a signal reveals time information and conceals frequency information, Fourier transform displays frequency information and conceals time information. To reverse the concealed time information, a compromise between time-based and frequency-based representation known as short-time Fourier transform (STFT) introduced by Dennis Gabor in 1946 is applied. This is done by defining the sinusoidal frequency and phase content variation of an initial signal overtime. This reveals the contained frequencies as well as the time intervals of the frequencies. STFT outlines important class of time-frequency distributions responsible for specifying intricate amplitude Vis a Vis time and frequency for any signal which is represented as mathematical function (Eq...). if we represent the evenly sampled f(t) and g(t) function by f[n] and g[n], then the discrete STFT (D) over a densely supported g window function can be derived as,

$$Fgf_{[n,m]} = \sum_{m=0}^{M-1} f[n-m]g[m] \in k[m]\ ,\ \epsilon_k[m] = e^{-2\pi}m\frac{k}{N} \tag{3}$$

where,

M = window length of g

N = No of samples in f

Constant-Q Transform (CQT) is another useful tool for extracting chroma features. It is a bank of filters with geometrically spaced centre frequencies, which are denoted by a mathematical function;

$$fk = f0.2^{kb}(k = 0 \ldots), \tag{4}$$

where b = no of filters in an octave.

When the correct selection of minimal center frequency (*fo*) and *b*, they directly relate to musical notes. This is why the CQT is so handy in the extraction process. When performing CQT, the time resolution increases toward the higher frequencies, mimicking the human auditory system. However, Discreet Fourier Transform (DFT) is favoured over CQT due to some draw backs. Computing CQT is more tedious than when computing DFT and it does not have an inverse transform that enables the seamless restoration of the original signal from the transformed coefficients. Also, because of the time resolution variation for diverse frequency bins in CQT, sampling of the diverse frequency bins is not synchronized. Lastly, compared to spectrograms derived from STFT in a consecutive time frames, working with the data structure resulting from CQT is more tedious.

Digitally speaking, an octave in music represents a distance between two frequencies with a variation factor of two divided into twelve units spaced logarithmically. Musical MIDI notation contains 128 pitches numbered from 0-127 with MIDI pitch p equals to 69 corresponding to pitch A4 usually used as a standard measure for tuning musical instruments with a Centre frequency of 440 Hertz represented by a formula;

$$\mathrm{F}_{\mathrm{p_{it\,ch(p)}}=2(\mathrm{p}-69)\mathrm{y}_{12\cdot 44}\ \text{where p}\ \varepsilon[0:27]} \tag{5}$$

The logarithmic perception of frequency motivates the use of a time-frequency representation with a logarithmic frequency axis labeled by the pitches of the equal-tempered scale. To derive such a representation from a given spectrogram representation, the basic idea is to assign each spectral coefficient X (m, k) to the pitch with center frequency that is closest to the frequency Fcoef(k). More precisely, a definition for each pitch p $\varepsilon[0:127]$ the set $\mathrm{P(p)} \coloneqq \left\{\mathrm{k}\epsilon[0:\mathrm{K}]:\mathrm{F}_{\mathrm{p_itch(p-0.5)\leq F_{C\circ ef(k)<Fpi^{tch}_{(P+0\cdot 5)}}}}\right\}$ is done. From this, a log-frequency spectrogram $\mathrm{YLF}:\mathrm{Zx}[0:127] \rightarrow \mathrm{R} \geq 0$ defined by equation (6) is done.

$$y_{LF(m,p)} \coloneqq \sum_{k\in P(p)} |x(m,k)|^2. \tag{6}$$

When defined as such, the frequency is logarithmically partitioned and linearly labeled as per MIDI pitches.

These features were then subjected to four series of experimental setups using two machine-learning models, Extreme Gradient Boost and Random Forest. The first experimental setup was Determining Intensities of the Eight-Emotion State from MFCC and Chroma features. Second setup involved determining Emotion Intensity of Emotions with Negative Valence from MFCC and Chroma features. Third setup was to determine Emotion Intensity of Emotions with Positive Valence from MFCC and Chroma features omitting Calm and the fourth setup involved determining Emotion Intensity of Emotions with Positive Valence from MFCC and Chroma features with Calm and Neutral Merged. The two models were then combined as base models in a late fusion based stacked approach where the best results from the experimental setup in each model was used to determine the emotion intensities.

4.3. Database and Evaluation Metrics

In our study, the database used to benchmark the experiments was the RAVDESS (Ryerson Audio-Visual database of Emotion Speech and Song) dataset [15]. A public gender balanced dataset built by the Ryerson University of Canada for the purpose of emotion study. It has 7356 files from 24 soundtracks of 12 male and 12 female vocalists. Each vocalist voiced two separate statements two times deliberately from the song and speech states mimicking eight emotional situations of happy, angry, disgust, sad, fearful, neutral, surprise and calm. The vocalization was done such that it gave three separate modality states of Audio-only, Video-only and Audio-Video. The song phrases were purposely uttered in six emotion states (happy, angry, disgust, sad, fearful and neutral) with five of the emotion states uttered in two emotion intensity levels of *strong* and *normal*, Table 2. The utterances contained a seven syllabi length harmonized in word frequency and familiarity using an MRC (Medical Research Council) psycholinguistic database.

The sung words were same as MIDI sounds of a piano having static acoustic intensity of six eight-note played at 300ms with a quarter-note finish at 600ms. The melody modes of major, neutral and minor were used to rate the utterances while each melody tonality synonymous with each emotion was made to emotionally associate consistently. To rate the perceived valence of the three melody modes, a self-assessment manikin (SAM) nine-point valence scale was used. Three major epochs of task-presentation recorded at 4500ms, counting epoch done at 2400ms and vocalization epoch done at 4800ms, resulting to a 13,700ms stimulus trial time marked the stimulus trial time.

The dataset was then tested on a late-fusion stacked model created from two base models of Extreme Gradient Boost and Random Forest. This leverages the strengths of the two diverse models by capturing various data patterns. XGBoost is efficient in handling structured data and complex interactions by using the MLP Classifier with a deep architecture (Multi-Layer Perceptron) which is excellent in capturing non-linear patterns, while Random Forest is robust in dealing with overfitting and good in handling high-dimensional data. For classification reports, precision, recall and F1-score were used to evaluate the model. These classification metrics helps to understand the performance of the model the diverse emotions. Confusion matrix on the other hand visualizes true versus predicted labels and highlights the correct and incorrect model predictions while accuracy score determines the overall model's performance measure.

Table 2. RAVDESS dataset sample

Row No.	Actor	Emotion State	Intensity	Actor Number and Path
0	Male	Neutral	Normal	1 Ravdess/Actor_1/03-01-01-01-01-01-01.wav
1	Male	Neutral	Normal	1 Ravdess/Actor_1/03-01-01-01-01-02-01.wav
2	Male	Neutral	Normal	1 Ravdess/Actor_1/03-01-01-01-01-01-01.wav
3	Male	Neutral	Normal	1 Ravdess/Actor_1/03-01-01-01-02-02-01.wav
4	Male	Calm	Normal	1 Ravdess/Actor_1/03-01-02-01-01-01-01.wav
…	…	…	…	…
1435	Female	Surprise	Normal	24 Ravdess/Actor_24/03-01-08-01-02-02-24.wav
1436	Female	Surprise	Strong	24 Ravdess/Actor_24/03-01-08-02-01-01-24.wav
1437	Female	Surprise	Strong	24 Ravdess/Actor_24/03-01-08-02-01-02-24.wav
1438	Female	Surprise	Strong	24 Ravdess/Actor_24/03-01-08-02-02-01-24.wav
1439	Female	Surprise	Strong	24 avdess/Actor_24/03-01-08-02-02-02-24.wav

1400 s x 5 columns

Table 3. Outcome of the XGBoost RF Models when tested with all emotions

Emotion Category	Precision		Recall		F1-Score		Support	
	XGB	RF	XGB	RF	XGB	RF	XGB	RF
Angry_normal	0.39	0.32	0.37	0.32	0.38	0.32	19	19
Angry_strong	0.80	0.75	0.84	0.95	0.82	0.84	19	19
Calm_normal	0.22	0.21	0.26	0.21	0.24	0.21	19	19
Calm_strong	0.23	0.21	0.35	0.30	0.28	0.25	20	20
Disgust_normal	0.19	0.24	0.21	0.26	0.20	0.25	19	19
Disgust_strong	0.54	0.39	0.37	0.37	0.44	0.38	19	19
Fear_normal	0.44	0.31	0.40	0.25	0.42	0.28	20	20
Fear_strong	0.31	0.53	0.26	0.47	0.29	0.50	19	19
Happy_normal	0.35	0.30	0.42	0.37	0.38	0.33	19	19
Happy_strong	0.45	0.43	0.53	0.32	0.49	0.36	19	19
Neural_normal	0.47	0.45	0.37	0.47	0.41	0.46	19	19
Sad_normal	0.33	0.59	0.26	0.68	0.29	0.63	19	19
Sad_strong	0.36	0.69	0.21	0.45	0.27	0.55	19	20
Surprise_normal	0.32	0.39	0.47	0.37	0.38	0.38	19	19
Surprise_strong	0.36	0.38	0.21	0.32	0.27	0.34	19	19
Accuracy					0.38	0.41	288	288
Macro avg	0.39	0.41	0.38	0.41	0.38	0.41	288	288
Weighted avg	0.39	0.41	0.38	0.41	0.38	0.40	288	288

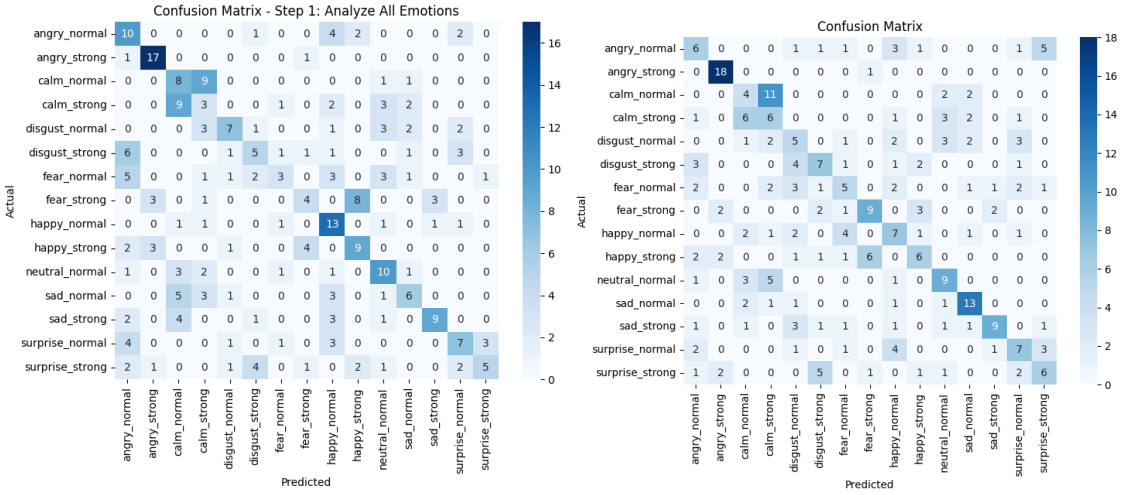
5. Results and Discussion

The results of the four experimental setups were depicted in a table format as indicated below. 80% of the dataset was used for training and 20% was used for testing. For equal contribution to the model training by all features, a feature scaling technique known as StandardScaler is used to scale features to zero mean and unit variance. To reduce the model complexity and avoid overfitting, Recursive feature elimination (RFE) was introduced by deploying RFE with a RandonForestClassifier which removes less impactful features and chooses the top n-features recursively. To solve class imbalance and avoid biased models Adaptive Synthetic (ADASYN) Sampling is used to create synthetic samples based on density distribution for minority classes. Each table indicates the precision, recall, F1-Score and the support feature

elements involved in the predictions of the two models.

5.1. Results from the 1st Setup: Determining Intensities of the Eight-Emotion State from MFCC+Chroma

After training and testing the models on the data, and using the eight emotions states (angry, calm, disgust, fear, happy, neutral, sad and surprise) as variables, the overall performance attained was 38%, on the XGBoost model, only posting a good recall of 84% in determining the intensity of angry_ strong and 53% recall on happy strong. The rest were below 50%. The RF model on the other hand attained a 41% accuracy with good hit of 95% recall on angry_ strong and 68% recall on sad_ normal. The rest were below 50% recall, Table 3. The high miss rate (fig.6a&b) and low accuracy on both models is attributed to the few numbers of support features on the many classification categories.



a: Confusion Matrix from XGBoost b: Confusion Matrix from RF

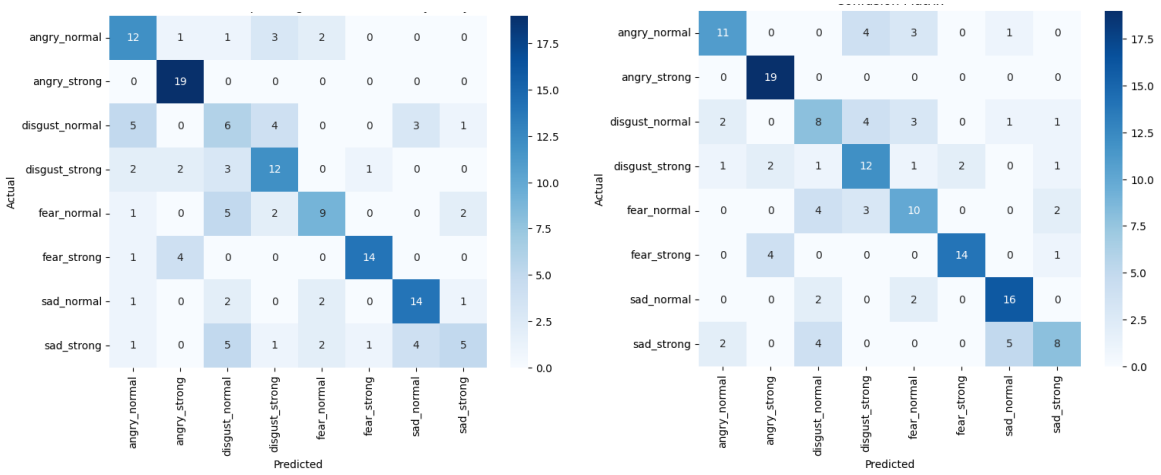
Fig.6. Confusion matrix from the analysis of all the emotions

Table 4. Outcome of XGB and RF models on Negative emotions

Emotion Category	Precision		Recall		F1-Score		Support	
	XGB	RF	XGB	RF	XGB	RF	XGB	RF
Angry_normal	0.62	0.69	0.53	0.58	0.57	0.63	19	19
Angry_strong	0.79	0.76	0.84	1.00	1.00	0.86	19	19
Disgust_normal	0.37	0.42	0.37	0.42	0.37	0.42	19	19
Disgust_strong	0.56	0.52	0.50	0.60	0.53	0.56	20	20
Fear_normal	0.42	0.53	0.42	0.53	0.42	0.53	19	19
Fear_strong	0.63	0.88	0.63	0.74	0.63	0.80	19	19
Sad_normal	0.62	0.70	0.75	0.80	0.68	0.74	20	20
Sad_strong	0.33	0.62	0.26	0.42	0.29	0.50	19	19
Accuracy					0.56	0.64	154	154
Macro avg	0.54	0.64	0.56	0.64	0.55	0.63	154	154
Weighted avg	0.54	0.64	0.56	0.64	0.55	0.63	154	154

5.2. Results from the 2nd Setup: Determining Intensity of Emotions with Negative Valence from MFCC + Chroma

The emotion states (Angry, Disgust, Fear, and Sad), exhibit negative characteristics hence grouped as containing negative valence. When introduced to the models, the performance increased from38% to 56% for the XGBoost model and 41% to 64% classification accuracy for the RF model. This is attributed to the reduced number of classifications. For the categories in the XGBoost, all attained a recall of over 50% except for disgust_ normal, fear_ normal and sad_ strong, which attained a recall of below 50%. The low recall on the three emotion classes is attributed to the almost similar characteristics exhibited by the emotion states. For the RF model, the hit rate was higher on the categories. Only two emotion classes posted a recall of 42%. These are disgust_ normal and sad_ strong. This again is attributed to the characteristic similarities of the two categories. The drop on the two models was as a result of an increase in the number of the classification categories from the earlier experimental set up as tabulated in Table 4. Figure 7 displays the hit and miss rates of the two models.



(a: Confusion Matrix from XGBoost) (b: Confusion Matrix from RF)

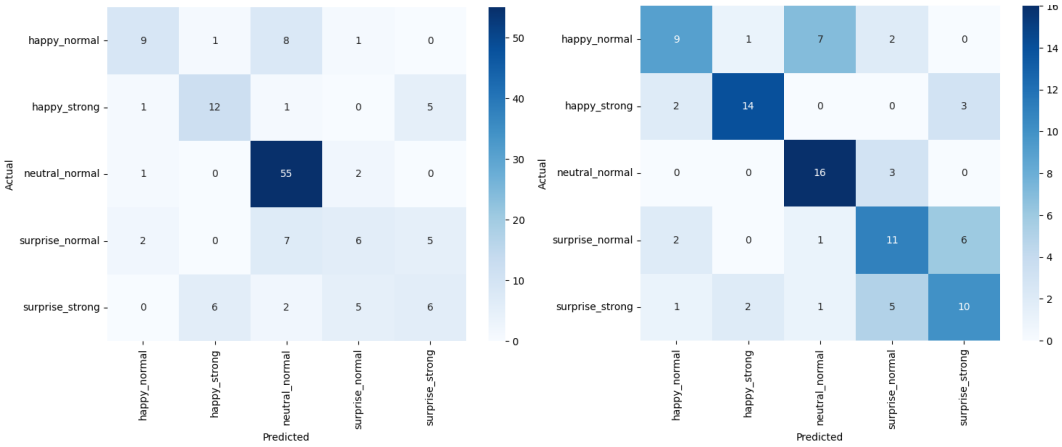
Fig.7. Confusion matrix from emotions with negative valence

5.3. Results from the 3rd Setup: Determining Emotion Intensity of Emotions with Positive Valence from MFCC + Chroma omitting Calm

Calm emotion state was omitted at this stage as it exhibits similar characteristics with Neutral state. Later the two were merge and the outcome determined. The model performance for the five emotion states (Happy_ Normal, Happy_ Strong, Neutral_ Normal, Surprise_ Normal and Surprise_ Strong), improved drastically from 56% to 64% classification accuracy for the XGBoost model. It posted a recall of over 50% for all the emotion states with happy_ strong and Neutral_ normal posing a recall of 87% and 84% respectively. The RF model posted an improvement from 64% classification accuracy to 67% classification accuracy. In Table 5, all the emotion states posted over 50% recall. This increase was attributed to the fewer categories being classified Vis a vise the number of the support features.

Table 5. Outcome of XGB and RF models when tested with positive emotions minus calm

Emotion Category	Precision		Recall		F1-Score		Support	
	XGB	RF	XGB	RF	XGB	RF	XGB	RF
Happy_normal	0.59	0.79	0.53	0.58	0.56	0.67	19	19
Happy_strong	0.82	0.89	0.87	0.84	0.78	0.86	19	19
Neutral_normal	0.64	0.59	0.84	0.84	0.73	0.70	19	19
Surprise_normal	0.53	0.52	0.50	0.55	0.51	0.54	20	20
Surprise_strong	0.61	0.62	0.58	0.53	0.59	0.57	19	19
Accuracy					0.64	0.67	96	96
Macro avg	0.64	0.68	0.64	0.67	0.63	0.67	96	96
Weighted avg	0.64	0.68	0.64	0.67	0.63	0.67	96	96

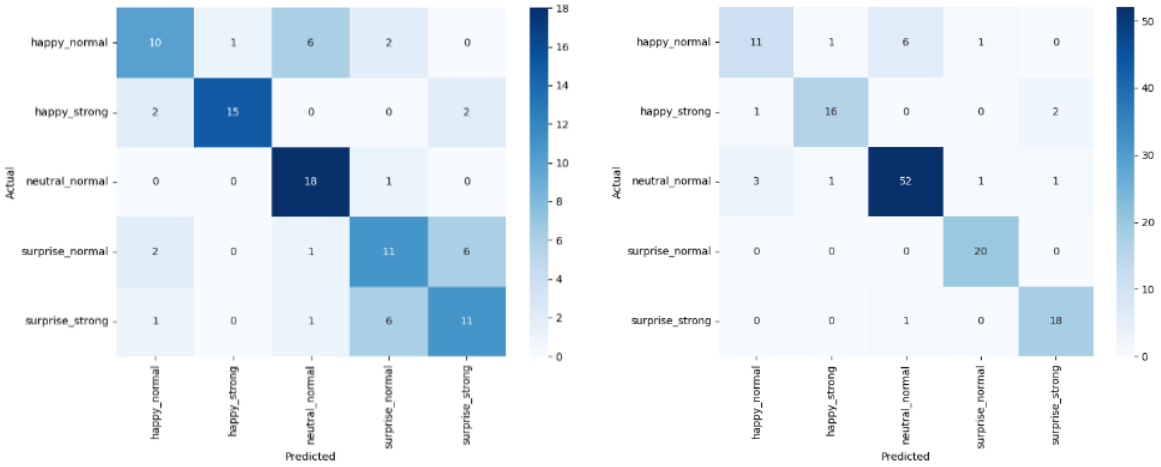


(a: Confusion Matrix from XGBoost) (b: Confusion Matrix from RF)

Fig.8. Confusion matrix from emotions with positive valence

5.4. Results from the 4th Setup: Determining Emotion Intensity of Emotions with Positive Valence from MFCC + Chroma with Calm and Neutral Merged

During this stage, the variables calm and neutral were merged due to the shared similarity of characteristics and neutral having only one intensity variable; normal. The model performance improved further to 68% classification accuracy with neutral_ normal posting an impressive recall of 95% on the XGBoost model. For the RF model, the improvement was even greater as it realized a classification accuracy of 87% with all the categories posting a recall of over 80% except happy_ normal which posted a 53% recall as depicted in Table 6. The drastic improvement from the previous outcome for both models was attributed again to the fewer classification categories used than those used in the 3rd experimental setup. Also combining Calm and Neutral emotion states, created a feature engineering effect, thus increasing the support features from the previous 19 to 58.



a: Confusion Matrix from XGBoost b: Confusion Matrix from RF

Fig.9. Confusion matrix from emotions with positive valence with calm and neutral merged

Table 6. Outcomes of XGB and RF models when tested with positive emotions with calm and neutral merged

Emotion Category	Precision		Recall		F1-Score		Support	
	XGB	RF	XGB	RF	XGB	RF	XGB	RF
Happy_normal	0.71	0.73	0.53	0.58	0.61	0.65	19	19
Happy_strong	0.76	0.89	0.68	0.84	0.72	0.86	19	19
Neutral_normal	0.76	0.88	0.95	0.90	0.85	0.89	58	58
Surprise_normal	0.47	0.91	0.35	1.00	0.40	0.95	20	20
Surprise_strong	0.41	0.86	0.37	0.95	0.39	0.90	19	19
Accuracy					0.68	0.87	135	135
Macro avg	0.62	0.85	0.58	0.85	0.59	0.85	135	135
Weighted avg	0.66	0.86	0.68	0.87	0.66	0.86	135	135

Table 7. Performance of the proposed model

Emotion Category	Precision	Recall	F1-score	Support
happy_normal	0.93	0.98	0.96	55
happy_strong	0.91	1.00	0.95	62
neutral_normal	0.97	0.97	0.97	58
surprise_normal	0.90	0.90	0.90	60
surprise_strong	0.97	0.80	0.87	56
accuracy			0.93	291
macro avg	0.93	0.93	0.93	291
weighted avg	0.93	0.93	0.93	291

Table 6 shows the best results of the two models from the fourth experimental setup. Extreme Gradient Boost attained a classification accuracy of 68% while Random Forest achieved a classification accuracy of 87%. These results formed the input of our proposed model where they were combined in a late-fusion stacking approach and the results are

shown in Table 7. Our proposed model achieved a classification accuracy of 93% and this was attributed to the combination of features from the base models increasing the number of support features, thus creating a feature engineering effect. Likewise, when the features of two emotion states depicting similar emotion characteristics were merged, the number of support features increased. Coupled with the fewer classification categories, the outcome of the model posted a recall of over 80% each in all categories. The confusion matrix in figure 10 shows the high hit rate due to the fused models.

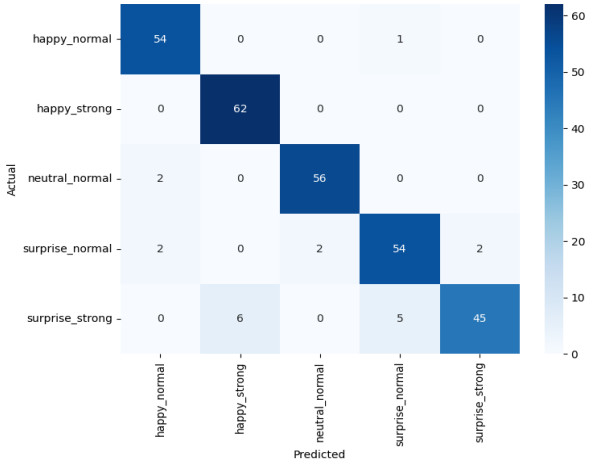


Fig.10. Confusion matrix from results of the proposed model

5.5. Comparison with Current Models

The performance of our proposed model was benchmarked on the works of two categories of research done using two different datasets. The first category of research work was on RAVDESS dataset shown in Table 8 and the second category of research work was on Kuet Bangla Emotion Speech (KBES), a self-developed corpus, Table 9. The two state-of-the art works deployed a single and cascaded deep learning frame models by integrating three speech signal transformation methods MFCC, STFT, and Chroma STFT on each dataset. On the RAVDESS dataset, the highest classification accuracy attained was 87.71% and on the KBES corpus, the highest classification accuracy attained was 71.67%.

Table 8. Model performs compared to the current models on RAVDESS dataset

Work Ref. Year	Feature Transformation	Classification Model	Accuracy (%)
Zhao et al., 2019	Log-Mel Spectograms	4 CNN + LSTM (Single DL)	75.43
		4 CNN + LSTM (Cascaded DL)	77.48
Mustaqeem and Kwon, 2020	Log-Mel Spectograms	7 CNN + 2 FC (Single DL)	67.25
		7 CNN + 2 FC (Cascaded DL)	72.51
Sultana et al., 2021	Log-Mel Spectograms	4 CNN + TDF+Bi-LSTM (Single DL)	80.12
		4 CNN + TDF+Bi-LSTM (Cascaded DL)	86.55
Islam et al., 2022	MFCC + STFT + Chroma STFT	4 3D-CNN+TDF+Bi-LSTM (Single DL)	78.65
		4 3D-CNN+TDF+Bi-LSTM (Cascaded DL)	87.71
Our Work	MFCC +Chroma	Late-Fusion RF and XGBoost	93.0

Table 9. Model performs compared to the current models on KBES dataset

Work Ref. Year	Feature Transformation	Classification Model	Accuracy (%)
Sultana et al., 2022	Log-Mel Spectograms	4 CNN+TDF+Bi-LSTM (Single DL)	57.22
		4 CNN+TDF+Bi-LSTM (Cascaded DL)	67.78
Islam et al., 2022	Log-Mel Spectograms	4 3D-CNN+TDF+Bi-LSTM (Single DL)	58.89
		4 3D-CNN+TDF+Bi-LSTM (Cascaded DL)	69.44
Masum et al., 2024	Log-Mel Spectograms	4 3D-CNN+TDF+Bi-LSTM+LSTM (Single DL)	61.11
		4 3D-CNN+TDF+Bi-LSTM+LSTM (Cascaded DL)	71.67
Our Work	MFCC +Chroma	Late-Fusion RF and XGBoost	93.0

In our study, we used the RAVDESS benchmark dataset by integrating the MFCC and Chroma features on a model build from two base models of XGBoost and Random Forest in a Late-Fusion approach. The proposed model achieved a classification accuracy of 93% and compared to the state-of-the-art works, there was an improvement of 5.29% against the results of the research works benchmarking on RAVDESS dataset while an improvement of 21.33% classification accuracy was attained in respect to the research works benchmarking on the KBES dataset.

6. Conclusions and Future Work

Table 10 displays the comparison of the base models Vis a Vis the proposed model. XGBoost attained a classification of 68%, RF posted 87% classification accuracy, while our proposed model attained a 93% classification accuracy. More notable also is the number of support features from the base models and the proposed model. The total support features of the base models are 135 each but when combined, the proposed model support features increased to 291. Fusing models comes with some advantages, especially in dealing with complex classification such as emotion intensity classification. XGBoost took care of the structured data and complex interactions while RF took care of overfitting and high-dimensional data. The deep architecture of the MLP Classifier took care of non-linear patterns in the dataset. For class balancing, the ADASYN (Adaptive Synthetic) sampling technique allowed the generation of synthetic samples to take care of the minority classes as per the density distribution. The complexity of the model, overfitting and performance was taken care of by the Recursive Feature Elimination (RFE) technique in conjunction with a RandomForestClassifier by removing non-essential features and selecting the top n_features. Finally, the blended predictions with majority voting improved the robustness and the overall predictive performance by merging predictions from the base models. This is achieved by using simple majority voting in selecting the most commonly predicted class from the base models.

Based on the analysis done on the results from the experimental setups, the following inferences can be argued. The lesser the classes to be classified, the higher the accuracy achieved by the model and the more the classes to be classified, the lower the accuracy achieved. The results from our first experimental setup yielded a 38% and 41% classification accuracy from XGB and RF models respectively. This was attributed to the many classification fields using fewer support features as compared to the results of the subsequent experimental setups where, the classification fields decrease.

Secondly, the model performance is highly influenced by the choice of machine learning algorithm in use. The number of support elements plays a crucial role in the performance of the models. This suggests that the determination of emotion intensities requires a large and well-balanced dataset for good results to be achieved. It requires a multi modal approach in order to achieve the best results [16]. This is informed by the many similar characteristics exhibited by each individual emotion. Emotions can be discreet or continuous [17]. They are also determined by the band width and frequency spectrum range where they are found.

This research opens a field of study for further research on the field of emotions. We performed individual detection and classification with two classification models on the dataset in our experiment. We further combined the best results of the two base models, in a late-fusion stacking approach to determine the emotion intensities. We looked at the discreet category of emotions and classified the intensities based on the valence of the emotion category. Further research on the continuous categories of emotions needs to be looked at. Parallel and serial models with two or more classifiers were other available options of determining emotion intensities. The use of the state-of-the-art machine learning models coupled with transfer learning and a different dataset can greatly improve the performance of the models used. This will be explored in our future work on this study.

Table 10. Comparison of the performance of the proposed model with the performance of the base models

Emotion Category	Precision			Recall			F1-Score			Support		
	XGB	RF	Proposed model	XGB	RF	Proposed model	XGB	RF	Proposed model	XGB	RF	Proposed model
Happy_normal	0.71	0.73	0.93	0.53	0.58	0.98	0.61	0.65	0.96	19	19	55
Happy_strong	0.76	0.89	0.91	0.68	0.84	1.00	0.72	0.86	0.95	19	19	62
Neutral_normal	0.76	0.88	0.97	0.95	0.90	0.97	0.85	0.89	0.97	58	58	58
Surprise_normal	0.47	0.91	0.90	0.35	1.00	0.90	0.40	0.95	0.90	20	20	60
Surprise_strong	0.41	0.86	0.97	0.37	0.95	0.80	0.39	0.90	0.87	19	19	56
Accuracy							0.68	0.87	0.93	135	135	291
Macro avg	0.62	0.85	0.93	0.58	0.85	0.93	0.59	0.85	0.93	135	135	291
Weighted avg	0.66	0.86	0.93	0.68	0.87	0.93	0.66	0.86	0.93	135	135	291

References

[1] Human Emotions and their manifestation: https://www.all-about-psychology.com/human-emotions-and-their-manifestation.html/ accessed 4th December, 2024

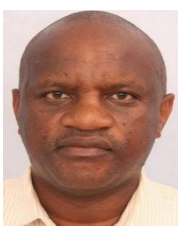
[2] Kendra Cherry. "*Basic Emotions and Their Effect on Human Behaviour*". https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976, (2019).

[3] Poria S., Majumder N., Mihalcea R., and Hovy E., "Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances". Michigan Institute for Data Science, US, 2019.
[4] Udahemuka, G., Djouani, K., & Kurien, A. M., "Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals": A Review. *Applied Sciences*, *14*(17), 2024.
[5] Joshi S. & Joshi F., "*Human Emotion Classification Based on EEG Signals Using Recurrent Neural Network and KNN"* DO - 10.48550/arXiv.2205.08419 ER –. (2022).
[6] Md Shad Akhtar, Asif Ekbal Erik Cambria: How Intense Are You? Predicting Intensities of Emotions and Sentiments Using Stacked Ensemble. https://sentic.net/predicting-intensities-of-emotions-and-sentiments.pdf Accessed on 24th December, 2024
[7] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks," IEEE Access, vol. 10, 2022.
[8] Md. Masum Billah, Md. Likhon Sarker, M. A. H. Akhand, Md Abdus Samad Kamal, "Emotion Recognition with Intensity Level from Bangla Speech using Feature Transformation and Cascaded Deep Learning Model". International Journal of Advanced Computer Science and Applications, (IJACSA) Vol. 15, No. 4, 2024.
[9] Islam, M.R.; Akhand, M.A.H.; Kamal, M.A.S.; Yamada, K., "Recognition of Emotion with Intensity from Speech Signal Using 3D Transformed Feature and Deep Learning". Electronics 2022, 11, 2362.
[10] Mustaqeem, Kwon S., "CLSTM Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network". *Mathematics*. 2020; 8(12):2133. https://doi.org/10.3390/math8122133
[11] Zhao, Z.; Li, Q.; Zhang, Z.; Cummins, N.; Wang, H.; Tao, J.; Schuller, B.W., "Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition". Neural Netw. 2021, 141, 52–60.
[12] Phan, Thi-Thu-Hong & Nguyen Doan, Dong & Nguyen, Huu-Du & Nguyen, Van & Pham-Hong, Thai. (2022). Investigation on New Mel Frequency Cepstral Coefficients Features and Hyper-parameters Tuning Technique for Bee Sound Recognition. Soft Computing. 10.1007/s00500-022-07596-6
[13] Alang Rashid, N. K., Alim, S. A., & Hashim, S. W. (2017). Receiver operating characteristics measure for the recognition of stuttering dysfluencies using line spectral frequencies. IIUM Engineenng Joumal, 18(1), 193-200.
[14] Ayush Kumar Shah & Araju Nepal: Chroma Feature Extraction, Department of Computer Science and Engineering, School of Engineering Kathmandu University, Nepal: https://www.academia.edu/42216949/Chroma_Feature_Extraction Accessed on 24th December, 2024
[15] Livingstone, S. R., & Russo, F. A. (2018). Retlieved July 2, 2022, from). The Ryerson audiovisual database of emotional speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north Amelican English. PLoS One, 13(5), e0196391. https:// doi.org/I O. 1371 /journal.pone.() 196391
[16] Patricia J. Bota, Chen Wang, Ana L. N. Fred and Hugo Placido da Silva (2019): A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. IEEE Open Access Journal.
[17] Liu D, Wang Z, Wang L and Chen L (2021): "*Multi-Modal Fusion Emotion Recognition Method of Speech Expression Based on Deep Learning*". School of Information Engineering, Shandong Youth University of Political Science, Jinan, China

Authors' Profiles

Simon Kipyatich Kiptoo: the Corresponding Author is a student at the Jomo Kenyatta University of Agriculture and Technology, pursuing a Masters' of Science degree in Computer Systems. with interest in Artificial Intelligence and computer systems.

Dr. Kennedy Ogada: PhD, co-Author, is Lecturer at Jomo Kenyatta University of Agriculture and Technology (JKUAT), Faculty of Pure and Applied Science, Department of Computing with Interests in Project Management, Penetration Testing, Quality Assurance, IT Audit and Strategy Development.



Tobias Mwalili: PhD, co-Author, is a Lecturer of Computer Science at Jomo Kenyatta University of Agriculture and Technology (JKUAT), Faculty of Pure and Applied Science, Department of Computing in Nairobi, Kenya, with a PhD, and his research focuses on areas like component-based systems and software quality.

How to cite this paper: Simon Kipyatich Kiptoo, Kennedy Ogada, Tobias Mwalili, "Determining Emotion Intensities from Audio Data Using Ensemble Models: A Late Fusion Approach", International Journal of Intelligent Systems and Applications(IJISA), Vol.17, No.6, pp.44-57, 2025. DOI:10.5815/ijisa.2025.06.04