

# Classifier model for lecturer evaluation by students using speech emotion recognition and deep learning approaches

Yesy Diah Rosita<sup>1,2</sup>, Wahyu Andi Saputra<sup>2</sup>

<sup>1</sup>Center of Excellence for Human Centric Engineering, Institute of Sustainable Society, Telkom University, Bandung, Indonesia

<sup>2</sup>Informatics Engineering Study Program, School of Computing, Telkom University, Purwokerto, Indonesia

## Article Info

### Article history:

Received Jul 31, 2024

Revised Sep 10, 2025

Accepted Oct 16, 2025

### Keywords:

Bi-LSTM

Energy

Evaluation

Lecturer

MFCC

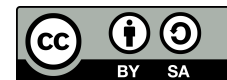
Student

Zero-crossing rate

## ABSTRACT

Lecturers play a crucial role in higher education, with their teaching behavior directly impacting learning and teaching quality. Lecturer evaluation by students (LES) is a common method for assessing lecturer performance, though it often relies on subjective perceptions. As a more objective alternative, speech emotion recognition (SER) uses speech technology to analyze emotions in the speech of lecturers during classes. This study proposes using deep learning-based SER, including convolutional neural network (CNN) and bidirectional long short-term memory (Bi-LSTM), to evaluate teaching quality by analyzing displayed emotions. Removing silence from audio signals is crucial for enhancing feature analysis, such as energy, zero-crossing rate (ZCR), and mel-frequency cepstral coefficients (MFCC). This method removes inactive segments, emphasizing significant segments, and improving accuracy in detecting voice and emotions. Results show that the 1D CNN model with Bi-LSTM, using MFCC with 13 coefficients, energy, and ZCR, performs excellently in emotion detection, achieving a validation accuracy of over 0.851 with an accuracy gap of 0.002. This small gap indicates good generalization and reduces the risk of overfitting, making teaching evaluations more objective and valuable for improving practices.

This is an open access article under the [CC BY-SA](#) license.



## Corresponding Author:

Yesy Diah Rosita

Informatics Engineering Study Program, School of Computing, Telkom University

St. D.I. Panjaitan No. 128, Purwokerto, Banyumas, Central Java-53147, Indonesia

Email: yesydr@telkomuniversity.ac.id

## 1. INTRODUCTION

Lecturers play a crucial role in higher education, where their teaching behavior directly impacts the learning process and ultimately determines the quality of education provided. This role is vital as the quality of teaching affects students' learning experiences and their academic outcomes. To ensure that teaching standards remain high, many higher education institutions have implemented lecturer evaluation by students (LES) systems to assess lecturer performance during classes [1]. These evaluations typically cover aspects such as lecturer discipline, subject mastery, and their interactions with students.

LES is presented in the form of a questionnaire that students complete at the end of the semester. This questionnaire aims to provide an overview of the teaching quality delivered by lecturers, and the results of this evaluation impact the course grades listed on students' transcripts [2]. However, this method tends to be subjective because the assessment is based on each student's personal perception, which can be influenced by factors such as mood, personal experiences, or individual interactions with the lecturer. Consequently, the

results of LES may not fully reflect the objective quality of teaching and are often inadequate as a sole measure of lecturer performance.

As an alternative for a more objective assessment of teaching quality, emotion analysis-based approaches can be utilized. One promising method is speech emotion recognition (SER), which leverages speech recognition technology to analyze emotions [3], [4] present in lecturers' speech during classes. SER relies on extracting features from audio speech signals to determine the types of emotions expressed by lecturers. This technology offers potential for a more objective evaluation since the emotions captured in speech can provide deeper insights into the lecturer's mood and attitude while teaching. Previous research indicates that emotions can generally be categorized into three classes: positive, negative, and neutral [5]. Using SER in this context allows for a more holistic assessment of how lecturers display their emotions during teaching. By identifying feature extraction patterns and appropriate model configurations, SER can provide accurate data on the percentage of emotions expressed by lecturers throughout a class session. This paves the way for a more objective evaluation method that relies not only on students' subjective perceptions but also on empirical data generated from audio analysis.

In the context of technological development, the use of deep learning has become an increasingly popular approach in SER. Deep learning algorithms, particularly deep neural networks, can process and analyze feature data more effectively than conventional methods. Convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have proven highly efficient in recognizing patterns in speech and emotion data [6], [7]. The application of these techniques in SER enables enhanced accuracy and the model's ability to understand more complex emotional contexts. The combination of SER and deep learning offers an innovative solution for lecturer evaluation. By integrating emotion analysis technology with deep learning algorithms, we can gain deeper insights into teaching quality and classroom atmosphere. This approach not only enhances accuracy in assessment but also provides more valuable data for continuous improvement in teaching practices.

## 2. METHOD

The objective of this study is to evaluate the performance of a deep learning model capable of classifying lecturer performance in delivering lecture material through SER. In this context, lecturers' emotions are classified into three classes: positive (happy and surprised), neutral, and negative (angry and sad). The methodology involves several stages: data collection, preprocessing, feature extraction, model creation, and performance evaluation.

### 2.1. Data collection

The data consists of speech samples in Indonesian, totaling 1,600 samples with a duration of 3-5 seconds: 491 positive (250 happy and 241 surprised), 619 negative (337 angry and 282 sad), and 400 neutral. The audio files are in .wav format and mono channel. Data was collected using a clip-on wireless microphone placed on the respondent's chest to ensure stable recording. The equipment features include: up to 100 m wireless operating range, selectable mono/stereo output mode, 3.5 mm headphone jack for real-time monitoring, built-in omnidirectional microphone for 360° sound pickup, and compatible with smartphones, tablets, cameras, recorders, or other audio/video recording devices. The same equipment was used to record lecturers during their presentations, with audio samples lasting approximately 30-60 seconds for emotion analysis. A total of 30 samples were collected, corresponding to the number of active lecturers in the School of Computing, Telkom University. To provide a clearer picture of the dataset composition, Table 1 summarizes the distribution of emotion classes.

Table 1. Emotion class distribution in the dataset

Emotion	Category	N
Happy	Positive	250
Surprised	Positive	241
Neutral	Neutral	400
Angry	Negative	337
Sad	Negative	282

## 2.2. Preprocessing

This stage aims to obtain audio data with voice activity by applying a threshold of 0.001. Previous research often removed silence only from the beginning and end of speech data [8], but in this study, segments with values below the threshold are removed throughout the entire recording, including the beginning, middle, and end. Figure 1 provides a visual comparison between the original audio signal input and the signal after silence removal. The silence removal technique was implemented using the Librosa library in Python, which is widely adopted for audio processing tasks due to its flexibility and ease of integration. In this study, an amplitude threshold of 0.001 was applied to distinguish between speech and non-speech segments. Segments with amplitude values below this threshold were considered silent and thus excluded from further analysis. The selection of the 0.001 threshold was not arbitrary. It was informed by prior research, which demonstrated that such a value effectively removes low-energy, non-informative segments while preserving the essential speech content necessary for reliable feature extraction and classification.

Figure 1(a) illustrates a portion of the audio signal where there is no evident voice activity. The amplitude remains consistently close to zero, clearly indicating the presence of silence, as defined by the 0.001 threshold. This segment does not contribute meaningful acoustic features and, therefore, is identified for removal [9]. As a result, Figure 1(b) displays the modified signal after the silence has been removed, showcasing only the relevant speech segments retained for further processing. This preprocessing step is crucial in enhancing the quality of input data, reducing noise, and improving the performance of subsequent feature extraction and classification stages in SER systems.

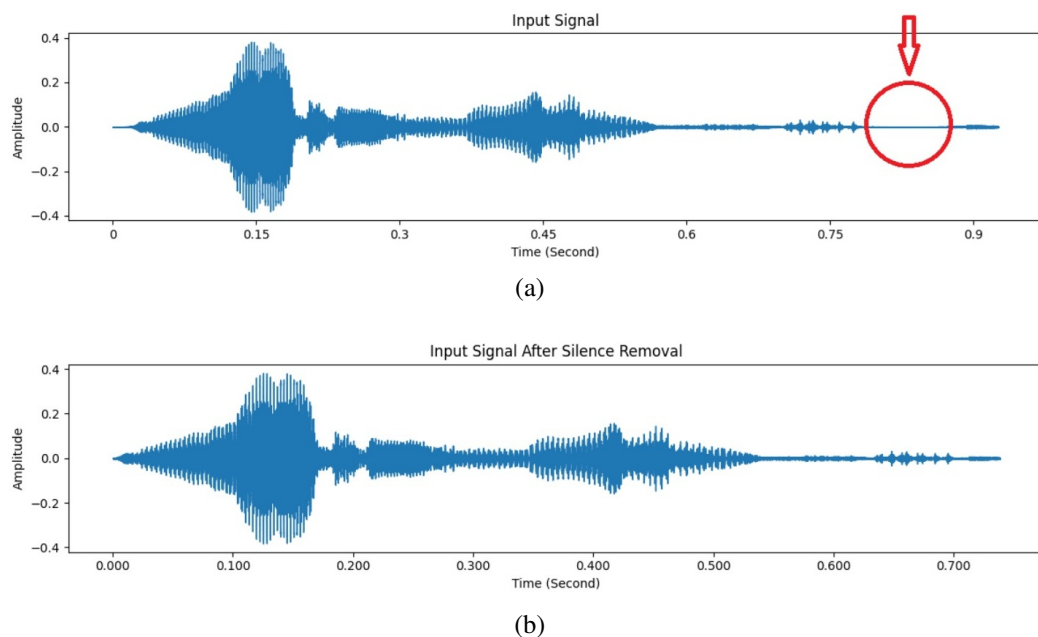


Figure 1. The difference in signal: (a) without silence removal and (b) with silence removal

## 2.3. Feature extraction

This stage aims to obtain audio data with voice activity by applying a threshold of 0.001. Previous research often removed silence only from the beginning and end of speech data [8]. Still, in this study, segments with values below the threshold are removed throughout the entire recording, including the beginning, middle, and end. The next stage involves feature extraction, which includes three types. First, mel-frequency cepstral coefficients (MFCC) [10] with varying coefficients (12 coefficients as in [11]–[13]; 13 coefficients as in [14]; 40 coefficients as in [15]–[17], combined with energy and zero-crossing rate (ZCR) [4], [14], [18].

Additionally, comparisons are made with combinations of MFCC coefficients, Chroma [18]–[21], and mel-spectrogram [18], [21]. This results in 40 feature combinations for model development. These dynamic features enhance sensitivity to temporal changes in speech, which can signal emotional transitions. This stage reveals the characteristics of the voice from various perspectives and assesses the performance of

each characteristic. The energy after silence removal is usually higher compared to the original signal energy, primarily because quiet or silent parts are removed, leaving only the louder or voice-containing sections. However, if only the silent parts are removed, the total energy may not change significantly, but the energy distribution per frame might. Similarly, with the ZCR feature, silence in the original signal may contain small fluctuations that cause zero-crossings. When silence is removed, these fluctuations disappear, resulting in a lower ZCR. After silence removal, the remaining parts may be more consistent or stable, meaning fewer rapid changes crossing zero, leading to a decrease in ZCR.

Like ZCR, silence in the original signal can also affect the spectral representation captured by MFCC. MFCC is a crucial feature in voice signal analysis used to capture rich spectral information. When silence is removed, MFCC analysis becomes more focused on the relevant parts of the voice, improving accuracy in recognizing voice patterns and emotions. By removing silence, we eliminate segments that do not carry important information, making the resulting MFCC more representative of the true characteristics of the voice. Visualization of MFCC before and after silence removal will show differences in spectral representation, where MFCC after silence removal will be more stable and reflect clearer and more consistent voice patterns. The MFCC feature also shows significant changes after silence removal. MFCC is an important representation in voice signal analysis that captures rich spectral information. When silence sections are removed, MFCC analysis becomes more focused on the relevant voice parts, potentially enhancing accuracy in recognizing voice patterns and emotions. Removing silence eliminates segments that do not provide important information, resulting in MFCC that is more representative of the true characteristics of the voice.

### 2.3.1. Energy

It is one of the most fundamental acoustic features in SER. It quantifies the overall strength or power of the speech signal in the time domain, reflecting how loudly or forcefully a person is speaking. Vocal intensity, captured by energy, often corresponds with emotional arousal and activation levels: for instance, high-arousal emotions like anger, joy, or fear tend to be expressed with greater energy, while low-arousal states like sadness or boredom result in quieter speech. Many studies in SER therefore incorporate energy as a reliable indicator of emotional expression, and frequently apply statistical functionals (e.g., mean, variance, and extremes) over energy contours to characterize emotion [19], [22], [23]. These temporal-energy patterns help classifiers distinguish between high-intensity emotional states and more subdued expressions, enhancing the overall robustness and performance of emotion detection systems.

Signal energy is a fundamental measure that quantifies the total power contained within an audio signal over time. It reflects how ‘strong’ or ‘loud’ the signal is, which is essential for tasks like voice activity detection and emotion analysis. The signal input represents the amplitude of the signal at the samples and the total number of samples in the frame. By squaring the amplitude, we ensure that both positive and negative values contribute positively to the total energy, thereby providing an accurate measure of signal strength. This discrete-time definition is widely used in audio processing due to its simplicity and computational efficiency.

In practice, the integral is approximated by summing over finite-duration frames, as shown above, because real-world signals are finite. This form is based on continuous-domain theory but is rarely used directly in digital signal processing due to discretization. Signal energy correlates with perceived loudness, though loudness perception is more complex and frequency-dependent. This conversion allows audio engineers to handle very large variations in signal energy more conveniently, aligning more closely with human perception. Signal energy is a core feature in emotion recognition systems since more intense vocal expressions (like anger or excitement) exhibit higher energy, whereas calmer speech (like sadness) tends to have lower energy. In feature extraction pipelines, energy is often used alongside MFCC and ZCR to provide a more holistic representation of the emotional content of speech.

Figure 2 visualizes the energy feature without silence removal as shown in Figure 2(a), and with silence removal as presented in Figure 2(b) by plotting a short-time energy contour directly beneath the raw waveform, with time on the x-axis and energy magnitude on the y-axis. This representation clearly highlights where speech segments occur, peaks correlate with voiced, high-intensity speech, while valleys indicate silence or quieter, low-arousal states like sadness or boredom. This visualization is especially useful for voice activity detection and emotion analysis: the temporal patterns when energy spikes or dips, help in characterizing emotional states over time. Typically, a sliding window of 10–30 ms (e.g., 160–320 samples at 16 kHz) is used to balance time resolution and smoothing of rapid amplitude changes.

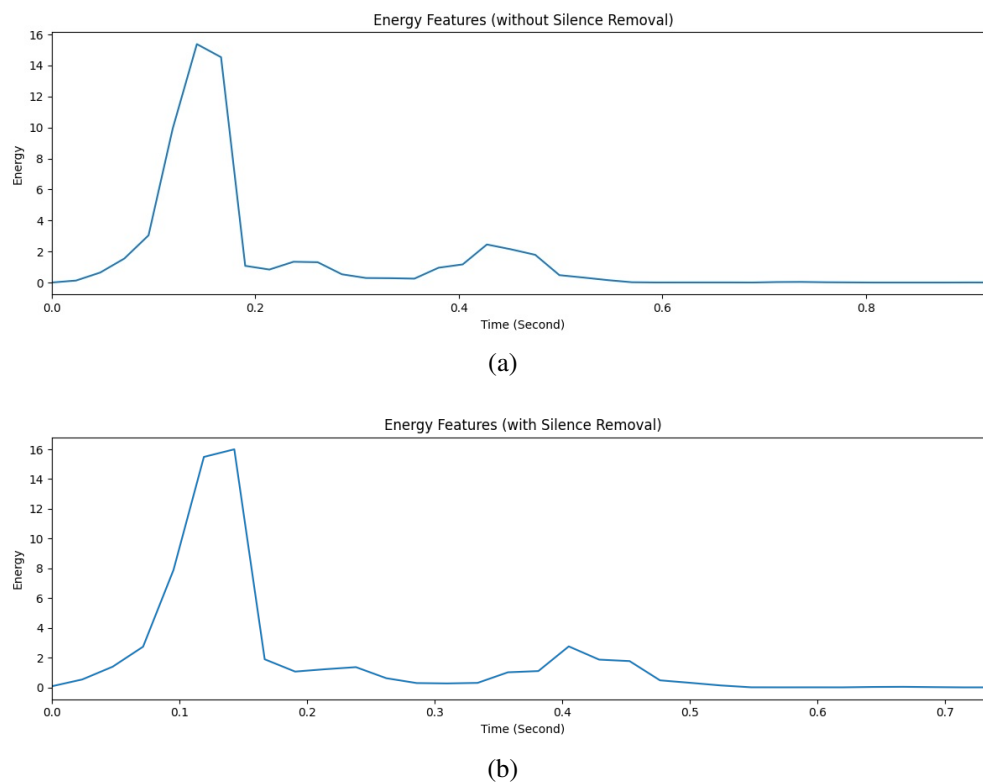


Figure 2. The difference in energy: (a) without silence removal and (b) with silence removal

### 2.3.2. Chroma

It is chosen because they capture harmonic and pitch-related information—attributes that go beyond what typical spectral features (like MFCC or ZCR/energy) represent. By encoding the distribution of energy across the twelve pitch classes, chroma features reveal tonal characteristics and musicality within speech, such as subtle pitch modulations, intonation patterns, and harmonic structure, that are often linked to the expression of emotions [24]. In fact, studies show that adding chroma to traditional feature sets consistently improves SER performance: for example, they notably contribute to emotion discrimination across datasets like RAVDESS and TESS, helping models distinguish emotional nuances that would be missed by MFCC alone. However, chroma's performance can vary depending on the dataset and emotional content, and it still requires further tuning, like combining chroma with temporal or rhythmic context, to reach optimal accuracy in emotion classification tasks [24].

Chroma features represent the spectral energy distribution in the twelve musical pitch classes (C, C/D,..., B), which aggregates octave-independent pitch information. They are particularly valuable in audio analysis tasks, such as emotion recognition in speech, because they capture harmonic and tonal characteristics while being invariant to timbre, instrumentation, and octave shifts. This makes them robust descriptors for capturing pitch-related variations in spoken utterances.

Figure 3 demonstrates that the resulting chromagram without silence removal (Figure 3(a)) and with silence removal (Figure 3(b)) are the two-dimensional time–chroma matrix showing how the spectral content is distributed across pitch classes over time. This structure is highly effective at summarizing harmonic content, as notes with identical pitch class but different octaves contribute to the same bin, preserving musical color regardless of octave. Such octave invariance also ensures chroma features remain stable under pitch shifts or speaker variations.

Chroma features are robust to changes in timbre and dynamic range since they focus on pitch-class patterns rather than exact spectral shapes. In emotion analysis, this helps capture intonational melodies and pitch modulations associated with affective speech, even amidst background noise or speaker variability. Augmentations like harmonic pitch class profiles (HPCP) further enhance robustness by tuning alignment and energy normalization across octaves.

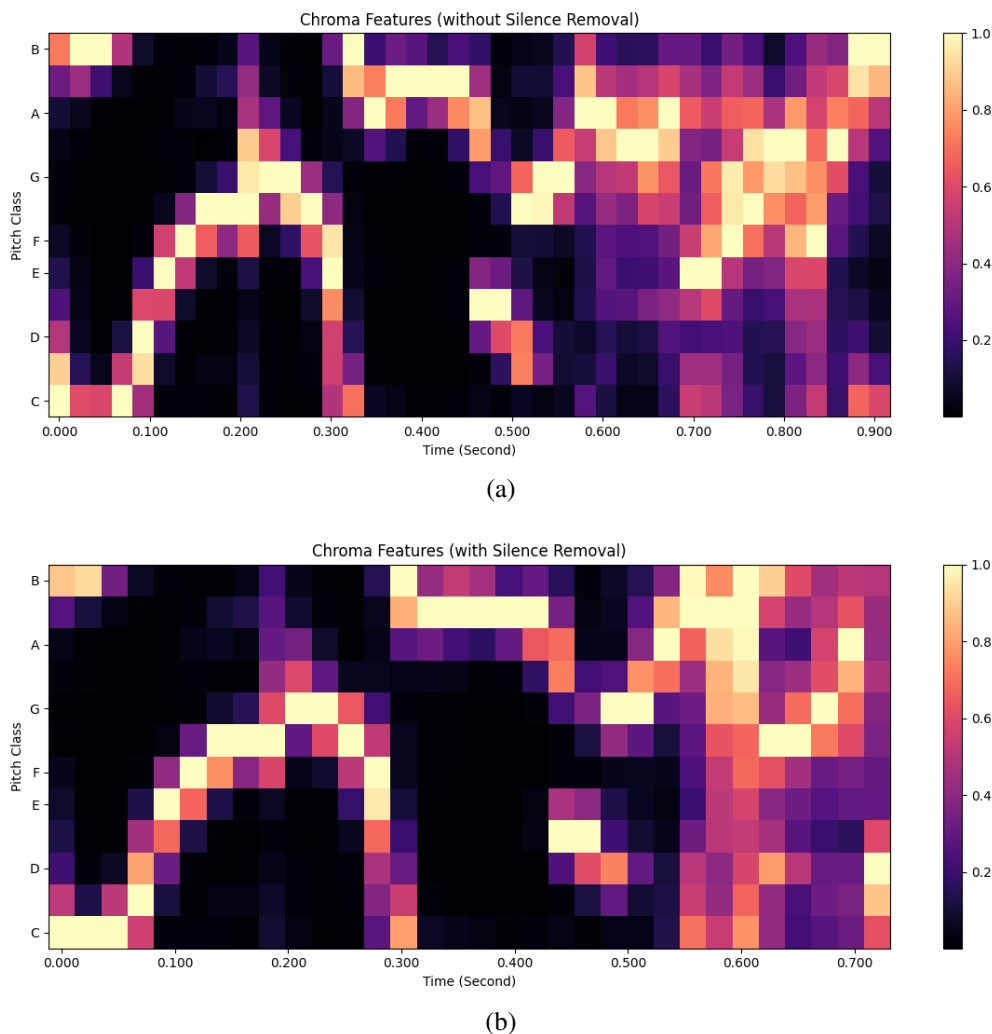


Figure 3. The difference in chroma: (a) without silence removal and (b) with silence removal

### 2.3.3. Mel-frequency cepstral coefficients

It is chosen because they effectively encode the physical characteristics of sound signals by simulating human auditory perception: they apply a mel-scale filter bank that emphasizes frequencies in a way humans perceive, take logarithmic compression to resemble loudness perception, and perform a discrete cosine transform to decorrelate filter outputs into compact coefficients. This structure enables MFCCs to extract phonetic content that is particularly valuable for emotion classification: prior work has shown that even a modest number of MFCC features [15], [10], [25] carry significant emotion discrimination power by capturing spectral variations tied to vocal tract dynamics. By forming a distinct feature map, these coefficients allow machine learning models to differentiate subtle emotional cues in speech, making MFCCs an effective choice for emotion recognition tasks. Figure 4 visualizes the characteristics of an audio signal, which is typically represented as a heatmap, a time-series visualization of MFCC coefficients without silence removal (Figure 4(a)) and with silence removal (Figure 4(b)). On the x-axis of the heatmap is time, while the y-axis shows the cepstral coefficient indices (e.g., MFCC 1–13). Each cell in the heatmap represents an amplitude value, darker or lighter depending on the color palette, corresponding to a specific time and coefficient index.

### 2.3.4. Mel-spectrogram

It is used because it provides a frequency representation on the mel scale with both time and frequency dimensions, which are suitable for processing with 2D convolutional kernels in deep learning models. By converting audio signals into image-like spectrograms, CNNs can effectively learn localized time-frequency

patterns, such as energy bursts, formant shifts, and pitch contours that are strongly associated with different emotional states. Recent research demonstrates that feeding mel-spectrograms into CNN architectures enables models to autonomously extract salient emotional cues, leading to improved classification performance compared to conventional approaches [26].

A mel-spectrogram is a perceptually motivated time–frequency representation of audio, widely used in speech and emotion recognition. It aligns with how humans perceive sound by emphasizing lower frequencies and compressing higher bands. As a result, it produces an image like matrix well-suited for deep learning applications, especially CNNs. The log transformation compresses the dynamic range, mimicking human loudness perception. Adding a small constant  $\epsilon$  prevents taking the log of zero. The result yields a stable feature representation for deep learning.

Figure 5 shows a heatmap of mel-spectrogram without silence removal (Figure 5(a)) and with silence removal (Figure 5(b)) that with time on the x-axis and mel-frequency (in Hz on the mel scale) on the y-axis, where color intensity represents magnitude in decibels (dB). Brighter bands on the heatmap indicate regions of high energy at specific frequencies and times, such as formant resonances or pitch harmonics, while darker areas show quieter portions. This image-like representation enables deep learning models, particularly 2D CNNs, to detect localized time-frequency patterns, such as energy bursts or frequency shifts, associated with emotional cues. The decibel scale (log amplitude) ensures the dynamic range is visually compressed, making both subtle and prominent audio features apparent.

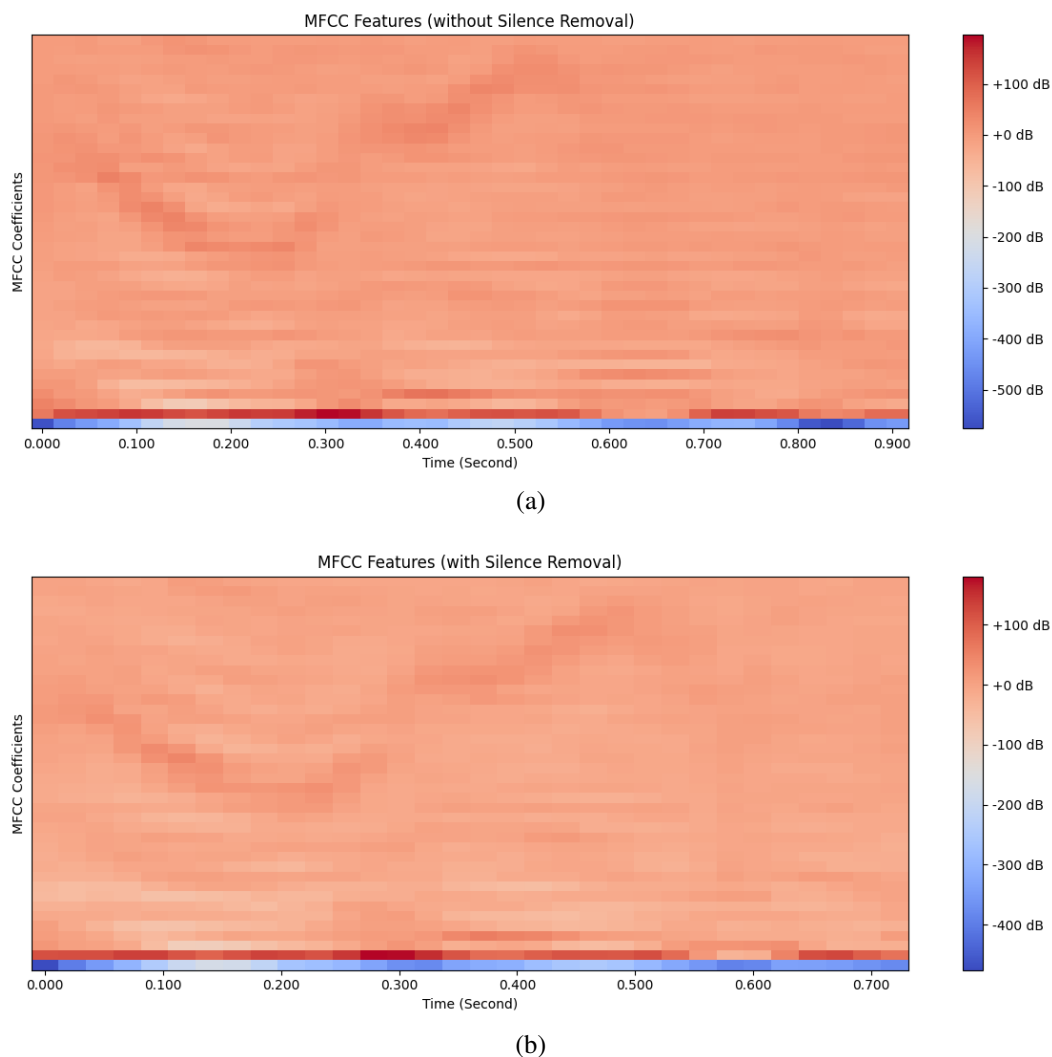


Figure 4. The difference in MFCC: (a) without silence removal and (b) with silence removal

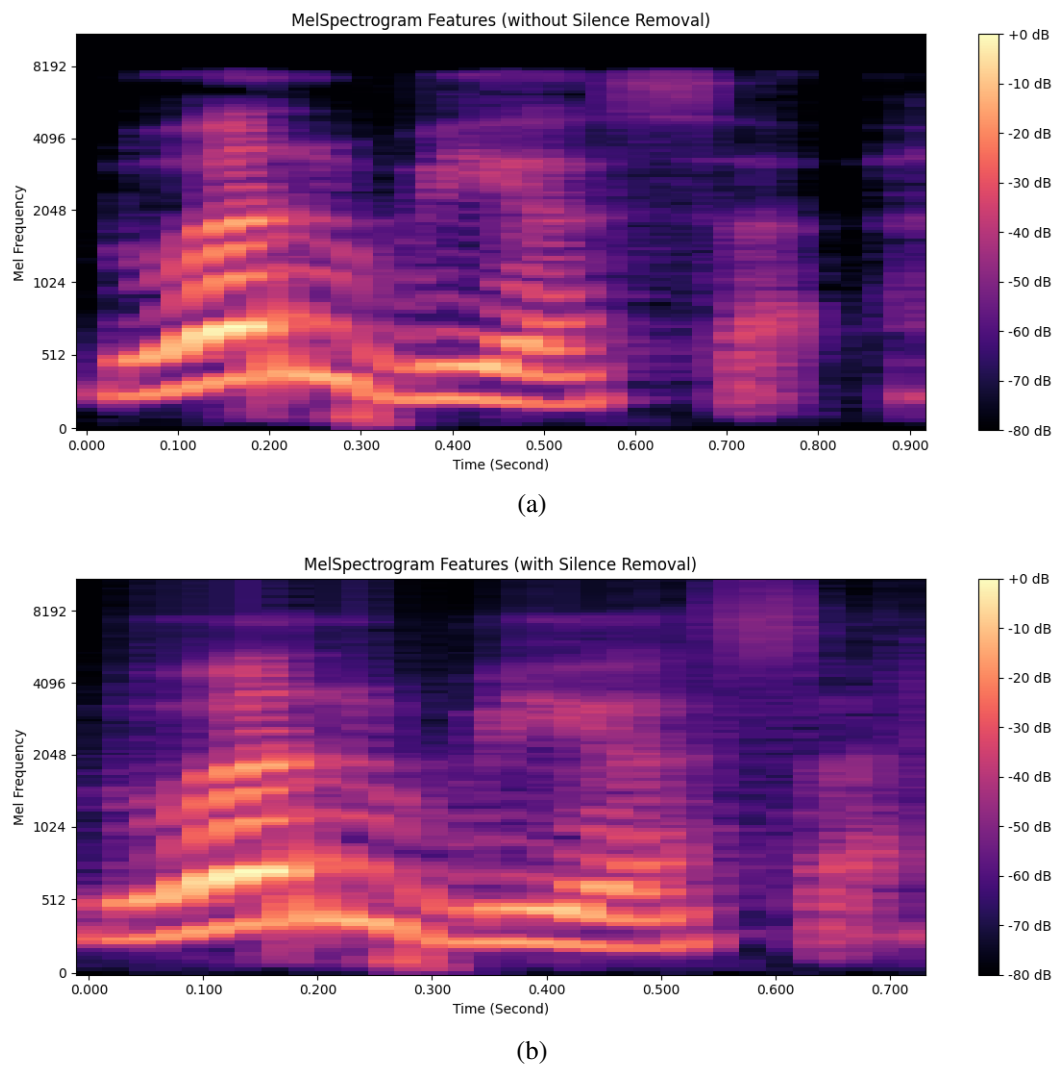


Figure 5. The difference in mel-spectrogram: (a) without silence removal and (b) with silence removal

The final mel-spectrogram matrix  $\{\tilde{S}_m(t)\}_{m=1,\dots,M}^{t=1,\dots,T}$  serves as an efficient and perceptually aligned input to 2D convolutional kernels. It preserves time–frequency locality, enabling neural networks to detect emotional cues such as formant shifts, pitch contours, and energy bursts. By combining mel-scale filtering, log compression, and spectral smoothing, mel-spectrograms outperform linear spectrograms in capturing emotive vocal patterns, making them ideal for emotion recognition architectures.

### 2.3.5. Zero-crossing rate

This feature measures the smoothness of the audio signal by counting how frequently it changes sign, crossing from positive to zero to negative, or vice versa, within a given time frame [27]. Also known as the number of zero-axis crossings per unit time [4], ZCR effectively captures the noisiness or smoothness of the signal: noisy or unvoiced segments typically exhibit higher ZCR, while voiced and more periodic regions yield lower values. Due to its clear association with spectral content, higher ZCR indicates richer high-frequency components, and lower values align with more periodic, low-frequency sounds. ZCR is widely used in voice activity detection, voiced/unvoiced frame classification, and even as an excitation level indicator in emotion recognition systems. Combined with features like energy and MFCCs, ZCR enhances spectral representations by providing insights into speech articulation dynamics and intensity fluctuations.

The Figure 6 ZCR visualization without silence removal (Figure 6(a)) and with silence removal (Figure 6(b)) that each of both using a dual-plot layout: the top plot displays the raw audio waveform



(amplitude vs. time), while the bottom plot shows the short-time ZCR over the same time axis. Peaks in the ZCR curve correspond to rapid sign changes, common during unvoiced sounds or noisy segments, while troughs align with voiced regions where the waveform oscillates smoothly and crosses zero less often. This visual alignment allows researchers to immediately identify voiced/unvoiced segments and associate sudden fluctuations with phonetic or emotional cues. Because ZCR is calculated per frame (e.g., 10–30 ms windows), the contour's temporal resolution effectively highlights dynamic speech features critical for emotion detection and voice activity tasks.

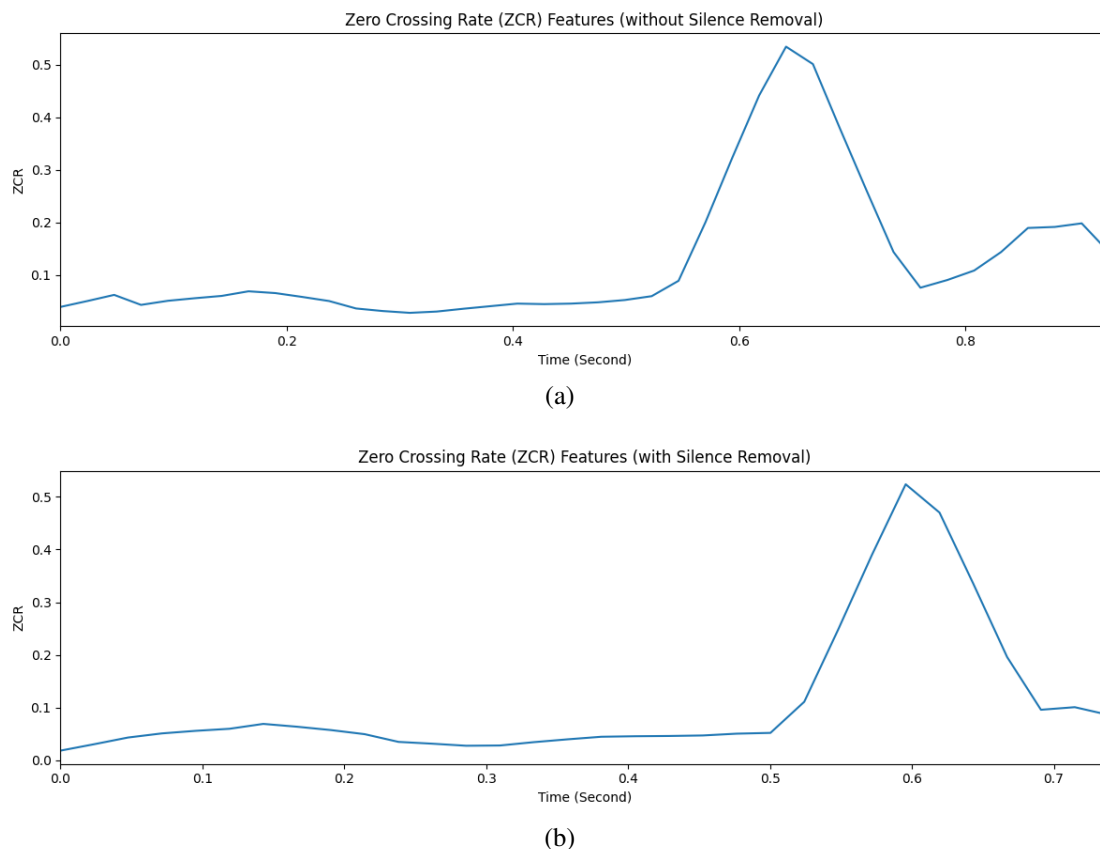


Figure 6. The difference in ZCR: (a) without silence removal and (b) with silence removal

Furthermore, in Figure 6, the ZCR contour is often depicted alongside a horizontal threshold line that classifies frames as voiced or unvoiced. Frames whose ZCR exceeds the threshold are marked as unvoiced (typically shown in one color), while those below are considered voiced (shown in another). This threshold-based segmentation is validated by prior work demonstrating that unvoiced segments generally exhibit higher ZCR and lower energy compared to voiced segments, where ZCRs are low and energies are high. Such delineation enables automated voice activity detection and helps the model focus on emotionally rich voiced regions. Moreover, the sharp contrast in ZCR trends between voiced and unvoiced regions offers visual cues about changes in speech excitation peaks in the ZCR curve often align with phonetic transitions or bursts, which can be critical indicators of emotional states or emphatic speech patterns.

#### 2.4. Model architecture

Experiments were conducted with several types of models, including CNN-1D, LSTM, bidirectional long short-term memory (Bi-LSTM), combinations of CNN-1D and LSTM, and CNN-1D and Bi-LSTM. Each model was tested with 8 feature extraction results from the previous stage, resulting in 40 model scenarios. The summary of model types and their respective layer compositions is presented in Table 2.

In this model, the frame size is derived from the audio's frame duration multiplied by the sample rate. At a standard sample rate of 22,050 Hz, using  $n_{\text{fft}}=2048$  results in each frame spanning approximately

93 ms, as Librosa applies an FFT window of that size by default. Meanwhile, a hop length of 512 samples is employed, which leads to approximately 75% overlap between successive frames. In practical terms, this configuration yields analysis shifts of roughly 23 ms per frame, promoting smoother temporal transitions during feature extraction.

The dataset is divided into 80% for training and 20% for testing (test\_size=0.2). Training is performed for up to 50 epochs, with early stopping enabled via EarlyStopping(monitor= 'val\_accuracy', patience=5 to halt training if validation accuracy does not improve over five consecutive epochs. We utilize the Adam optimizer, combined with the categorical\_crossentropy loss function and an initial learning rate of 0.001. The model is trained using a batch size of 32. Activation functions include rectified linear unit (ReLU) in the convolutional and dense hidden layers, and Softmax in the output layer for multi-class classification.

Table 2. The summary of model types

Architecture	Layers
CNN-1D	<ul style="list-style-type: none"> <li>- Conv1D(filters=x1, kernel=3, ReLU)</li> <li>- MaxPooling1D(pool=2)</li> <li>- Conv1D(filters=x2, kernel=3, ReLU)</li> <li>- MaxPooling1D(pool=2)</li> <li>- Flatten</li> <li>- Dense(units=x3, ReLU)</li> <li>- Dense(units=3, Softmax)</li> </ul>
LSTM	<ul style="list-style-type: none"> <li>- LSTM(x1 units, tanh, return_sequences=True)</li> <li>- LSTM(x2 units, tanh)</li> <li>- Dense(x3 units, ReLU)</li> <li>- Dense(3 units, Softmax)</li> </ul>
CNN-1D + LSTM	<ul style="list-style-type: none"> <li>- Conv1D(128, kernel=5, ReLU) + MaxPooling</li> <li>- Conv1D(64, kernel=5, ReLU) + MaxPooling</li> <li>- Dropout(0.3)</li> <li>- LSTM(128, return_sequences=True)</li> <li>- LSTM(64)</li> <li>- Dense(32, ReLU)</li> <li>- Dense(3, Softmax)</li> </ul>
CNN-1D + Bi-LSTM	<ul style="list-style-type: none"> <li>- Conv1D(32, kernel=3, ReLU, L2=1e-4), BatchNorm, MaxPool(2), Dropout(0.3)</li> <li>- Conv1D(64, kernel=3, ReLU, L2=1e-4), BatchNorm, MaxPool(2), Dropout(0.3)</li> <li>- Bi-LSTM(128, return_sequences=True, L2=1e-4), Dropout(0.3)</li> <li>- Bi-LSTM(64, L2=1e-4), Dropout(0.3)</li> <li>- Dense(output softmax) with L1=1e-5, L2=1e-4 regularization</li> </ul>

## 2.5. Experimental result

In this study, the dataset was partitioned into three subsets: 80% for training, 10% for validation, and 10% for testing. This stratified split not only ensures the model has ample data to learn underlying patterns but also provides a robust framework for evaluation. The validation set is used during training to monitor overfitting, tune hyperparameters, and guide early stopping, while the test set remains unseen until the very end to offer an unbiased measure of generalization performance. Adopting this split ratio aligns with standard machine learning practices, where an 80/10/10 partition is widely recommended to maintain representative class distributions and avoid biased estimates. Moreover, stratified sampling was applied to preserve the proportional representation of each emotion class across all subsets, which prevents class imbalance from skewing model performance evaluations.

## 2.6. Implementation result

The process of evaluating lecturer performance through SER starts with analyzing all audio data obtained from lecture recordings. The audio data undergoes consistent preprocessing to ensure accurate and reliable results. SER typically follows a structured pipeline. It begins by capturing and cleaning the raw audio signal, removing noise and dividing it into short, overlapping frames, often using pre-emphasis, endpoint detection, and framing techniques. Next, for each frame, acoustic features are extracted, which often include hand-crafted descriptors like MFCCs, pitch, energy, ZCR, spectral coefficients, or more advanced formant and wavelet features. In modern systems, these handcrafted features might be enhanced with deep representations (e.g., embeddings from wav2vec or HuBERT), sometimes using multi-stream fusion architectures to capture complementary information. The resulting features are then fed into classifiers, ranging from traditional models

like support vector machines (SVMs), random forests (RF), or Gaussian mixture model (GMMs), to advanced neural networks including CNN-RNN hybrids or transformer-based structures. Finally, frame-level emotion predictions are aggregated, either by counting votes or averaging probabilistic outputs, to yield a distribution over emotions, which is expressed as the percentage share of each emotion across the entire speech sample.

### 3. RESULTS AND DISCUSSION

According to several previous studies, CNN-1D has shown good performance. Therefore, this study experimented to observe the training performance using 5 classifier models: CNN-1D, LSTM, Bi-LSTM, CNN-1D with LSTM, and CNN-1D with Bi-LSTM. Parameters used in previous studies indicate that combinations of MFCC, spectrogram, and chroma features can effectively represent voice characteristics. Other studies also suggest that combining MFCC, energy, and ZCR can represent voice characteristics better. MFCC is represented by various numbers of coefficients as representatives of voice characteristics. Some of these coefficients include 12, 13, 25, and 40, resulting in 40 different scenarios for this observation as shown in Table 3. Based on the previously mentioned model configurations, the training process will stop if the model's performance on validation data does not improve for 5 consecutive epochs. This aims to prevent overfitting, which means the model is too detailed on training data and does not generalize well to new data. If the accuracy or loss values are not too far from the validation accuracy and loss values, the model performance is considered more ideal. Table 3 shows that the training accuracy values are almost identical, but there is a noticeable gap between accuracy and loss compared to validation results as shown in Figure 7.

Table 3. Summary of models with features, classifier, and performance metrics

ID	Features	Classifier	Loss	Acc_train	Loss_val	Acc_val
1	MFCC(12)+Energy+ZCR	CNN	0.183	0.934	0.452	0.828
2	MFCC(13)+Energy+ZCR	CNN	0.219	0.927	0.464	0.825
3	MFCC(25)+Energy+ZCR	CNN	0.063	1.000	0.598	0.868
4	MFCC(40)+Energy+ZCR	CNN	0.064	1.000	0.487	0.887
5	MFCC(12)+Mel+Chroma	CNN	0.178	0.969	0.615	0.848
6	MFCC(13)+Mel+Chroma	CNN	0.229	0.961	0.533	0.864
7	MFCC(25)+Mel+Chroma	CNN	0.070	1.000	0.410	0.904
8	MFCC(40)+Mel+Chroma	CNN	0.078	1.000	0.445	0.878
9	MFCC(12)+Energy+ZCR	LSTM	0.183	0.934	0.452	0.828
10	MFCC(13)+Energy+ZCR	LSTM	0.316	0.862	0.478	0.838
11	MFCC(25)+Energy+ZCR	LSTM	0.281	0.883	0.701	0.791
12	MFCC(40)+Energy+ZCR	LSTM	0.302	0.873	0.589	0.791
13	MFCC(12)+Mel+Chroma	LSTM	0.419	0.830	0.606	0.762
14	MFCC(13)+Mel+Chroma	LSTM	0.223	0.926	0.478	0.854
15	MFCC(25)+Mel+Chroma	LSTM	0.229	0.912	0.651	0.795
16	MFCC(40)+Mel+Chroma	LSTM	0.082	0.977	1.195	0.821
18	MFCC(13)+Energy+ZCR	Bi-LSTM	0.092	0.973	0.846	0.821
19	MFCC(25)+Energy+ZCR	Bi-LSTM	0.240	0.898	0.547	0.811
20	MFCC(40)+Energy+ZCR	Bi-LSTM	0.080	0.968	0.599	0.858
21	MFCC(12)+Mel+Chroma	Bi-LSTM	0.127	0.956	0.827	0.808
22	MFCC(13)+Mel+Chroma	Bi-LSTM	0.176	0.939	0.645	0.805
23	MFCC(25)+Mel+Chroma	Bi-LSTM	0.191	0.930	0.793	0.808
24	MFCC(40)+Mel+Chroma	Bi-LSTM	0.154	0.949	0.588	0.844
26	MFCC(13)+Energy+ZCR	CNN-LSTM	0.176	0.940	0.475	0.834
27	MFCC(25)+Energy+ZCR	CNN-LSTM	0.128	0.951	0.751	0.831
28	MFCC(40)+Energy+ZCR	CNN-LSTM	0.026	0.992	0.644	0.878
29	MFCC(12)+Mel+Chroma	CNN-LSTM	0.093	0.971	0.563	0.848
30	MFCC(13)+Mel+Chroma	CNN-LSTM	0.078	0.974	0.386	0.881
31	MFCC(25)+Mel+Chroma	CNN-LSTM	0.158	0.946	0.494	0.851
32	MFCC(40)+Mel+Chroma	CNN-LSTM	0.075	0.975	0.481	0.868
34	MFCC(13)+Energy+ZCR	CNN+Bi-LSTM	0.451	0.853	0.446	0.851
35	MFCC(25)+Energy+ZCR	CNN+Bi-LSTM	0.404	0.863	0.444	0.838
36	MFCC(40)+Energy+ZCR	CNN+Bi-LSTM	0.346	0.891	0.484	0.871
37	MFCC(12)+Mel+Chroma	CNN+Bi-LSTM	0.278	0.916	0.387	0.887
38	MFCC(13)+Mel+Chroma	CNN+Bi-LSTM	0.483	0.831	0.433	0.854
39	MFCC(25)+Mel+Chroma	CNN+Bi-LSTM	0.494	0.825	0.470	0.854
40	MFCC(40)+Mel+Chroma	CNN+Bi-LSTM	0.299	0.915	0.434	0.874

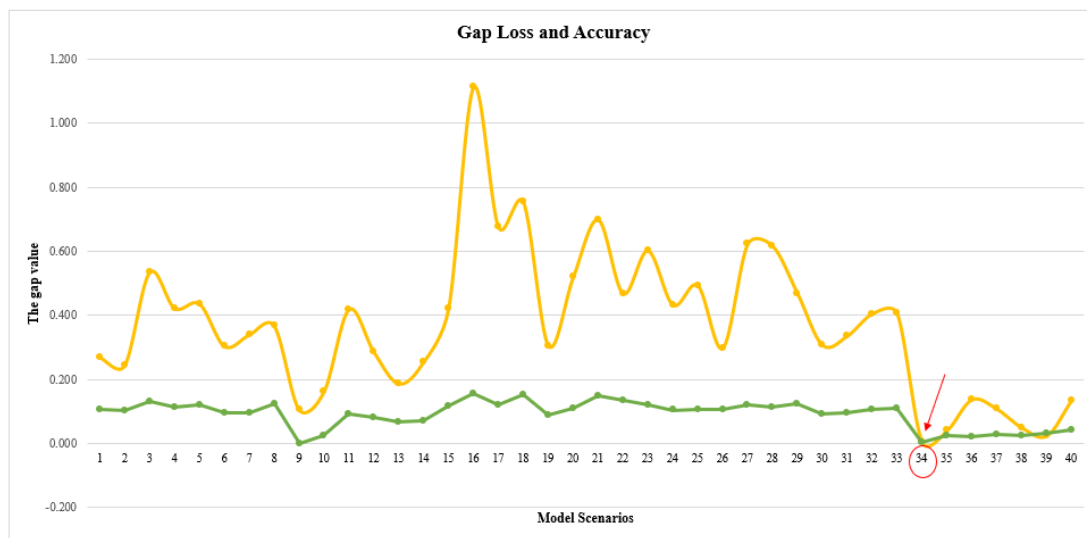


Figure 7. Observations of training results across all scenarios

The use of the CNN-1D with Bi-LSTM model (model 34) performed better with MFCC 13 coefficients, energy, and ZCR parameters. This combination effectively captured the temporal and spatial features of the voice data, resulting in better generalization on validation data. The optimal performance of this model reflects not only the strong learning capability of the important features but also the stability in maintaining alignment between training and validation data, indicating that the configuration and preprocessing of the data used were appropriate.

In the implementation phase, Figure 8 shows emotion recognition by model 34, revealing that the lecturer stayed mostly neutral with only brief moments of negative emotion (e.g., anger or sadness). This realtime emotional trend offers a clear, objective view of in-class affect, enabling educators and administrators to quickly spot and respond to stress or disengagement. Such transparency enhances feedback-driven teaching improvements without relying solely on student opinions, aligning with successful classroom emotion analytics approaches.

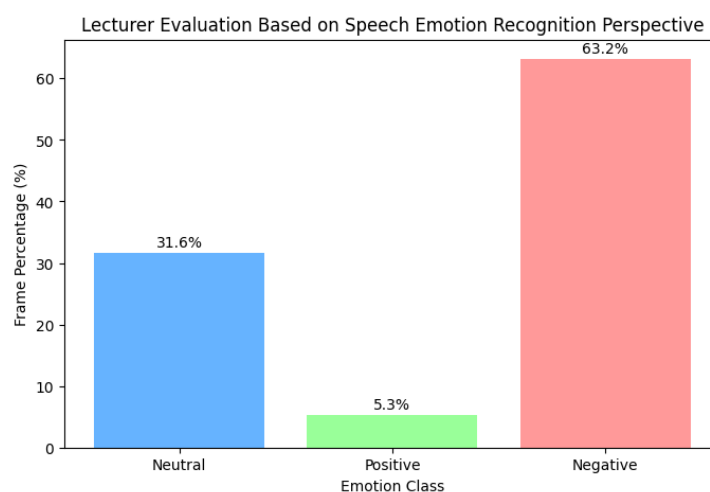


Figure 8. The example of emotional tone distribution of the lecturer evaluation by the student

Recent educational technology studies support this use of emotion analytics in the classroom. For instance, automated auditory emotion recognition systems have been employed to detect negative “moments” based on speech emotion patterns in teaching videos, revealing correlations between audio cues and classroom

climate. Similarly, real-time emotion monitoring in lecture settings has proven effective for assessing teaching quality and adjusting pedagogical approaches on the fly. These precedents highlight the value of visually driven SER outputs, like the one shown in Figure 8, for enabling timely, objective insights into in-class emotional dynamics.

4. CONCLUSION

Our extensive experiments reveal that the hybrid CNN-1D + Bi-LSTM architecture (model 34) stands out among the 40 evaluated configurations. By integrating spatial convolutional filters with bidirectional temporal modeling powered by a feature combination of 13 coefficient MFCCs, energy, and ZCR. The model effectively captures both the textures in the frequency domain and the temporal dynamics essential for emotion recognition. This synergy allowed Model 34 to significantly outperform the standalone CNN-1D, LSTM, and Bi-LSTM models in our tests. The robust generalization observed in Model 34 evidenced by strong validation performance and low signs of overfitting confirms the soundness of our feature selection and model design. By emphasizing both spectral and temporal patterns, the model avoids overemphasis on narrow signal characteristics while preserving essential emotional cues. This mirrors best practices in the field: hybrid CNN-Bi-LSTM networks are now clearly recognized as superior for processing rich time-series audio features. In summary, our research demonstrates that using a hybrid architecture like CNN-1D + Bi-LSTM, combined with the foundational acoustic characteristics (MFCC, energy and ZCR), offers a durable, high-performing framework for SER. This approach not only matches, but in several cases, surpasses current state-of-the-art methods. Future work may explore the integration of attention mechanisms, additional modalities such as chroma features or spectrogram derivatives, or data enhancement strategies to further enhance accuracy and adaptability.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to the Research and Community Service Agency, Telkom University, for their invaluable support and contributions throughout this research. Their assistance has been instrumental in the completion of this study, and we are deeply appreciative of their commitment to fostering research and development.

FUNDING INFORMATION

This article is an additional output of the internal research grant funded by LPPM Telkom University under Grant ID IT Tel6667-b/LPPM-000/Ka. LPPM/III/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Yesy Diah Rosita	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓
Wahyu Andi Saputra						✓	✓	✓	✓		✓		✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal Analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project Administration

Fu : Funding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

## INFORMED CONSENT

The authors declare that they have no conflicts of interest regarding the publication of this article.

## DATA AVAILABILITY

The data supporting this study consist of audio recordings of university lecturers during their teaching sessions. These recordings were collected with informed consent from all participants and in compliance with institutional ethical regulations. Due to privacy considerations involving human subjects, the data are not publicly accessible. However, they are available from the corresponding author upon reasonable request.




## REFERENCES

- [1] I. Noben, J. F. Deinum, and W. H. A. Hofman, "Quality of teaching in higher education: reviewing teaching behaviour through classroom observations," *International Journal for Academic Development*, vol. 27, no. 1, pp. 31–44, Jan. 2022, doi: 10.1080/1360144X.2020.1830776.
- [2] Y. D. Rosita, Z. Salsabila and A. R. P. Pamungkas, "Lecturer evaluation from the perspective of speech emotion recognition with deep learning," *2025 International Conference on Data Science and Its Applications (ICoDSA)*, Jakarta, Indonesia, 2025, pp. 565–571, doi: 10.1109/ICoDSA67155.2025.11157430.
- [3] Z. Kexin and L. Yunxiang, "Speech emotion recognition based on transfer emotion-discriminative features subspace learning," *IEEE Access*, vol. 11, pp. 56336–56343, 2023, doi: 10.1109/ACCESS.2023.3282982.
- [4] B. Paul, S. Bera, T. Dey, and S. Phadikar, "Machine learning approach of speech emotions recognition using feature fusion technique," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 8663–8688, Jan. 2024, doi: 10.1007/s11042-023-16036-y.
- [5] E. A. Alkhamali, A. Allinjawi, and R. B. Ashari, "Combining transformer, convolutional neural network, and long short-term memory architectures: a novel ensemble learning technique that leverages multi-acoustic features for speech emotion recognition in distance education classrooms," *Applied Sciences*, vol. 14, no. 2, 2024, doi: 10.3390/app14125050.
- [6] B. Bakariya, A. Singh, H. Singh, P. Raju, R. Rajpoot, and K. K. Mohbey, "Facial emotion recognition and music recommendation system using CNN-based deep learning techniques," *Evolving Systems*, vol. 15, no. 2, pp. 641–658, Apr. 2024, doi: 10.1007/s12530-023-09506-z.
- [7] T. N. M. Aris and C. Ningning, "Integration of CNN and LSTM networks for behavior feature recognition: an analysis," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 14, no. 5, pp. 1793–1799, Oct. 2024, doi: 10.18517/ijaseit.14.5.10116.
- [8] N. Barsainyan and D. K. Singh, "Optimized cross-corpus speech emotion recognition framework based on normalized 1D convolutional neural network with data augmentation and feature selection," *International Journal of Speech Technology*, vol. 26, no. 4, pp. 947–961, Dec. 2023, doi: 10.1007/s10772-023-10063-8.
- [9] N. Niyozmatova, K. Jalelov, B. Samijonov, and M. Madrahimova, "Eliminating noise from a speech signal based on a pair of filters," *International Journal of Science and Research Archive*, vol. 13, no. 2, pp. 401–410, Nov. 2024, doi: 10.30574/ijrsra.2024.13.2.2058.
- [10] R. J.-Moreno and R. A. Castillo, "Deep learning speech recognition for residential assistant robot," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 2, pp. 585–592, Jun. 2023, doi: 10.11591/ijai.v12.i2.pp585-592.
- [11] N. Choudhury and U. Sharma, "Enhanced emotion recognition from spoken assamese dialect: a machine learning approach with language-independent features," *Traitement du Signal*, vol. 40, no. 5, pp. 2147–2160, Oct. 2023, doi: 10.18280/ts.400532.
- [12] E. Mancini, A. Galassi, F. Ruggeri, and P. Torroni, "Disruptive situation detection on public transport through speech emotion recognition," *Intelligent Systems with Applications*, vol. 21, Mar. 2024, doi: 10.1016/j.iswa.2023.200305.
- [13] S. Murugaiyan and S. R. Uyyala, "Aspect-based sentiment analysis of customer speech data using deep convolutional neural network and BiLSTM," *Cognitive Computation*, vol. 15, no. 3, pp. 914–931, May 2023, doi: 10.1007/s12559-023-10127-6.
- [14] A. Bastanfard and A. Abbasian, "Speech emotion recognition in Persian based on stacked autoencoder by comparing local and global features," *Multimedia Tools and Applications*, vol. 82, no. 23, pp. 36413–36430, Sep. 2023, doi: 10.1007/s11042-023-15132-3.
- [15] K. Chauhan, K. K. Sharma, and T. Varma, "A method for simplifying the spoken emotion recognition system using a shallow neural network and temporal feature stacking & pooling (TFSP)," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11265–11283, Mar. 2023, doi: 10.1007/s11042-022-13463-1.
- [16] K. Mao, Y. Wang, L. Ren, J. Zhang, J. Qiu, and G. Dai, "Multi-branch feature learning based speech emotion recognition using SCAR-NET," *Connection Science*, vol. 35, no. 1, Dec. 2023, doi: 10.1080/09540091.2023.2189217.
- [17] P. Tiwari, H. Rathod, S. Thakkar, and A. D. Darji, "Multimodal emotion recognition using SDA-LDA algorithm in video clips," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 6585–6602, Jun. 2023, doi: 10.1007/s12652-021-03529-7.
- [18] S. K. Panda, A. K. Jena, M. R. Panda, and S. Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42763–42781, 2023, doi: 10.1007/s11042-023-15275-3.
- [19] A. A. Alnuaim *et al.*, "Human-computer interaction with detection of speaker emotions using convolution neural networks," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/7463091.
- [20] B. NaserSharif, M. Ebrahimpour, and N. Naderi, "Multi-layer maximum mean discrepancy in auto-encoders for cross-corpus speech emotion recognition," *The Journal of Supercomputing*, vol. 79, no. 12, pp. 13031–13049, 2023, doi: 10.1007/s11227-023-05161-y.
- [21] A. Marik, S. Chattopadhyay, and P. K. Singh, "A hybrid deep feature selection framework for emotion recognition from human speeches," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11461–11487, Mar. 2023, doi: 10.1007/s11042-022-14052-y.
- [22] I. G. B. A. P. Paramitha, H. B. Kusnawan, and M. Ernawati, "Performance comparison of deep learning algorithm for speech emotion recognition," *Journal of Computer Science and Informatics Engineering (J-Cosine)*, vol. 6, no. 2, pp. 99–106, Dec. 2022, doi: 10.29303/jcosine.v6i2.443.




- [23] G. Assunção, P. Menezes, and F. Perdigão, "Speaker awareness for speech emotion recognition," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 4, pp. 15–22, Apr. 2020, doi: 10.3991/ijoe.v16i04.11870.
- [24] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, Jul. 2021, doi: 10.1016/j.ins.2021.02.016.
- [25] K. Kaur and P. Singh, "Impact of feature extraction and feature selection algorithms on punjabi speech emotion recognition using convolutional neural network," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 5, pp. 1–23, Sep. 2022, doi: 10.1145/3511888.
- [26] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Communication*, vol. 120, pp. 11–19, Jun. 2020, doi: 10.1016/j.specom.2020.03.005.
- [27] S. Jothamani and K. Premalatha, "MFF-SAUG: multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons & Fractals*, vol. 162, Sep. 2022, doi: 10.1016/j.chaos.2022.112512.

## BIOGRAPHIES OF AUTHORS



**Yesy Diah Rosita**    received a Master of Computer from Institute of Science and Technology Surabaya, Indonesia. Her thesis about classification of infant's cry sounds using an artificial neural network was published on IEEE, 2016. Now, she is a lecturer in the School of Computing at Telkom University, Indonesia. She is interested in speech recognition, deep learning, and decision support system. She is also a contributor to Mathworks, which is a company that specializes in mathematical computing software. She can be contacted at email: yesydr@telkomuniversity.ac.id.



**Wahyu Andi Saputra**    received a Master of Engineering from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2018. His thesis is about identifying malaria using artificial neural networks and was published in IEEE. He is currently a lecturer at the School of Computing at Telkom University, Indonesia. He has also worked on several projects related to computer vision and software engineering. He can be contacted at email: andiwahyu@telkomuniversity.ac.id.