

# Speech emotion recognition with light gradient boosting decision trees machine

Kah Liang Ong<sup>1</sup>, Chin Poo Lee<sup>1</sup>, Heng Siong Lim<sup>2</sup>, Kian Ming Lim<sup>1</sup>

<sup>1</sup>Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia

<sup>2</sup>Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia

---

## Article Info

### Article history:

Received Jun 21, 2022

Revised Oct 2, 2022

Accepted Oct 13, 2022

### Keywords:

Light gradient boosting machine

Machine learning

Speech

Speech emotion

Speech emotion recognition

---

## ABSTRACT

Speech emotion recognition aims to identify the emotion expressed in the speech by analyzing the audio signals. In this work, data augmentation is first performed on the audio samples to increase the number of samples for better model learning. The audio samples are comprehensively encoded as the frequency and temporal domain features. In the classification, a light gradient boosting machine is leveraged. The hyperparameter tuning of the light gradient boosting machine is performed to determine the optimal hyperparameter settings. As the speech emotion recognition datasets are imbalanced, the class weights are regulated to be inversely proportional to the sample distribution where minority classes are assigned higher class weights. The experimental results demonstrate that the proposed method outshines the state-of-the-art methods with 84.91% accuracy on the Berlin database of emotional speech (emo-DB) dataset, 67.72% on the Ryerson audio-visual database of emotional speech and song (RAVDESS) dataset, and 62.94% on the interactive emotional dyadic motion capture (IEMO-CAP) dataset.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



---

## Corresponding Author:

Chin Poo Lee

Faculty of Information Science and Technology, Multimedia University

Melaka, Malaysia

Email: cplee@mmu.edu.my

---

## 1. INTRODUCTION

Speech emotion recognition aims to identify the affective aspects of the audio waveforms regardless of the content of the utterances. Speech emotion recognition requires knowledge of signal processing and machine learning. From the perspective of signal processing, the audio waveforms are represented as features in the temporal and frequency domain. Thereafter, the features are classified into their respective class by machine learning methods. The performance of the machine learning methods highly relies on the quality of the features. Apart from that, speech emotion recognition also faces data scarcity problems where there are limited data samples.

In this work, the audio waveforms are represented by seven discriminative temporal and frequency features. Besides that, data augmentation techniques are applied to solve data scarcity problems. After the preprocessing and feature extraction, the features are thereafter classified into the emotion class by the light gradient boosting machine. As the sample distributions of the emotion classes are skewed, the class weights are adjusted according to the class distributions to alleviate imbalance dataset problems. In addition, to determine the optimal settings for the light gradient boosting machine, hyperparameter tuning with a comprehensive grid search is conducted on all datasets used. The contributions of this paper are as follows. i) Machine learning

methods generally require large amount of data for learning, therefore, data augmentation techniques, including time stretching and pitch shifting, are proposed to synthesize more audio samples. With data augmentation, the machine learning method learns on more data samples, thus reducing the overfitting and improving the generalization capability of the machine learning method. ii) The raw audio signals are hardly feasible for emotion recognition. Thus, several discriminative features in the temporal and frequency domains are used to represent the audio signals. The features include mel-frequency cepstral coefficients, mel spectrogram, root mean square, chroma features, zero-crossing rate, wavelet transform, and kurtosis. And iii) While deep learning methods demonstrate high accuracy in many applications, they usually impose enormous computational costs. In view of this, the light gradient boosting machine is proposed for the classification of speech emotions. The empirical results show that the light gradient boosting machine can deliver comparable performance at a faster training speed and lower computational costs compared to deep learning methods.

The existing works in speech emotion recognition [1]–[3] could be categorized into three types: machine learning, convolutional neural networks (CNNs), and recurrent neural networks (RNNs). In the machine learning methods, the features are handcrafted to represent the audio signals. Shegokar and Sircar [4] fed the continuous wavelet transform features into the support vector machines (SVM) classifier. In 2018, Guan *et al.* [5] proposed speech emotion recognition by feeding the local dynamic features for the model training. Jin *et al.* [6] extracted acoustic and lexical features as the input of the SVM classifier for speech emotion recognition. Farooq *et al.* [7] implemented the deep CNN correlation-based feature selection to select the features for the SVM classifier. Liu *et al.* [8] proposed a phoneme clustering classification by recognizing the unlabeled phonemes' formant characteristics of speech signals. In 2020, Koduru *et al.* [9] proposed an enhanced feature extraction algorithm by applying the feature selection for the features and then fed them into the decision tree classifier. Furthermore, Slimi *et al.* [10] proposed a one-hidden-layer neural network classification by extracting the log mel-spectrogram features and feeding them into the classifier. Most researchers about the machine learning methods only considered one dataset except Farooq *et al.* [7] and Liu *et al.* [8] who both conducted experiments on three speech emotion datasets. hgjg

Other than the conventional machine learning methods, CNNs are also widely used for speech emotion recognition [11]–[13]. Anvarjon *et al.* [14] proposed a lightweight CNN to overcome the complexity problems by using only a few parameters. The proposed architecture used rectangular kernels and a modified pooling strategy to extract the deep frequency features. The proposed model contains eight convolutional layers with a rectified linear unit (ReLU), three max-pooling layers, batch normalization, and two fully connected layers with a SoftMax classifier.

Issa *et al.* [15] proposed CNN classification by using five features. The proposed model contains one-dimensional convolutional layers with dropout, batch normalization, and activation layers. Badshah *et al.* [16] applied the fast fourier transform (FFT) to convert the speech signal into a spectrogram image. The spectrogram image is then passed to a CNN model for emotion classification. The CNN model consists of convolutional layers, fully connected layers, and a SoftMax layer. Tripathi *et al.* [17] extracted the features by using the center loss and reconstruction as regularization. They used convolution layers with max pooling layers to extract the features from each parallel convolution path. The extracted features were then fed into fully connected layers with batch normalization.

Tripathi *et al.* [18] used speech features and transcriptions to improve the recognition rates. The speech was represented as spectrograms and mel-frequency cepstral coefficients (MFCC) to retain emotion-related characteristics. The transcriptions captured the semantic meaning. A multi-channel CNN model was deployed for emotion recognition. The speech channel alternated between spectrogram and MFCC. The speech channel consists of four parallel 2D-CNN layers with kernels of different sizes. The transcription channel takes word embeddings as the input. Outputs from the channels are passed to the fully connected layers and classification layer. Yenigalla *et al.* [19] used the spectrogram and phoneme embedding as the input to the multi-channel CNN model for emotion classification. The phoneme channel received phoneme embedding followed by four parallel convolution layers of different kernels. The spectrogram channel consists of four parallel 2D convolution layers with different filter sizes. Both outputs from the phoneme and spectrogram channel were passed to two fully connected layers and a classification layer.

RNNs are suitable for the classification, processing, and forecasting that involves sequential data [20], [21]. Unlike feedforward neural networks, RNNs contain feedback connections that allow the bi-directional passage of information. Chernykh and Prikhodko [22] proposed a connectionist temporal classification method based on RNNs. The networks classified the sequence of acoustic features into

speech emotion classes. Lee and Tashev [23] adopted RNNs where the model extracted the high-level representation features from the temporal dynamics for speech emotion recognition. Moreover, Latif *et al.* [24] proposed transfer learning-based deep belief networks for speech emotion recognition. They conducted cross-language and cross-corpora experiments to investigate the performance of deep belief networks. Mustaqeem *et al.* [25] implemented key sequence segment selections based on radial basis function networks to extract the specific features from the spectrogram. The features were then classified using a bi-directional long short-term memory model.

## 2. METHOD

In this section, the proposed speech emotion recognition with light gradient boosting machine (LightGBM) is referred to as the emo-LGBM method. The proposed emo-LGBM mainly consists of three steps: data augmentation, feature extraction, and classification. Firstly, the data samples are resampled at the sampling frequency of 44.1 kHz. Subsequently, two data augmentation techniques, namely time stretching, and pitch shifting are performed on the data samples to increase the number of samples for the model learning. After that, seven frequency domain and time domain features are extracted from the augmented audio samples. Eventually, the extracted features are fed into the LightGBM method to recognize the speech's emotional state. Figure 1 depicts the system flow of the proposed emo-LGBM method.

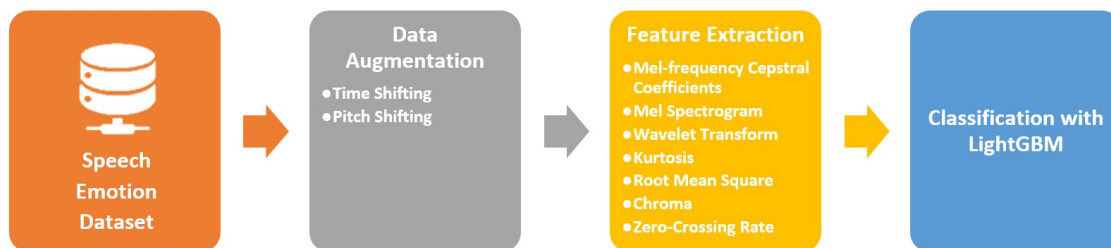


Figure 1. Process flow of speech emotion recognition with LightGBM

### 2.1. Data augmentation

Data augmentation is usually utilized to synthesize more training samples based on the original samples. In this work, two data augmentation techniques, namely time stretching and pitch shifting are leveraged. Both time stretching and pitch shifting are applied at four different factor values: 0.8, 0.9, 1.1, and 1.2, for synthesizing eight times more training samples than the original number of samples.

#### 2.1.1. Time stretching

Time stretching aims to change the speed of an audio signal to speed up or slow down the audio signal. In the time stretching process, the duration of the audio signal will be scaled and modified. Increasing the speed of the audio signal diminishes the output of the audio signal where some audio segments are deleted and respliced to shorten the duration of the audio signal. Contrarily, decreasing the speed of the audio signal extends the output of the audio signal. In order to lengthen the duration of the audio signal, some similar audio segments are generated and rebuilt together with the existing audio segments.

During the time stretching process, a short-time fourier transform (STFT) is first applied to convert the audio signal from the time domain into the frequency domain signal. Then, the frequency domain signal is divided and windowed by discrete fourier transforms (DFT) over the Hanning windows. To calculate the stretched STFT matrix, the Hanning windows are used to determine the reflection padding from the frame edges to simplify the time grid of the sample index and frame index. For the stretched output, an inverse STFT is used to reconstruct the frequency domain signal back to the time domain signal. Figure 2 depicts the sample waveforms of time stretching with different factor values.

#### 2.1.2. Pitch shifting

Pitch shifting aims to increase or decrease the pitch of the audio signal without changing the total duration of the signal. In the pitch shifting process, only the pitch of the audio signal is modulated while the

speed of the signal remains unchanged. The pitch shifting modifies the linear-frequency spectrogram vertically to reflect the pitch of the audio signal. Pitch shifting involves the phase vocoder shifting the pitch of the audio signal by taking the sample rate of the audio signal and the given fractional factor value. The fractional factor value of  $m$  determines whether the shifted pitch will become sharper or duller. Given the  $m$  value larger than 1, the pitch of the audio signal will be shifted sharper. Inversely, given the  $m$  value smaller than 1, the pitch of the audio signal will be shifted duller. Figure 3 depicts the sample waveforms of pitch shifting with different factor values. Given the original audio signal  $l_t$ , the pitch shifting output  $l'_t$  is defined by (1).

$$l'_t = \frac{l_t}{m} \quad (1)$$

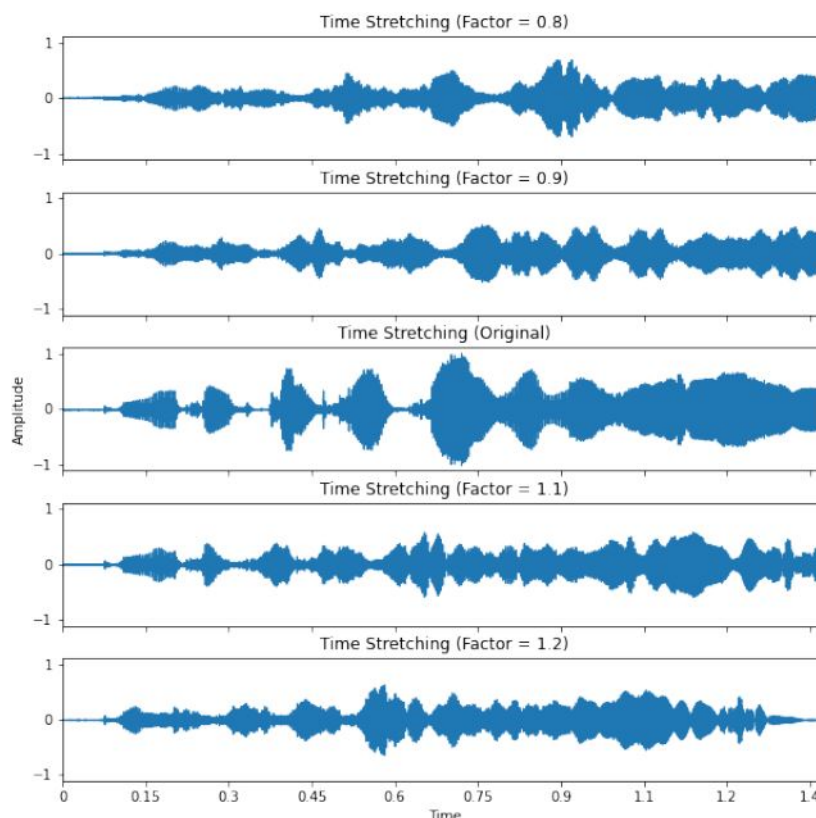


Figure 2. Sample waveforms with different time stretching factors

## 2.2. Feature extraction

Feature extraction plays an important role in finding the representations that best describe the unique characteristics of the class. Feature extraction encodes the audio signal into an understandable format for model learning. There are seven features of the audio signals that are extracted as the input for model learning, namely mel-frequency cepstral coefficients, mel-spectrogram, wavelet transform, kurtosis, root mean square, chroma, and zero-crossing rate.

In this work, four frequency domain features are leveraged, namely mel-frequency cepstral coefficients, mel spectrogram, wavelet transform, and kurtosis. The mel-frequency cepstrum captures the short-time power spectrum of the sound waves. The mel-frequency scale mimics how the human auditory responds to the frequencies. A mel spectrogram depicts the frequency spectrum of the sound wave in the mel frequency scale. Wavelet transform is used to divide and convert the continuous time signal into components of different scales, thus capturing different details of the audio signals. Wavelet transform also provides simultaneous time and frequency domain localization for speech emotion analysis. Kurtosis is a statistical domain feature used to calculate the intensity distribution of the audio signal. From the peak and flat distributions, kurtosis helps in characterizing speech emotions.

Apart from frequency domain features, time domain features are also adopted. The time domain features include root mean square, chroma features, and zero-crossing rate. The root mean square calculates the square root of the arithmetic mean of the squared amplitude values of the audio signal. The root mean square amplitude values are useful in evaluating the reflectivity in the zone of interest for speech emotion recognition. The chroma features are compact descriptors that represent the tonal information of an audio signal. The chroma features encode the profile of the pitch classes into the audio waveform, which is effective in differentiating the intensity between the audio signals. The zero-crossing rate counts the number of zero crossings (from positive to negative, and vice versa) of the audio signal within a duration. The zero-crossing rate is a good basic property of an audio signal to compare the sign of each pair of consecutive samples. Representing the audio signals in both the frequency domain and temporal domain, it provides a detailed description of the audio signal, thus improving the performance of speech emotion recognition.

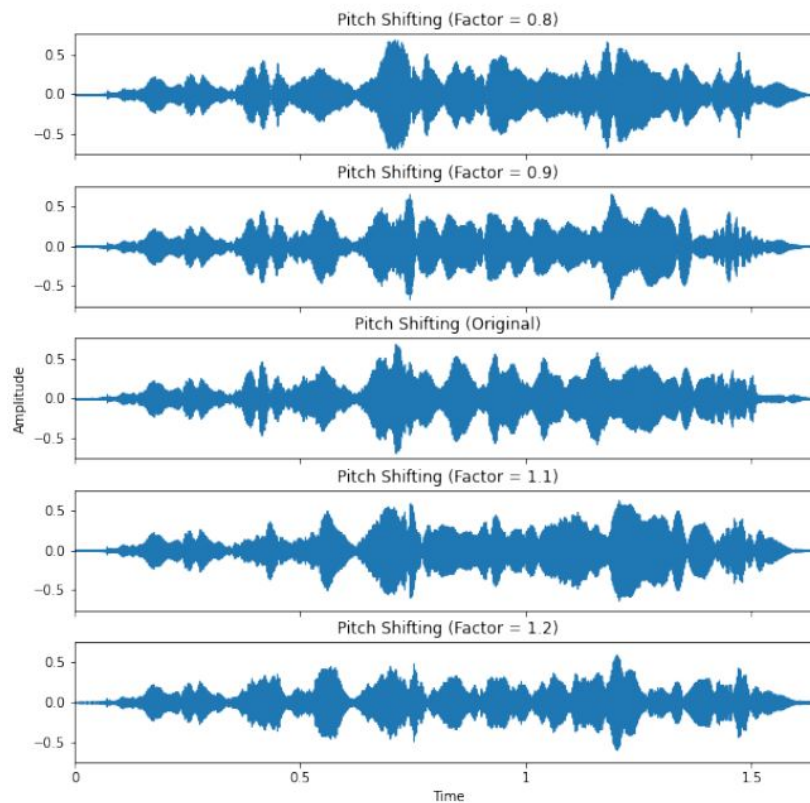


Figure 3. Sample waveforms with different pitch shifting factors

### 2.3. Classification with light gradient boosting machine

LightGBM [26] is an efficient gradient-boosting method that uses tree-based learning. In the gradient boosting framework, the trees are built one after another, unlike the random forest where the tree is created for each sample. One of the advantages of using LightGBM is the faster training speed and higher efficiency to handle the missing values between the nodes. Moreover, LightGBM uses a leaf-wise tree growth algorithm. The leaf-wise tree growth algorithm splits the nodes based on the contribution to the global loss. Therefore, LightGBM can avoid growing into a very deep tree in the architecture. Figure 4 illustrates the leaf-wise tree growth algorithm.

As the speech emotion datasets are normally imbalanced where some emotion classes may have a much larger sample size than others, the class weights of the LightGBM classifier are adjusted accordingly. To reduce the performance degradation caused by the skewed sample distribution, the minority classes are assigned higher class weights while the class weights of the majority classes are reduced. In doing so, the LightGBM classifier imposes a higher penalty for the misclassification made by the minority class. By doing so, the LightGBM classifier focuses more on reducing the errors of the minority class.

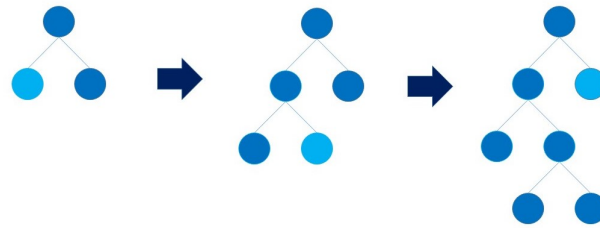


Figure 4. Leaf-wise tree growth algorithm of LightGBM

## 2.4. Datasets

This section describes the speech emotion datasets that are used to evaluate the performance of the proposed emo-LGBM method. Specifically, the Berlin database of emotional speech (emo-DB), Ryerson audio-visual database of emotional speech and song (RAVDESS), and interactive emotional dyadic motion capture (IEMOCAP). To have a fair comparison with the existing works, each speech emotion dataset is divided into 80% training set and 20% testing set.

The emo-DB [27] is a dataset for speech emotion recognition in the German language. Five male and five female professional speakers took part in the data recording. The dataset consists of 535 utterances with 7 emotions: anger, boredom, anxiety, happiness, sadness, disgust, and neutral.

The Ryerson audio-visual database of emotional speech and song (RAVDESS) [28] comprises 1,440 utterances. There are 8 emotions in the dataset: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. The sentences and songs were recorded by twelve male and twelve female professional actors in English.

The interactive emotional dyadic motion capture (IEMOCAP) [29] dataset consists of 5,507 utterances recorded by 5 male and 5 female actors. There are mainly 4 emotions that are used in the speech emotion recognition field, namely neutral, happiness, sadness, and anger. In this view, only four emotions were considered in this work. Table 1 summarizes the datasets with their emotion classes and the total number of samples.

Table 1. Summary of datasets

Datasets	Emotions	Number of Samples
Emo-DB	Anger, Boredom, Anxiety, Happiness, Sadness, Disgust, Neutral	535
RAVDESS	Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised	1440
IEMOCAP	Neutral, Happiness, Sadness, Anger	5507

## 2.5. Hyperparameter tuning

Hyperparameter tuning is essential to optimize the performance of the classifier. In this work, the boosting type of the LightGBM model is set to gradient boosting decision trees (GBDT). GBDT is a traditional method to improve the learning process that minimizes the residuals from previous predictions and a small learning rate is always preferred to achieve the optimal solution. The tuning of other hyperparameters is performed by grid search using the MLJAR library integrated with Optuna. The MLJAR library evaluates a predefined set of values for different hyperparameters, including the learning rate of the model, the maximum tree leaves for the base learning, the regularization weight value of leaves, the data and frequency of subset features in each iteration, the minimal number of data in one leaf, and the extra trees evaluating node. The hyperparameter tuning is conducted on each speech emotion dataset and the hyperparameter values that yield the highest test accuracy are selected as the optimal settings. Table 2 presents the summary of the hyperparameter tuning.

The experimental results of the datasets with and without data augmentation are presented in Table 3. It is observed that data augmentation has improved the performance of all three datasets. The data augmentation promotes a huge leap in accuracy, specifically 15.10% on the emo-DB dataset. On the RAVDESS dataset, incorporating data augmentation increases the recognition rate of the proposed emo-LGBM method from 59.65% to 67.72%. On the IEMOCAP dataset, the data augmentation has improved the accuracy from 61.22% to 62.94%. The enhancements in the performance demonstrate that time stretching and pitch-shifting techniques are effective in synthesizing more samples for model learning. The improvements also affirm the significance of data augmentation in boosting the generalization capability of the machine learning method.



Table 2. Summary of hyperparameter tuning

Hyperparameter	Optimal Values		
	Emo-DB	RAVDESS	IEMOCAP
learning_rate	0.1	0.1	0.05
num_leaves	721	1709	538
lambda_l1	0.0294555967	0.0000000269	0.0000339968
lambda_l2	0.0039221660	0.0000146529	0.0000647275
feature_fraction	0.8906030742	0.9178903409	0.6992103133
bagging_freq	0.8872461301	0.9569325817	0.5457987388
bagging_freq	3	3	6
min_data_in_leaf	31	30	57
extra_trees	True	True	False

Table 3. Experimental results with and without data augmentation

Data Augmentation	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
Without Data Augmentation	69.81	59.65	61.22
With Data Augmentation	84.91	67.72	62.94

### 3. EXPERIMENTS AND ANALYSIS

This section presents the comparison results of the proposed emo-LGBM method with the state-of-the-art methods. As observed in Table 4, the proposed emo-LGBM method outperforms the existing works on the Emo-DB, RAVDESS, and IEMOCAP datasets. On the emo-DB dataset, the existing methods recorded accuracy in the range of 58.86% - 82.73%. In comparison, the proposed emo-LGBM method records a higher accuracy of 84.91%. Likewise, on the RAVDESS dataset, the best existing method, i.e., CNNs [30] yielded an accuracy of 65.67% which is 2.05% lower than the proposed emo-LGBM method. The (-) in the results are due to the RAVDESS dataset not being used in the existing works. The performance of all methods is inferior on the IEMOCAP due to the relatively large sample size and multi-speaker speech. The existing works achieved an accuracy of 50.17% - 62.74%. Nevertheless, the proposed emo-LGBM method yields a higher accuracy of 62.94% despite the large dataset size and multi-speaker challenges. The experimental results corroborate the performance of the proposed emo-LGBM method. The data augmentation with pitch shifting and time stretching synthesizes more training samples for better LightGBM model learning, hence improving the generalization capability of the model.

Table 4. Comparative results on emo-DB, RAVDESS, IEMOCAP dataset

Method	Accuracy (%)		
	Emo-DB	RAVDESS	IEMOCAP
Phoneme Clustering with RF [8]	71.05	49.20	62.01
Phoneme Clustering with KNN [8]	60.05	43.24	59.28
Phoneme Clustering with MLP [8]	72.91	61.02	61.91
Deep Neural Network [31]	82.73	-	62.74
CNN [30]	58.86	65.67	55.24
LSTM [30]	59.67	53.97	56.99
CNN with LSTM [30]	69.72	53.08	50.17
Emo-LGBM (Proposed)	84.91	67.72	62.94

### 4. CONCLUSION

In this paper, enhanced speech emotion recognition with LightGBM is presented. The method first performs data augmentation by time stretching and pitch shifting to increase the sample size for better model learning. The augmented audio samples are thereafter encoded as the frequency and temporal domain features, including Mel-frequency Cepstral Coefficients, Mel spectrogram, Wavelet transform, Kurtosis, Chroma, Root Mean Square, and Zero-Crossing rate. The extracted features are finally passed to the LightGBM method for speech emotion recognition. As the datasets are imbalanced, the class weights are adjusted to assign more weights to the minority class. Not only that, hyperparameter tuning is also performed on the LightGBM method to determine the optimal hyperparameter settings. The empirical results demonstrate that the proposed emo-

LGBM method outshines the existing methods with the highest accuracy of 84.91%, 67.72%, and 62.94% on the emo-DB, RAVDESS, and IEMOCAP datasets, respectively.

## ACKNOWLEDGEMENT

The research in this work was supported by the Telekom Malaysia Research and Development Grant RDTC/221045, Fundamental Research Grant Scheme of the Ministry of Higher Education FRGS/1/2021/ICT02/MMU/02/4 and Multimedia University Internal Research Grant MMUI/220021.

## REFERENCES





- [1] M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, "Speech emotion recognition based on SVM and KNN classifications fusion," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 2, pp. 1259–1264, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1259-1264.
- [2] H. K. Palo and M. N. Mohanty, "Classification of emotional speech of children using probabilistic neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 2, pp. 311–317, Apr. 2015, doi: 10.11591/ijece.v5i2.pp311-317.
- [3] A. Agrima, I. Mounir, A. Farchi, L. Elmaazouzi, and B. Mounir, "Emotion recognition from syllabic units using k-nearest-neighbor classification and energy distribution," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 6, pp. 5438–5449, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5438-5449.
- [4] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *2016 10<sup>th</sup> International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec. 2016, pp. 1–8, doi: 10.1109/ICSPCS.2016.7843306.
- [5] H. Guan, Z. Liu, L. Wang, J. Dang, and R. Yu, "Speech emotion recognition considering local dynamic features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10733, 2018, pp. 14–23.
- [6] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 4749–4753, doi: 10.1109/ICASSP.2015.7178872.
- [7] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors*, vol. 20, no. 21, Oct. 2020, doi: 10.3390/s20216008.
- [8] 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.
- [9] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 45–55, Mar. 2020, doi: 10.1007/s10772-020-09672-4.
- [10] A. Slimi, M. Hamroun, M. Zrigui, and H. Nicolas, "Emotion recognition from speech using spectrograms and shallow neural networks," in *Proceedings of the 18<sup>th</sup> International Conference on Advances in Mobile Computing and Multimedia*, Nov. 2020, pp. 35–39, doi: 10.1145/3428690.3429153.
- [11] K. Zheng, Z. Xia, Y. Zhang, X. Xu, and Y. Fu, "Speech emotion recognition based on multi-level residual convolutional neural networks," *Engineering Letters*, vol. 28, no. 2, pp. 559–565, 2020.
- [12] P. Jiang, H. Fu, and H. Tao, "Speech emotion recognition using deep convolutional neural network and simple recurrent unit," *Engineering Letters*, vol. 27, no. 4, pp. 901–906, 2019.
- [13] P. S. Tan, K. M. Lim, C. P. Lee, and C. H. Tan, "Acoustic event detection with MobileNet and 1D-convolutional neural network," in *2020 IEEE 2<sup>nd</sup> International Conference on Artificial Intelligence in Engineering and Technology (IICAJET)*, Sep. 2020, pp. 1–6, doi: 10.1109/IICAJET49801.2020.9257865.
- [14] T. Anvarjon, Mustaqeem, and S. Kwon, "Deep-Net: a lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, Sep. 2020, doi: 10.3390/s20185212.
- [15] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, May 2020, doi: 10.1016/j.bspc.2020.101894.
- [16] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)*, Feb. 2017, pp. 1–5, doi: 10.1109/PlatCon.2017.7883728.
- [17] S. Tripathi, A. Ramesh, A. Kumar, C. Singh, and P. Yenigalla, "Learning discriminative features using center loss and reconstruction as regularizer for speech emotion recognition," *Workshop on Artificial Intelligence in Affective Computing*, pp. 44–53, 2019.
- [18] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," *arXiv preprint arXiv:1906.05681*, Jun. 2019.
- [19] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram and phoneme embedding," in *Interspeech*, Sep. 2018, pp. 3688–3692, doi: 10.21437/Interspeech.2018-1811.
- [20] Y. Fang, H. Fu, H. Tao, X. Wang, and L. Zhao, "Bidirectional LSTM with multiple input multiple fusion strategy for speech emotion recognition," *IAENG International Journal of Computer Science*, vol. 48, no. 3, pp. 1–6, 2021.
- [21] Z. Hu, S. Tang, Y. Luo, F. Jian, and X. Si, "3Dacrn model based on residual network for speech emotion classification," *Engineering Letters*, vol. 29, no. 2, pp. 400–407, 2021.
- [22] V. Chernykh and P. Prikhodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, Jan. 2017.
- [23] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech*, Sep. 2015, pp. 1537–1540, doi: 10.21437/Interspeech.2015-336.
- [24] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," in *Interspeech*, Sep. 2018, pp. 257–261, doi: 10.21437/Interspeech.2018-1625.







- [25] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.
- [26] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, pp. 3147–3155, 2017.
- [27] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, Sep. 2005, pp. 1517–1520, doi: 10.21437/Interspeech.2005-446.
- [28] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, May 2018, doi: 10.1371/journal.pone.0196391.
- [29] C. Busso *et al.*, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.
- [30] J. Parry *et al.*, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Interspeech*, Sep. 2019, pp. 1656–1660, doi: 10.21437/Interspeech.2019-2753.
- [31] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7194–7198, doi: 10.1109/ICASSP40776.2020.9053581.

## BIOGRAPHIES OF AUTHORS







**Kah Liang Ong**     received his bachelor's degree in Information Technology (Hons.) Artificial Intelligence from Multimedia University, Malaysia, in 2021. Currently, he is a full-time master's student and his current research interest is speech emotion recognition which mainly involves audio pre-processing, feature extraction, and emotion classification. He can be contacted at 1161200577@student.mmu.edu.my.







**Chin Poo Lee**     is a senior lecturer in the Faculty of Information Science and Technology at Multimedia University, Malaysia. She completed her Master of Science and Ph.D. in Information Technology in the area of abnormal behaviour detection and gait recognition. She is a certified Professional Technologist since 2018, a member of the International Association of Engineers since 2020 as well as an Outcome-Based Education Consultant and Trainer. Her research interests include action recognition, computer vision, gait recognition, natural language processing, and deep learning. She can be contacted at cplee@mmu.edu.my.



**Heng Siong Lim**     received his B.Eng. (Hons) Degree in Electrical Engineering from Universiti Teknologi Malaysia in 1999. He obtained his M.Eng.Sc. and Ph.D. in Engineering focusing on signal processing for wireless communications from Multimedia University in 2002 and 2008 respectively. He is currently a professor at the Faculty of Engineering and Technology, Multimedia University. His current research interests are in the areas of signal processing for advanced communication systems, with emphasis on detection and estimation theory as well as their applications. He can be contacted at hslim@mmu.edu.my.



**Kian Ming Lim**     received B.IT. (Hons.) in Information Systems Engineering, Master of Engineering Science, and Ph.D. (I.T.) degrees from Multimedia University. He is currently a Lecturer with the Faculty of Information Science and Technology, Multimedia University. His research and teaching interests include machine learning, deep learning, and computer vision and pattern recognition. He can be contacted at kmlim@mmu.edu.my.