2022 25th International Conference on Computer and Information Technology (ICCIT)
17-19 December, Cox's Bazar, Bangladesh

# Multimodal Emotion Recognition Using Heterogeneous Ensemble Techniques

A.M. Esfar-E-Alam
*Department of Computer Science and Engineering and Engineering*
*BRAC University*
Dhaka-1207, Bangladesh
esfar.alam@bracu.ac.bd

Mehran Hossain
*Department of Computer Science*
*BRAC University*
Dhaka-1207, Bangladesh
mehran.hossain@g.bracu.ac.bd

Maria Gomes
*Department of Computer Science*
*BRAC University*
Dhaka-1207, Bangladesh
maria.gomes@g.bracu.ac.bd

Rafidul Islam
*Department of Computer Science*
*BRAC University*
Dhaka-1212, Bangladesh
sheikh.md.rafidul.islam@g.bracu.ac.bd

Ramisha Raihana
*Department of Computer Science*
*BRAC University*
Dhaka-1207, Bangladesh
ramisha.raihana@g.bracu.ac.bd

2022 25th International Conference on Computer and Information Technology (ICCIT) | 979-8-3503-4602-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCIT57492.2022.10054720

***Abstract*—Emotion recognition and sentiment analysis serve several purposes, from analyzing human behavior under specific conditions to the enhancement of customer experience for various services. In this paper, a multimodal approach is used to identify 4 classes of emotions by combining both speech and text features to improve classification accuracy. The methodology involves the implementation of six models for both audio and text domains combined using four different heterogeneous ensemble techniques - hard voting, soft voting, blending and stacking. The effects of each ensemble method on the accuracy for the multimodal classification task are also investigated. The results of this study show that the usage of ensemble learning to combine modalities greatly improves classification, with stacking being the best-performing ensemble technique for the selected collection of models. The proposed model outperforms several existing methods for 4-class emotion detection on the IEMOCAP dataset, obtaining a weighted accuracy of 81.2%.**

***Index Terms*—multimodal, ensemble learning, emotion recognition, stacking, IEMOCAP**

## I. Introduction

Automated emotion recognition has become increasingly relevant due to its contributions to areas such as social media and product management, advertisement and marketing techniques. Emotions can be classified based on facial cues, gestures, speech, text or even EEG signals from the brain. However, due to the limited access to human features that most computers may have in real-life scenarios, this study focuses on speech and text as they are more attainable compared to facial expressions or brain signals.

Previous unimodal approaches focus solely on a single channel of expression. However, this overlooks the contribution of different features if used in conjunction with each other. To elaborate, the sentiments found in text can be interpreted in different ways depending on the tone and vocal features. The same can be said for speech, where similar speech features can be found in different expressions and can only be differentiated by investigating the spoken content, in other words, the lexical data.

A unimodal approach would, therefore, result in a loss of context needed for classification in certain scenarios. To utilize the correlation between text and speech, a bimodal approach has been used, which seeks to use both sets of features to train the base classifiers separately. The classifiers are then combined using four different ensemble techniques, after which the best performing one is chosen for a combined prediction.

The novelty of our work lies within the way we joined the two modalities. To do so, we implemented four different ensemble techniques which can combine results from different models. Among them, voting techniques like hard and soft voting have been tried before in previous approaches like in [15] and [9]. However, two-layer ensemble methods such as Stacking and Blending have not been suggested in other works for this task to the best of our knowledge.

Considering the challenges faced in existing methods and the motivation behind this study, the research objectives of this paper are:

- Using double-layered ensemble techniques such as Stacking and Blending to combine modalities alongside old Voting methods
- Comparing the performance of different heterogeneous ensemble techniques
- Merging acoustic and linguistic information to recognize emotions of a speaker.

This paper is structured as follows: Section 2 briefly discusses existing methods that have been used for emotion recognition. Section 3 focuses on the dataset that was chosen for this study and its preprocessing, and Section 4 explains the features that have been extracted from the data, divided into two parts for each modality. Section 5 involves a detailed description of the different ensemble techniques used, the

results of which are compared in Section 6. Lastly, Section 7 concludes the paper.

## II. Related Work

While earlier approaches involved hidden Markov models [1], [2], Gaussian mixture models [3], [4], and SVMs [5], recent studies have focused more on neural network architectures which have proven to be effective for this task [6], [7]. For instance, J.Joy et al. [8] explored emotion recognition from speech using MultiLayer Perceptron (MLP) networks where results showed the importance of selecting appropriate speech features like MFCCs. In [7], E.Batbaatar et al. used CNN and BiLSTM models on text data. The authors compared different models and found that deep learning models outperform traditional machine learning models, with their proposed model giving the best scores.

Multimodal models have also been proposed by different scholars and have shown improvements over their constituent unimodal methods [9]–[11]. For example, G. Sahu achieved impressive results in [9] by soft voting twelve classifiers, six for each mode. Using prosodic features and text transcriptions, the author classified six emotions on the IEMOCAP dataset [12] with approximately 14% higher accuracy compared to either mode alone. S.Yoon et al. [13] used dual recurrent neural networks (RNNs) to process speech and text data simultaneously while S.Tripathi et al. [14] used a different fusion method where the best architecture for each model classification was found first and fusion was performed at the final layer.

Ensemble models have been shown to be more effective than individual models in previous works [15]–[18]. An ensemble of Random Forest (RF), Gradient Boosting (GB) and Hist GB was tried by N.T.Ira et al. [15] with a voting classifier which improved results by 10 to 13% when compared to independent models. Another study by C.Zheng et al. [18] used weighted voting in an ensemble system of three deep learning models focusing on different aspects of emotion recognition using local features of spectrograms and local statistical features, where they found that two models were capable of correcting misclassifications of the third model to achieve better scores.

We incorporate multimodality along with ensemble learning as they enhance performance. Furthermore, we closely investigate the effect of different ensemble learning techniques on our model.

## III. Dataset and Preprocessing

The IEMOCAP [12] dataset was used for the multimodal classification task as it consisted of 9 emotions and over 10,000 audio recordings, both scripted and improvised, with transcripts across several sessions. The availability of both audio and its corresponding textual data makes it a good fit for the multimodal approach used in this study. Due to the imbalanced nature of the dataset, sparse emotions such as disappointment, surprise, frustration and fear were dropped along with ambiguous categories such as 'xxx' and 'oth' (others). Excited and happy labels were merged to happy as well due to the closeness of the categories.

Preprocessing thus resulted in a dataset containing 4 major emotions- anger, neutral, happiness and sadness across 5531 samples. The raw audio data was used for speech feature extraction and the textual data for each audio was processed further by removing stop words and punctuation while also lowercasing characters. Some symbols that were indicative of emotions such as '?' and '!' were not removed.

## IV. Feature Extraction

### A. Speech Features

To extract meaningful information from the audio files, spectral features such as MFCCs and Mel-Filter Banks were extracted alongside pitch, which is a prosodic feature. Features extracted from the vocal tract system are called spectral features, which are in the frequency domain. They provide information regarding the movement of articulators and the nature of the vocal tract. On the other hand, prosodic features are concerned with rhythm, stress and intonation. They can extract emotional expression or excited behaviors. The conversations available in the dataset are broken down into sentences, with each sentence having its corresponding labeled file, from which the features were extracted. The following features were extracted for this study:

- MFCC: Mel-Frequency Cepstral Coefficients provide information regarding the shape of the speech signal spectrum. After windowing the speech signal, discrete Fourier transforms are applied. The log of the magnitudes are taken and the frequencies are warped on the Mel scale, after which an inverse discrete cosine transform is used. The first 40 MFCCs were used in this study.
- Mel-Filter Banks: They mimic the nonlinear frequency feature of human ears, which is the ability in humans to better differentiate between lower frequencies than between higher frequencies. The Mel scale thus makes it easier to identify differences in lower frequencies with the help of filter banks that separate the input signal into multiple components.
- Chroma: It represents pitches from the 12 different pitch classes in audio, providing the tonal content of the signal.

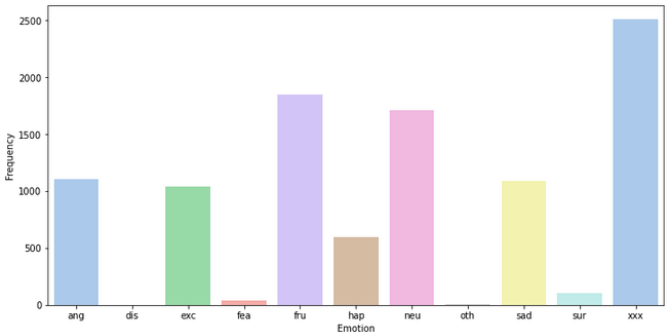


Fig. 1. Frequency of the categories before preprocessing

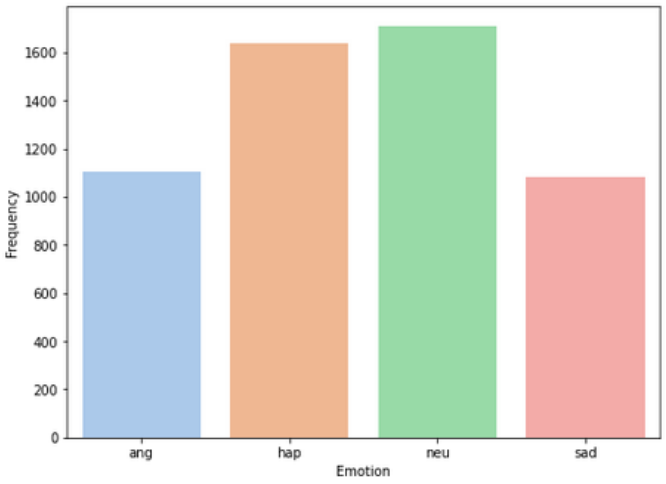


Fig. 2. Frequency of the categories after preprocessing

- Root Mean Square Energy: It is the overall energy of the speech signal, which can also denote loudness. It can be a good indicator of angry or sad emotions.
- Zero Crossing Rate: It is the number of times the signal changes its polarity.
- Spectral Flux: For an audio signal, this is the spectral change between two consecutive frames. It is calculated by finding the difference between the spectral magnitude of two successive windows and squaring that value.
- Spectral Roll-Off: It gives an idea of the frequency under which a certain amount of energy can be found.
- Pitch: It is a measure of the frequency of sound.
- Contrast: It is defined as the difference in decibels between spectral peaks and spectral valleys in a speech signal.

### *B. Text Features*

Several text feature extraction techniques exist, such as Bag of Words, TF-IDF and Word2Vec. The extraction of text features involves vectorization, which is the process of encoding raw text data into floating-point or integer values to use as inputs in machine learning algorithms. In this study, the Count Vectorization method was chosen to transform text data into vectors on the basis of word frequency only.

After the initial preprocessing stage, the tokens, which consisted of individual words in the text, were converted to vectors to build the vocabulary. In this implementation, both unigrams and bigrams were taken into account as a pair of words often conveys more meaning than a singular one. Words with a document frequency lower than 5 were ignored while building the vocabulary as well.

## V. Proposed Methodology

A multimodal approach has been used in this study, which involves six popular base classifiers that learn from features in the speech and text domain. The base classifiers used are Logistic Regression, Random Forest, Support Vector Machine,

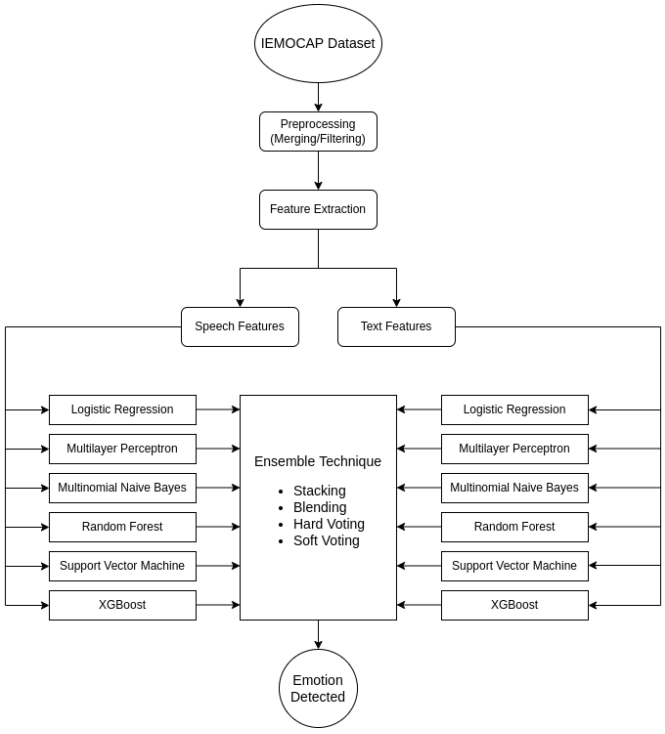


Fig. 3. Workflow of the proposed model

Naive Bayes, XGBoost and a neural network model, the Multilayer Perceptron. As stated in Section 1, emotional information in text features provides additional context with acoustic features that is necessary to identify the correct emotion. Features found in either acoustic or linguistic domains alone are limited, resulting in a lack of context. Moreover, it is evident from similar approaches such as [9], [14], [17], [19] that speech features complement the lexical data and lead to a more accurate classification. In this study, each base classifier is trained on the two modalities separately, resulting in twelve trained models. They are then combined on a decision level using an ensemble technique, which provides the final prediction.

Ensemble techniques aim at improving the accuracy of results by combining multiple models. They utilize the advantages of several different methods to counteract each model's individual weaknesses and can be either homogeneous or heterogeneous. The former combines multiple instances of the same type of model, such as bagging or boosting. However, as the purpose of this research is to combine different models, the following heterogeneous ensemble techniques have been used, after which the best-performing one was chosen as the final model:

- Stacking: First talked about in [20], this method has a series of base classifiers on level-0 and a meta-classifier on level-1 which performs the job of combining the output of the base models. The base models are trained and their predicted probabilities are used as inputs for training the meta-model. A k-fold-like approach is taken

here to cover the whole training set without overfitting. Each base classifier is then trained a second time on the whole training set.

- Blending: It was introduced as a modified stacking model which won the Netflix Grand Prize in 2009 [21]. It is similar to stacking, but instead of a k-fold approach, a portion of the training data is held out as a validation set. Each base model is then trained on the training set and predicted on the validation set. The predicted probabilities are given to the meta-classifier as training data.
- Hard Voting: By far the simplest method of combining models, predictions are done on all base models and the emotion that was chosen by the majority of the models is the final output.
- Soft Voting: In this method, all the probabilities obtained from each base model for predictions are averaged and the emotion with the highest probability afterwards is the final prediction.

## VI. Experimental Results & Discussion

The performance of each model after carrying out a 5-fold cross-validation is given in Table 1. It was found that each ensemble method outperforms the base models, with the stacking ensemble (E4) giving the highest weighted accuracy (WA) of 81.2%. Hard Voting performed the worst with a WA of 74.9%, while Soft Voting and Blending showed a very similar performance giving a WA of approximately 79%. Due to the imbalance in the dataset, WA was chosen as the core metric to evaluate the models. The macro-averaged F1 score for each model was also used as a secondary metric.

As seen in Table 1, both blending (E3) and stacking (E4) outperform the voting methods in terms of weighted accuracy, which shows that the use of a meta-model in combining classifiers can yield better results in an ensemble system. Stacking gives a better result compared to blending which can be explained by its use of the k-fold approach, allowing the meta-classifier to be trained on a larger portion of the dataset as opposed to a hold-out validation set only.

TABLE I
Performance of each model (Weighted Average and F1 scores)

| **Model** | **WA(%)** | | **F1(%)** | |
|---|---|---|---|---|
| | **Text** | **Speech** | **Text** | **Speech** |
| Logistic Regression | 65.2 | 63.3 | 66.2 | 63.2 |
| MLP | 65.2 | 64.5 | 65.9 | 64.0 |
| Naive Bayes | 63.8 | 46.9 | 64.6 | 46.4 |
| Random Forest | 63.5 | 61.5 | 64.7 | 68.6 |
| SVM | 62.6 | 67.3 | 64.3 | 67.0 |
| XGBoost | 63.6 | 68.4 | 64.7 | 68.6 |
| E1 (Hard Voting) | 74.9 | | 75.5 | |
| E2 (Soft Voting) | 78.7 | | 79.5 | |
| E3 (Blending) | 78.9 | | 79.0 | |
| **E4 (Stacking)** | **81.2** | | **81.5** | |

The confusion matrix for E4 has been given in Fig. 4. It shows similar accuracies for the sad, happy, and angry classes. The neutral class tends to be misclassified as happy or sad, giving it a lower score. Furthermore, the other emotions are most commonly misclassified as neutral. This could be due to the fact that while other emotions may have more pronounced features and characteristics, the neutral class does not. As this class signifies the absence of any specific emotion, the classification task may be more difficult due to its ambiguous nature. In contrast, the model shows better performance when identifying the other classes of emotions.

In Table 2, it can be seen that the stacking ensemble (E4) outperforms most implementations from recent years tested on four emotions on IEMOCAP. Papers using a multimodal approach such as [11], [22]–[24] focus on combining individual models for text and speech while incorporating deep learning methods. However, the results of this paper show that using a larger number of models for each modality leads to better performance. Furthermore, we confirm that simpler machine learning models in an ensemble system can perform on par with state-of-the-art methods.

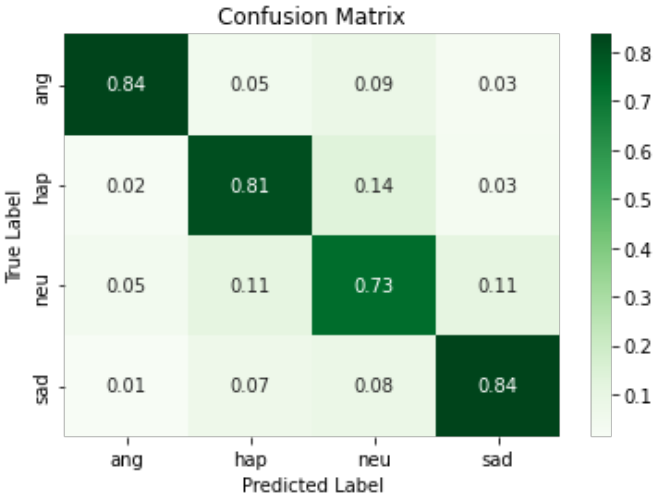


Fig. 4. Normalized Confusion Matrix showing per-class accuracy of Stacking

TABLE II
Comparison with state-of-the-art results for classifying four emotions on the IEMOCAP dataset

| **Reference** | **WA(%)** | **UA(%)** |
|---|---|---|
| Tripathi (2019) [22] | 69.5 | 76.1 |
| Cai (2019) [11] | 70.4 | 71.3 |
| Xu (2020) [10] | 70.4 | 69.5 |
| Mustaqeem (2020) [6] | - | 72.3 |
| Makiuchi (2021) [23] | 73.5 | 73 |
| Zheng (2019) [18] | 75.0 | 75.0 |
| Atmaja (2019) [19] | - | 75.5 |
| Yoon (2019) [25] | 76.5 | 77.6 |
| Lian (2020) [24] | 82.7 | - |
| **Proposed (E4)** | **81.2** | **80.8** |

## VII. Conclusion

In this paper, a multimodal approach towards emotion recognition has been taken with the use of four different ensemble methods. Six base classifiers were trained on speech and text data separately and put together in a single model. The findings show that each ensemble performs better than the individual base models, with the stacking ensemble giving the highest accuracy of 81.2%. Although the number of samples for training was limited and transcriptions may not always match the given emotion label, the proposed implementation achieved results which surpass previous research on the same dataset, showing that stacking simple machine learning models is highly effective for increasing overall performance in emotion classification tasks. The research draws attention to the efficacy of hetergeneous ensemble techniques for emotion recognition and serves as a stepping stone for further investigation of this domain. In the future, other datasets can be incorporated to test the consistency of the proposed model or include more deep learning models in the ensemble as they have been proven to be very effective for the task of emotion recognition.

## References


[1] B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), 2003, pp. I-401, doi: 10.1109/ICME.2003.1220939.

[2] K. Lu and Y. Jia, "Audio-visual emotion recognition with boosted coupled HMM," Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2012, pp. 1148-1151.

[3] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in Ninth International Conference on Spoken Language Processing, 2006.

[4] L. He, M. Lech, N. Maddage, S. Memon and N. Allen, "Emotion Recognition in Spontaneous Speech within Work and Family Environments," 2009 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009, pp. 1-4, doi: 10.1109/ICBBE.2009.5162772.

[5] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1057–1070, 2011.

[6] Mustaqeem, M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM" in IEEE Access, vol. 8, pp. 79861-79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[7] E. Batbaatar, M. Li and K. H. Ryu, "Semantic-Emotion Neural Network for Emotion Recognition From Text," in IEEE Access, vol. 7, pp. 111866-111878, 2019, doi: 10.1109/ACCESS.2019.2934529.

[8] J. Joy, A. Kannan, S. Ram and S. Rama, "Speech Emotion Recognition using Neural Network and MLP Classifier", 2020 International Journal of Engineering Science and Computing (IJESC), 2020, Vol. 10 No. 4, pp. 25170-25172.

[9] G. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution", ArXiv, 2019, doi: 10.48550/arXiv.1904.06022.

[10] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng and X. Li, "Learning Alignment for Multimodal Emotion Recognition from Speech", Interspeech 2019, 2019, pp. 3569-3573, doi: 10.21437/Interspeech.2019-3247.

[11] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks", Mathematical Problems in Engineering, 2019, pp.1-9, doi: 10.1155/2019/2593036

[12] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.

[13] S. Yoon, S. Byun and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 112-118, doi: 10.1109/SLT.2018.8639583.

[14] S. Tripathi., S. Tripathi, and H. Beigi "Multimodal emotion recognition on IEMOCAP with neural networks", 2019, arXiv:1804.05788v3.

[15] N. T. Ira and M. O. Rahman, "An Efficient Speech Emotion Recognition Using Ensemble Method of Supervised Classifiers," 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), 2020, pp. 1-5, doi: 10.1109/ETCCE51779.2020.9350913.

[16] W. Zehra, A.R. Javed, and Z. Jalil, "Cross corpus multi-lingual speech emotion recognition using ensemble learning", Complex Intell. Syst. 7, pp. 1845–1854, 2021, https://doi.org/10.1007/s40747-020-00250-4.

[17] K. Noh, J. Lim, S. Chung, G. Kim and H. Jeong, "Ensemble Classifier based on Decision-Fusion of Multiple Models for Speech Emotion Recognition," 2018 International Conference on Information and Communication Technology Convergence (ICTC), 2018, pp. 1246-1248, doi: 10.1109/ICTC.2018.8539502.

[18] C. Zheng, C. Wang. and N. Jia, "An Ensemble Model for Multi-Level Speech Emotion Recognition", Applied Sciences, 10(1), p.205, doi:10.3390/app10010205.

[19] B. T. Atmaja, K. Shirai and M. Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 519-523, doi: 10.1109/APSIPAASC47483.2019.9023098.

[20] D.H. Wolpert, "Stacked generalization", Neural Networks, 1992, vol. 5 issue 2, pp. 241-259, doi: 10.1016/S0893-6080(05)80023-1.

[21] J. Sill, G. Takacs, L. Mackey, D. Lin, "Feature-Weighted Linear Stacking", arXiv, 2009, doi: 10.48550/arXiv.0911.0460.

[22] S. Tripathi, A. Kumar, A. Ramesh, C.Singh, P. Yenigalla, "Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions", CICLing 2019: 20th International Conference on Computational Linguistics and Intelligent Text Processing, 2019, doi: 10.48550/arXiv.1906.05681.

[23] M. R. Makiuchi, K. Uto and K. Shinoda, "Multimodal Emotion Recognition with High-Level Speech and Text Features," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 350-357, doi: 10.1109/ASRU51503.2021.9688036.

[24] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, R. Li, "Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition", Proc. Interspeech 2020, 394-398, doi: 10.21437/Interspeech.2020-1705.

[25] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, doi: 10.1109/ICASSP.2019.8683483.