www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1156-1162

e-ISSN : 2583-1062
Impact Factor : 7.001

SPEECH EMOTION RECOGNIZATION USING MACHINE LEARNING-A REVIEW

Divakar Sathivada[1]

[1]Computer Science Engineering, GMR Institute of Technology Rajam,
sathivadadivakar@gmail.com

ABSTRACT

The present paper does an exhaustive study of speech emotion recognition systems (SER), elaborating on various methodologies for detecting emotions from spoken audio signals. The main objective of SER is a vital component of Human-Computer Interaction (HCI) that identifies a user's correct emotion through speech. In the study, multiple machine learning approaches have been combined, including SVM, RF, CNN, and GMM. It applies feature extraction techniques like Mel-frequency cepstral coefficients and modulation spectral features to improve emotion classification. The research uses several datasets, among them the Berlin Database of Emotional Speech, Spanish database, and Ravdess dataset. The proposed methods have proved the efficacy of fusion processes between the use of feature selection, speaker normalization, and advanced machine learning means for the improvement of SER performance. These findings can have added to the growing research in speech processing and emotion recognition, with a myriad of possible applications in areas such as virtual assistants, call center analytics, and emotional analysis in psychotherapy.

Keywords: Speech Emotion Recognition, Machine Learning,Feature Extraction, Support Vector Machines, Gaussian Mixture Model, Convolutional Neural Network, Emotion Classification.

1. INTRODUCTION

Speech Emotion Recognition (SER) is the key part of the human-computer interaction that enables machines to recognize the emotional tone in a speech and react accordingly. This has a huge possibility in many areas like virtual assistants, customer service, healthcare, and education. Recognizing emotions in speech helps SER systems to increase the user experience by empathetically responding and personalizing the interactions. SER is also used in call centers to evaluate customer satisfaction, in psychotherapy to supervise patients' emotional states, and in education to promote emotionally responsive learning environments.

The difficulty of SER is the inconsistency of emotional expressions in different speakers, accents, and contexts. Emotions can be manifested in subtle ways, may overlap with other emotions, and range from cultural to backgrounds, which adds to the SER systems' difficulties in generalizing effectively. Machine learning and deep learning algorithms have stood out as the most predominant SER tools, with the models representing Convolutional Neural Networks (CNN), Random Forest (RF), Gaussian Mixture Models (GMM), and Support Vector Machines (SVM) being the most successful.

Each of them has its own specificity which is better for some applications than the others. CNN's are particularly good at SER as they can analyze spectrograms of the speech, which are some of the patterns that are connected with the pitch, tone, and intensity. Nevertheless, Random Forests can handle large datasets and are robust through the use of classifier ensembles. GMMs are probability-based methods that are computational, therefore, they are suitable for real-time applications. SVMs are the most useful for small datasets, where they build borders that divide data into emotional classes with high-dimensional features.

Despite progress, the most challenging issues include handling real-world noise, emotional variability, and cross-cultural nuances. This article reviews CNN, RF, GMM, and SVM as the methods, reveals the pros and cons of using the models in SER, and gives the plan for the future research to be more durable, flexible, and multi-modal emotion recognition systems.

2. LITERATURE SURVEY

Speech Emotion Recognition (SER) is an emerging field of machine learning and artificial intelligence research, which helps in improving human-computer interaction through the analysis of vocal cues that help machines realize humans' emotions. Using the plethora of research study, it has been established that various models, datasets, and techniques of feature extraction are possible to be used to improve the accuracy and robustness of SER systems. This literature review of SER's basic methodologies is therefore dedicated to four of the most popular models: CNN, RF, GMM, and SVM.

2.1 Convolutional Neural Networks (CNN) in SER

Convolutional Neural Networks have shown huge potential in SER in that they can process audio files transformed into spectrograms. Spectrograms are graphical representations of sound. Fayek et al. (2017) applied CNNs to spectrograms

from the IEMOCAP dataset; they achieved high classification accuracy in emotion in CNN, as CNN captures spatial hierarchies in data. CNNs exploit layers of convolutional filters meant to recognize emotional cues within frequency and time domains. The experiments showed CNNs to work best in situations of complex emotions and emotions that overlapped with each other, albeit at the cost of much computation and a high demand for training datasets

2.2 Random Forest (RF) in SER

Random Forest is an ensemble learning method of decision trees, which has become very popular for the task of SER because of its robustness and capability to deal with large high-dimensional data. Bertero et al. have recently applied the RF classifier to classify emotions from speech features extracted from RAVDESS and Emo-DB datasets. It shows that RF well manages diversified acoustic features, including Mel-Frequency Cepstral Coefficients (MFCC) and Chroma. RF mitigates over-fitting by using a number of decision trees casting votes on the final output, although it comes with some limitations in the interpretability of the decision and inability to handle complicated patterns in emotional data.

2.3 Gaussian Mixture Models in SER

Gaussian Mixture Models are statistical models that model data as a mixture of multiple Gaussian distributions. GMMs can be exploited in SER as the capability of modelling variations in emotional expression may become available. In Lee et al. (2011), GMMs were used for MFCC features derived from the Berlin Database of Emotional Speech. Therefore, while using the features from this database, there was good accuracy in the differentiation of basic emotions such as happiness, sadness, anger, and fear. GMMs are computationally efficient. This makes them well suited for real-time applications of emotion recognition. However, these also have limitations when dealing with non-linear and complex emotion boundaries

2.4 SVM in SER

Support Vector Machines are usually used in emotion classification, especially when the data is small. SVM models create a decision boundary that maximizes the margin between different emotion classes, making them pretty effective for high-dimensional feature spaces. SVMs have been used successfully to classify emotions from MFCC features in studies using the Berlin and RAVDESS datasets. Zhang et al. (2016) established that SVM is strong enough to perform in a linear separable case, but its performance does decrease when it deals with complex, overlapping classes and large datasets

2.5 Feature Extraction Techniques in SER

Feature extraction is the most prominent step throughout these models in SER. The most widely used feature characteristics are:

MFCC (Mel-Frequency Cepstral Coefficients): These coefficients capture characteristics of the speech signal closely related to human auditory perception.

Chroma Features: Encodes the 12 distinct pitch classes, which are useful for extracting harmonic content in speech.

Spectrograms and Mel-Spectrograms: Depict displays of audio frequencies over time, which CNNs learn to process well.

Tonnetz and Spectral Contrast: Encode tonal and harmonic properties that offer supplemental information to help distinguish between emotions

2.6 Datasets in SER Research

There are a few datasets that are generally used in the papers to train and test SER models are

RAVDESS: It provides highly recorded emotional speech in a database that can be used for model benchmarking.

IEMOCAP: Available multimodal data including speech and video. The researchers can dive deeper into multimodal SER.

Berlin Database of Emotional Speech (Emo-DB): This is a simple dataset for SER research having clear labeling of emotions and standard recordings.

SAVEE and MELD: Other datasets which can supplement SER research in addition to diversity in language, context, and demographics of the speaker.

3. DESIGN

The design principles and architecture behind SER systems using CNN, Random Forest, GMM, and SVM:

Step 1: Data Preprocessing: Audio data is typically normalized and segmented.

Step 2: Feature Extraction: Common features used are Mel-Frequency Cepstral Coefficients (MFCC), Mel-spectrograms, Chroma, Spectral Contrast, and Tonnetz.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 1156-1162
e-ISSN : 2583-1062
Impact Factor : 7.001

Step3: Model Selection: The selection criteria for each model, such as the dataset size, computational resources, and target performance metrics, are discussed.

Step4: System Architecture:

- For CNN: 2D convolutional layers with pooling and fully connected layers.
- For Random Forest: A series of decision trees trained on random subsets.
- For GMM: Probabilistic Gaussian distributions for emotion classification.
- For SVM: High-dimensional separation hyperplane using kernel functions.

4. METHODOLOGY

4.1 Convolutional Neural Network (CNN)

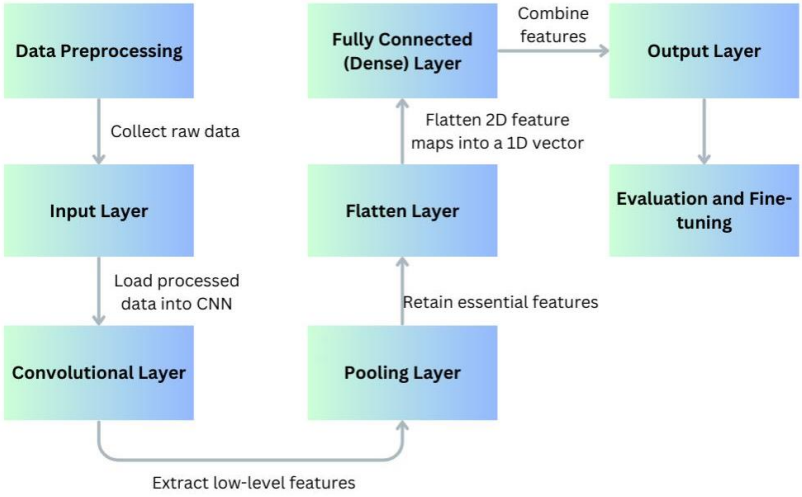


Figure 1 Convolutional Neural Network

https://www.canva.com/design

Data Preprocessing: Convert raw audio signals into spectrograms or Mel-spectrograms. These are 2D visual representations of the audio frequencies over time, ideal for CNNs.

Normalize spectrograms to standardize input data for CNN layers.

Feature Extraction and Model Architecture: Convolutional Layers: These layers extract local patterns in the spectrograms, capturing frequency and time-based features.

Pooling Layers: Apply max-pooling to down-sample the data, reducing dimensionality and preserving important features.

Fully Connected Layers: After convolutional layers, the data is passed to fully connected layers to combine features and classify the data.

Training: The CNN is trained using backpropagation with a cross-entropy loss function, and optimized with gradient descent.

Output: The model outputs a classification label for each audio sample, representing the detected emotion.

4.2 Random Forest (RF)

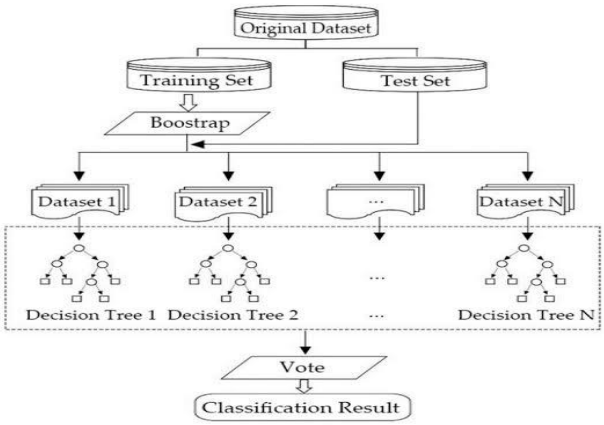


Figure 2 Random Forest[18]

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1156-1162
e-ISSN : 2583-1062
Impact Factor : 7.001

Feature extraction and construction of a feature set for a sample SER starts with an audio signal using a Random Forest classifier, typically taking into consideration the following features: Pitch, energy, and spectral properties and the MFCCs. The raw audio data forms the basis of a feature set constructed to represent each sample.

Once ready, these features are plugged into the Random Forest. An ensemble learning method, it consists of several decision trees. Each of the decision trees making up the forest works individually by partitioning feature space according to a set of rules at nodes, such that it learned portions of the overall feature-emotion relationships of the data. The trees are trained on the random subsets of data and features so that each tree specializes a little and captures various patterns within the data-like change in pitch or intensity, associating different emotions with it.

The random forest makes use of the prediction of each tree in the forest and then uses a majority vote to decide which emotion class the input sample belongs to. Random Forest combines all of the decisions for each tree into one prediction by finding the most voted emotion class. Unlike traditional ensemble methods, overfitting problems are not as problematic because of the diversity in the different decision trees, which are more likely to capture as wide a variety of emotional cues to then generate a far more accurate prediction.

Aggregated predictions from multiple decision trees yield very well-balanced classification outcome by the Random Forest classifier in SER.

4.3 Gaussian Mixture Model (GMM)

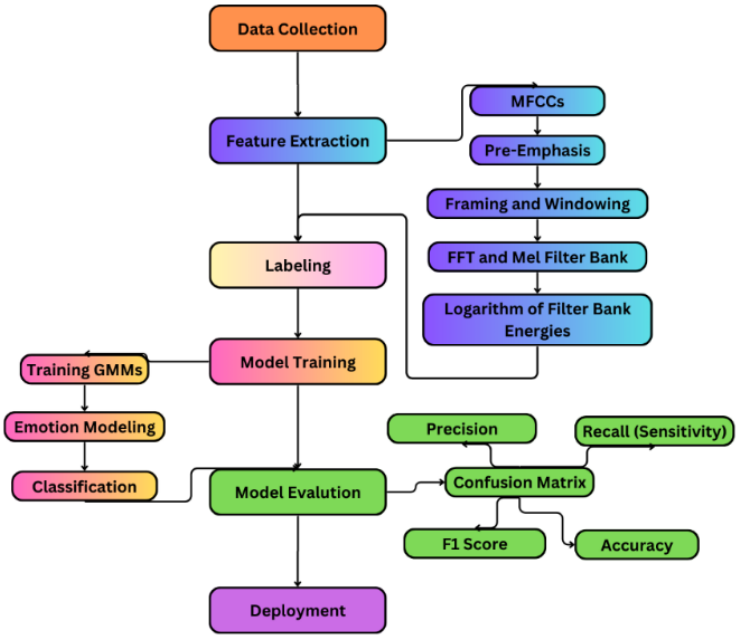


Figure 3 Gaussian Mixture Model

https://www.canva.com/design/DAGUeaIwoj4/Gm5tNEV9lIhYOyJE0A6llw/edit

- **Feature Extraction**: MFCCs or Mel-spectrogram features.
- **Model Structure**: The data is represented as a mixture of multiple Gaussian distributions.
- **Training**: Uses Expectation-Maximization (EM) to iteratively estimate parameters for each Gaussian component.
- **Gaussian Probability Density Function**:

Each component in the GMM is a Gaussian distribution, defined as

$$p(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right)$$

Equation 1https://brilliant.org/wiki/gaussian-mixture-model/

where:

- x is the feature vector (MFCCs).
- μ_i is the mean vector of the i-th Gaussian component.
- Σi is the covariance matrix of the i-th component.
- d is the dimensionality of x.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
www.ijprems.com
editor@ijprems.com
Vol. 04, Issue 11, November 2024, pp : 1156-1162
e-ISSN : 2583-1062
Impact Factor : 7.001

The probability of a feature vector x given the GMM is a weighted sum of the probabilities from each Gaussian component:

$$p(x|\lambda) = \sum_{i=1}^{K} \pi_i \cdot p(x|\mu_i, \Sigma_i)$$

Equation 2https://brilliant.org/wiki/gaussian-mixture-model/

where λ represents the GMM parameters (means, covariances, and mixture weights), π_i is the weight of the i-th Gaussian, and K is the number of Gaussian components.

Expectation-Maximization (EM) Algorithm:

- To train the GMM, the EM algorithm is used to maximize the likelihood of the data. This involves two main steps:
- **E-Step (Expectation Step)**: Calculate the responsibility $\gamma_{i|}(x)$ for each component, which represents the probability of the i-th component given the data point x:

$$\gamma_i(x) = \frac{\pi_i \cdot p(x|\mu_i, \Sigma_i)}{\sum_{j=1}^{K} \pi_j \cdot p(x|\mu_j, \Sigma_j)}$$

Equation 3https://brilliant.org/wiki/gaussian-mixture-model/

$$\pi_i = \frac{1}{N} \sum_{n=1}^{N} \gamma_i(x^{(n)})$$

Equation 4https://brilliant.org/wiki/gaussian-mixture-model/

M-Step (Maximization Step): Update the parameters (weights, means, and covariances) using the responsibilities computed in the E-step:

Update weights:

$$\mu_i = \frac{\sum_{n=1}^{N} \gamma_i(x^{(n)}) x^{(n)}}{\sum_{n=1}^{N} \gamma_i(x^{(n)})}$$

Equation 5https://brilliant.org/wiki/gaussian-mixture-model/ Update means:

$$\Sigma_i = \frac{\sum_{n=1}^{N} \gamma_i(x^{(n)})(x^{(n)} - \mu_i)(x^{(n)} - \mu_i)^T}{\sum_{n=1}^{N} \gamma_i(x^{(n)})}$$

Equation 6https://brilliant.org/wiki/gaussian-mixture-model/ Update covariances:

- **Classification**: Emotions are classified based on the probability of feature vectors belonging to each Gaussian distribution.

4.4 Support Vector Machine (SVM)

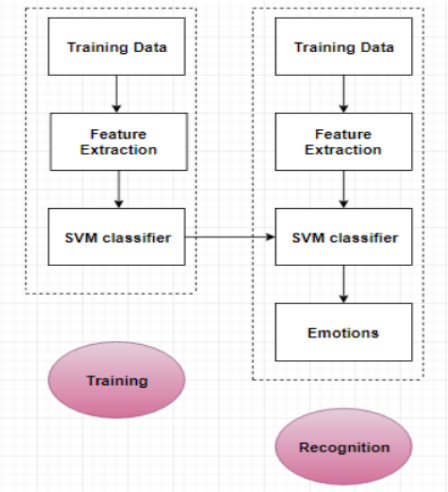


Figure 4 Support Vector Machine[17]

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 1156-1162
e-ISSN : 2583-1062
Impact Factor : 7.001

Using a Support Vector Machine (SVM) in Speech Emotion Recognition (SER), there is an onset of the process of feature extraction that is extracted from the audio. Raw audio signals, in general, are transformed into informative features that provide emotional cues like pitch, energy, and spectral features or MFCCs, and formant frequencies. Each of these informative

features captures the multiple facets of the audio signal significant for emotion discrimination.

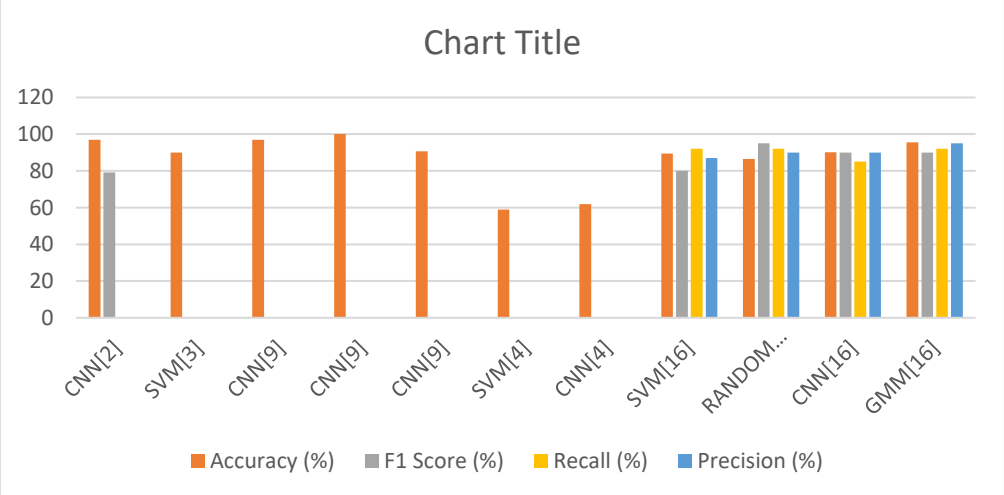
After extracting these features, they are passed on to an SVM-a type of supervised learning algorithm that focuses on finding the best hyperplane which can separate data points based on different emotion categories. In SER, the defined hyperplane maximizes the margin among other emotion category data points and hence gives a clear boundary. If such data is not linearly separable, the use of kernel functions, like radial basis function, polynomial, and linear, enables the features to be mapped into higher dimensionality spaces where separation becomes possible for SVM.

Then the SVM classifier trains itself to be able to differentiate features associated with each emotion class from each other, with resulting creations of a boundary that generalizes well to the new data. When the SVM is making the prediction, it would assign the emotion class to the data point depending on which side of the decision boundary it falls on. The SVM, therefore, can classify the speech to categories such as happy, sad, angry, etc. Therefore, this method allows the SVM to capture tiny patterns of emotion within the audio data. This is also why SVM robust performance often occurs in SER tasks when the quality and well-defined features are available.

5. RESULTS

SNO	Reference No.	Model(s) Used	Dataset(s) Used	Accuracy (%)	F1 Score (%)	Recall (%)	Precision (%)
1	2	CNN	RAVDESS	96.97	79.14	-	-
2	3	SVM	Berlin, Spanish	90-94	-	-	-
3	9	CNN	Ravdess	96.88			
4	9	CNN	TESS	100			
5	9	CNN	SAVEE	90.62			
6	4	SVM	MOSI	58.91			
7	4	CNN	MOSI	61.87			
8	16	SVM	Berlin Database of Emotional Speech	89.45	80	92	87
9	16	RANDOM FOREST	Berlin Database of Emotional Speech	86.45	95	92	90
10	16	CNN	Berlin Database of Emotional Speech	90.12	90	85	90
11	16	GMM	Berlin Database of Emotional Speech	95.6	90	92	95

Graphical Reprentation



6. CONCLUSION

In Speech Emotion Recognization final conclusion, is seen as an increasingly growing research area of immense potential in multiple applications, including healthcare, education, and human-computer interaction. Generally speaking, extraction of speech features from an audio recording and application of machine learning algorithms for emotion categories of speech are said to classify speech. SERs are typically rated based on the accuracy parameter as the prime measure.

The main outcome of SER research studies is the key results with possible interpretations that allows people to see the advantages and shortcomings of proposed systems. It appears that the future of SER research lies in increasing precision and robustness of the developed systems, explorations of new multimodal sources of data, and studying of the influence of cultural and linguistic differences on emotions' recognition.

7. REFERENCES

[1] Raja, K. S., & Sanghani, D. D. (2024). Speech Emotion Recognition Using Machine Learning. Educational Administration: Theory and Practice, 30(6 (S)), 118-124

[2] Harár, P., Burget, R., & Dutta, M. K. (2017, February). Speech emotion recognition with deep learning. In 2017 4th International conference on signal processing and integrated networks (SPIN) (pp. 137-140). IEEE.

[3] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. Social Media and Machine Learning [Working Title].

[4] Shang, Y., & Fu, T. (2024). Multimodal fusion: a study on speech-text emotion recognition with the integration of deep learning. Intelligent Systems with Applications, 200436.

[5] Chowanda, A., Iswanto, I. A., & Andangsari, E. W. (2023). Exploring deep learning algorithm to model emotions recognition from speech. Procedia Computer Science, 216, 706-713.

[6] Dixit, S., Low, D. M., Elbanna, G., Catania, F., & Ghosh, S. S. (2024). Explaining Deep Learning Embeddings for Speech Emotion Recognition by Predicting Interpretable Acoustic Features. arXiv preprint arXiv:2409.09511

[7] Rezapour Mashhadi MM, Osei-Bonsu K (2023) Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. PLoS ONE 18(11): e0291500

[8] Abdusalomov, A., Kutlimuratov, A., Nasimov, R., & Whangbo, T. K. (2023). Improved speech emotion recognition focusing on high-level data representations and swift feature extraction calculation. Computers, Materials & Continua, 77(3), 2915-2933.

[9] Ottoni, L. T. C., Ottoni, A. L. C., & Cerqueira, J. D. J. F. (2023). A deep learning approach for speech emotion recognition optimization using meta-learning. Electronics, 12(23), 4859.

[10] Shang, Y., & Fu, T. (2024). Multimodal fusion: a study on speech-text emotion recognition with the integration of deep learning. Intelligent Systems with Applications, 200436.

[11] Hosain, M., Arafat, M., Islam, G. Z., Uddin, J., Hossain, M. M., & Alam, F. (2023). Emotional Expression Detection in Spoken Language Employing Machine Learning Algorithms. arXiv preprint arXiv:2304.11040.

[12] Nigar, N. (2024). Speech Emotion Recognition Using CNN and Its Use Case in Digital Healthcare. arXiv preprint arXiv:2406.10741.

[13] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE access, 7, 117327-117345.

[14] Shaila, S. G., Sindhu, A., Monish, L., Shivamma, D., & Vaishali, B. (2023, May). Speech Emotion Recognition Using Machine Learning Approach. In International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022) (pp. 592-599). Atlantis Press.

[15] CHAKHTOUNA, A., SEKKATE, S., & Abdellah, A. D. I. B. (2024). Unveiling embedded features in Wav2vec2 and HuBERT msodels for Speech Emotion Recognition. Procedia Computer Science, 232, 2560-2569.

[16] Koti, V. M., Murthy, K., Suganya, M., Sarma, M. S., Kumar, G. V. S., & Balamurugan, N. (2024). Speech Emotion Recognition using Extreme Machine Learning. EAI Endorsed Transactions on Internet of Things, 10.

[17] Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. arXiv preprint arXiv:2002.07590.

[18] Chen, Wei & Luo, Yu-Feng & Wen, Xinyu & Zhang, Chunze & Yin, & Wu, Yingdan & Yao,. (2019). Pre-evacuation Time Estimation Based Emergency Evacuation Simulation in Urban Residential Communities. International Journal of Environmental Research and Public Health. 16. 4599. 10.3390/ijerph16234599.