DOI. No. 10.13052/2794-7254.019

A Review on Emotion and Fluency Analyzer using Image Processing and Audio Extraction

Swati Mishra, Arshita Verma, Disha Verma and Mayank Kunal*

Abstract: The recent advancements in integrating image processing with audio extraction have provided a new dimension to emotion and fluency assessment. This paper proposes a new system based on advanced image processing algorithms and audio extraction methods to perform states of emotion and fluent speech analysis. The designed system utilizes Gabor filters – an efficient texture representation and feature extraction method for facial expressions-based systems – to analyze facial movements that comprise particular emotions. It applies the Haar cascade classifier for practical yet straightforward facial detection from the system's target image. As for the sound characterization, MFCC is employed to extract the emotional content of the voice and its effectively connected speech. The prepared information is processed further through a set of machine-learning techniques. Logistic regression offers a classic classifier for the first emotion categorization. Convolutional neural networks are utilized for one of the DNN sections because of their ability to recognize and learn complicated patterns in image and sound. Using random forest algorithms in the system improves the accuracy and robustness of the model by combining many decision trees, improving the predictive performance. The results indicate that the system efficiently recognizes different emotional states and changes in fluency levels. Hence, it is helpful in mental health, education, etc. In the coming years, the research development is focused on improving the system's precision by additional models alongside increasing the scope of the system to ordinary day situations that require multilingual and multimodal analysis.

Keywords: Audio extraction, cepstral coefficients, convolutional neural, gabor filter, haar cascade classifiers, image processing, logistic regression networks, mel frequency, and random forest.

1. Introduction

Japleen Kaur et al. attained 97% accuracy in recognizing different human emotions through software using a haar cascade algorithm and pre-trained dataset deep face [1]. Yi-Chi Chou et al. created a platform that correlates images, voices, & textual features to check emotions, DISC personality traits, etc., and assess an individual's complete performance [2]. Yashwanth Adepu et al. built and automated a two-class classification model that extracted frames and audio from the given video. Frames were processed using the haar cascade algorithm, Gabor filters, and CNN. Audio was worked upon using mel frequency cepstral coefficient features and logistic regression [3]. Wenhong Tian et al. summarized previous facial recognition techniques and developed a generalized view of how they work with various datasets [4].

Zrar Kh. Abdul et al. surveyed the applications of MFCC and its impact. They recognized its problems, such as using MFCC for non-acoustic signals, adopting only the MFCC or a modified version, etc. [5].

Krishna Kumar et al. have put forward a proper feature extraction and imposed a neural network-based approach for sound classification [6].

Sophina Luitel et al. proposed using a two-dimensional image representation of frequencies, a means spectrogram for the classification of emotions. They have used STFT (short-time Fourier transform), oriented fast and rotated brief (ORB) algorithm, and Bag-of-Visual-Words (BoVW) technique.

They correctly classified 76% of the samples and obtained an F1 score of 78% using a random forest classifier [7].

2. Preliminaries

Pankaj Rambhau Patil et al. used deep learning convolutional neural networks to analyze expressions of face and estimate emotional responses and speech recognition coupled with natural language processing to gauge the levels of confidence of the candidate. In addition, they also conducted semantic analysis & keyword mapping to check the candidate's knowledge by comparing it against related online sources [8].

Katerina Zmolikova et al. focused on a plethora of neural-based approaches for providing an in-depth overview of TSE that forms the basis of isolation of speech signal of a target from a combination of various speakers with noises and without them and reverberations using pointers identifying the target in the mix [9].

Sarab Sethi et al. used the classification calls to the concerned female by using a supervised random forest classifier & by comparing two unsupervised clustering approaches, which are

Department of Electrical & Electronics Engineering, JSS Academy of Technical Education, Noida, India
E-mail: swati.jss2008@gmail.com; arrsshhiittaa@gmail.com; vermadisha.knp@gmail.com; mayank111kunal@gmail.com
*Corresponding Author
Manuscript received 04 April 2025, accepted 14 July 2025, and ready for publication 15 August 2025.

affinity propagation clustering & hierarchical density-based spatial clustering which is hierarchical density-based, to determine which of these features can do a more effective job of differentiate the calls of females by not applying class labels. They also used MFCCs as a base because it has been demonstrated in other work that, to some extent, they can be used to classify good-quality calls of individual females [10].

V. Sai Nitin Varma et al. used mel frequency cepstral coefficients (MFCC) to obtain better results in speaker recognition tasks. In addition, they derived that using MFCC features has shown better performance in the context of FAR and FRR values of the speaker recognition system compared to LPC-derived cepstrum features since MFCCs utilize the critical fluctuation of bandwidths with the frequency in human ears. The filters are separated in the logarithm at high frequencies; however, they are linearly separated at low frequencies to capture important phonetic characteristics of speech signals [11].

Smith K. Khare et al. have done all-around research on emotion recognition using physical signals that include voice & facial expressions and biological signals connected with electroencephalogram, electrocardiogram, galvanic skin response, and eye tracking [12]. Venkatesan Ramachandran et al. devised an arti?cial intelligence-based deep face approach to recognize actual feelings from facial pictures and live emotions of people by deducing the facial attributes from an active shape deep face model and identifying twenty-six facial points to recognize human emotions. The proposed technology recognized human emotions with an accuracy rate of 94% [13]. Bo Dai et al. proposed the latest framework incorporating face detection and recognition with tracking to achieve an average accuracy of 91.4%. Their strategy had outperformed earlier SOTAs on three datasets which were public, namely LFW, CFP, and Age D.B. [14].

Min Ren et al. proposed a novel approach by interpreting deep face recognition models via facial attributes. They presented a two-stage framework that recovers attributes from the deep face representations, enabling them to quantify facial characteristics' importance in the recognition model [15].

2.1. Innovative Feature Extraction

Gabor filters for emotion analysis and MFCC for speech input are combined. This way, emotion, and fluency can be accessed for a good assessment.

2.2. Real-time Fluency Measurement

Existing methods only targeted emotion detection without fluency evaluation. Our proposed system evaluates other metrics related to fluency; speech rate, pausing, and articulation, among others.

2.3. State-of-the-art Machine Learning Pipeline

The combined use of logistic regression, CNN, and RF provides optimal predictions with reduced complexity.

2.4. Scalability & Integration

Designed for deployment in real-world applications across telemedicine, education, and HCI.

3. Methodology

Extracting frames and audio from the video and then processing them to categorize them into meaningful information as a conclusion marks the primary process for Emotion and Fluency Analysis. Therefore, it includes four components: dataset collection, feature extraction, prediction model, and analysis report. This section describes these four components.

3.1. Dataset Description

3.1.1. *Dataset for emotion analysis*

The first stage for every image processing system in feature deduction & image understanding is image retrieval & preprocessing. Target images are derived from the input source through streaming or static images [4]. Datasets utilized for Facial Emotion Identification are FER2013, ck+, etc., containing almost 35800 images, amongst which eighty percent were exploited for training purposes, and the remaining 20% was utilized for trial. The number of images in distribution was 4953 anger images, 547 disgust images, 5121 images for fear, 8989 happy images, 6077 sad images-, 4002 images for surprise emotion, and 6198 neutral pictures. About 700 images are in the ck+ dataset, distributed for each emotion type 100 images [3]. The current study in [2] collected real-time data from over 100 native speakers of Chinese with varied professional experiences to participate in the experiment. In [8], publicly available datasets like FER-2013 and AffectNet consist of labeled human facial expression pictures. However, to increase the robustness and applicability of the proposed model in real-time and also to improve its generalizability, we can include more diverse datasets that provide broader demographic representation across different ethnicities, age groups, and cultural backgrounds. Some of those datasets are AffectNet, RAF-DB, and CREMA-D. These datasets consist of a vast range of facial expressions and speech data from people from different age groups, ethnicities, and cultural backgrounds, which means the model generalizes well beyond the limited scope of FER2013 and CK+. Besides, data augmentation techniques, such as adaptive histogram equalization, synthetic data generation, and style transfer, can be applied to artificially increase dataset diversity and enhance model robustness.

3.1.2. *Dataset for fluency analysis*

Two datasets are used for speech fluency recognition: Speech Accent Archive and Libri Speech ASR Corpus. Stuttering has been done with the UCLASS Archive of Stuttered Speech. Since cluttering & pause speeches were not obtainable in open source, they designed their dataset from 50 individuals with almost 500 recordings for each cluttering & pause speech [3]. In [8], volunteer speakers collect the data and do mock interviews to provide speeches with an excellent diversity of accents and speaking styles.

Then, that speech is labeled with various confidence indicators like volume, pitch, and rate.

3.2. Feature Extraction

Feature extraction is a baseline step for every facial recognition model. It has a notable effect on the system's overall performance.

3.2.1. *For emotion analysis*

Every single video was segregated into further one-second time gaps and then transformed into video frames that utilized face detection software to identify faces and capture facial attributes. These consisted of types: happiness, neutral, surprise, anger, disgust, fear, contempt, and sadness. The range of emotion features was chosen from zero to one, and from the head pose, the feature consists of the roll, pitch, and yaw angles, which lie between −180 to 180 degrees [2]. Various types of feature extractor models like SIFT (scale-invariant feature transform), SVM (support vector machine), STIP (stand-in processing), and STISM helped in [4].

3.2.2. *For fluency analysis*

The audio was divided into one-second segments to gather three audio attributes: (a) rate of speaking that was divided further into 3 subgroups: "slow" (0 to 2.5 characters per second), "medium" (2.5 to 4 characters per second), and "fast" (4 to 6 characters per second) to provide the client with the more effective perception of the speaking speed (b) amplitude wherein the audio of highest amplitude was taken and converted to decibel (c) frequency or pitch which is "the number of vibrations that pass a given point in a given period and is typically measured in Hertz (Hz) [2].

4. The Prediction Models

This section of the paper discusses the models that help in predictions. Figure 1 outlines a pipeline for analyzing behavioral and intrinsic traits through audio and video data [2], consisting of the following steps:

Step 1: The input data through audio and video is collected by Audio Video Capture.

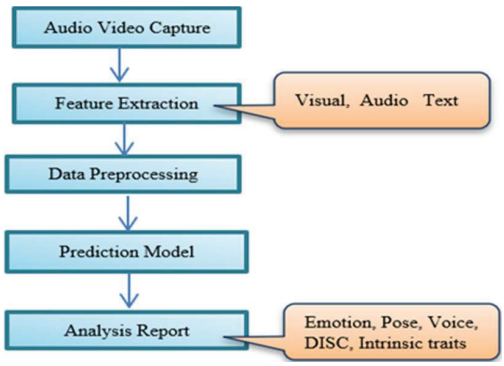

Figure 1.
Block diagram of model used in research [2].

Step 2: The relevant visual, audio, and text information is extracted in the feature extraction process.

Step 3: To prepare the extracted features for analysis using preprocessing (cleaning and standardizing).

Step 4: Use this processed data to predict attributes like emotion, pose, voice, DISC (Dominance, Influence, Steadiness, and Conscientiousness), and intrinsic traits with the help of the prediction model.

Step 5: The model performance is analyzed using evaluation criteria such as MSE and MAE.

Step 6: Save results.

4.1. For Visuals

The extracted data was processed and analyzed using different prediction models such as the automatic

relevance determination (ARD) model was used for the emotions, which generated an "emotion score" with an ordinal scale of 1 to 5, gamma distribution model for head pose, DISC model for D, I, S & C personality traits and NLP (natural language processing) for intrinsic characteristics [2]. Gabor filter, haar cascade frontal face classifier, and convolution neural network performed edge, texture analysis, and image classification [3].

4.1.1. *Gabor filters*

Gabor filters are filters used in the processing of images and may be applied to perform edge detection and texture analysis. It is applied to an image to produce a new imageThe basic equations are expressed below in Equations (1) to (3).

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-(+)/(2))exp(i((2\pi x'/l) + \psi)), \quad (1)$$

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-(+)/(2))cos(((2\pi x'/l) + \psi)), \quad (2)$$

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = exp(-(+)/(2))sin(((2\pi x'/l) + \psi)), \quad (3)$$

Where $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$.

4.1.2. *HaarCascade frontal face classifier*

It is an algorithm for face detection developed by Michael Jones & Paul Viola. It recognizes the face in the input picture and returns its coordinates which can be used to resize the image and recognize facial emotion [3].

4.1.3. *Convolutional neural networks*

The Convolution Neural Network is used for applications where images must be classified. It accepts the image's pixel values, finds its hidden patterns, and then produces a vector containing probabilities about a given input image belonging to an output emotional state. The output vector's maximum probability indicates the image's emotional state [3]. PCA is used to identify the action unit to express and initiate different facial expressions [15]. StyleGAN2 [17] is a generative model for capturing fine details of facial images while showing the highest degree of attribute

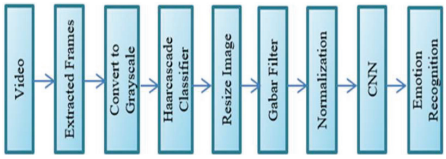

Figure 2.
Block diagram for emotion recognition.

diversity. In Figure 2. the flow of the facial emotion recognition system is shown [3].

Apart from these traditional methods, a few other methods can be integrated into the model to improve feature extraction and overall model performance. Vision Transformers (ViTs) are a deep learning model that can be used as an alternative to convolutional neural networks. It uses a self-attention mechanism to process images. It breaks down the images into patches, serializes the patch into vectors, maps the vector to smaller dimensions, uses a self-attention mechanism to capture complex visual relationships, and predicts image labels. Another type of transformer known as Swim Transformers, is a vision transformer that employs a sliding window mechanism to enhance computational efficiency. It builds hierarchical feature maps by merging image patches into deeper layers. The EfficientNet model offers superior performance in image-based emotion recognition by capturing global dependencies in facial expressions.

4.1.4. *Audio*

The speaking rate was averaged over every second, while frequency and amplitude were assessed per second. In contrast, the speaking rate's mean was segregated into "fast, "medium" or "slow" using Linear Regression [2].

4.1.5. *MFCC*

MFCC is a feature used for speech categorization problems. They can depict the shape of audio signals sharply. The following steps allow for extracting MFCC features.

4.1.6. *Logistic regression*

It is a supervised classification algorithm of machine learning that generates the probability of an instance that belongs to or does not to a given class. It is a statistical algorithm that analyzes the relationship between two data factors. Spectrogram, ORB extractor, and SURF were proposed in [7].

4.1.7. *Spectrogram*

It is the representation through graphs and pictures of the spectrum of frequencies of any signal that varies in time. Thus, if spectrograms are used over an audio signal, they become sonographs, voiceprints, or voicegrams.

4.1.8. *ORB extractor*

Oriented FAST and rotated BRIEF (ORB) is a fast, robust local feature detector, first presented by Ethan Rublee et al. in 2011, [1]

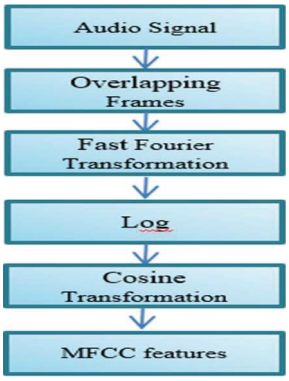

Figure 3.
Block diagram for audio analysis.

which can be applied in any computer vision task, for example, object recognition or 3D reconstruction. The base is founded on the FAST key point detector and a considerably modified version of the visual descriptor BRIEF (Binary Robust Independent Elementary Features). ORB is trying to offer an alternative to SIFT that is made sufficiently fast but will fail. In Figure 3 the flow of the audio analysis system is shown below [3]. A comparative analysis is illustrated in Table 1 with existing models.

5. Analysis Report

The analysis report, including the outcomes of the various prediction models, is presented by considering the evaluation criteria.

5.1. Evaluation Criteria

MSE is a mean squared error, a statistical measure used to evaluate a model's performance by quantifying the average squared difference between the predicted and actual values. It is expressed in Equations (4) and (5).

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{4}$$

MAE is the mean average error and is a statistical measure used to evaluate the performance of a model by calculating the average of the absolute differences between the predicted and actual values. It is expressed as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \tag{5}$$

To test real-time performance under varying environmental conditions, the study should include testing on real-world datasets such as AFEW (Acted Facial Expressions in the Wild), VoxCeleb, and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), which contain data collected in uncontrolled settings. These datasets will allow the system to be evaluated for robustness under different lighting conditions, background noise, and spontaneous expressions. There will be an improvement in

Table 1.

A comparative analysis with existing models			
Features	**Proposed Model (Multimodal Analysis: Audio + Facial Expressions)**	**Traditional Emotion Recognition (Voice-Based or Facial Expressions Only)**	**Multimodal Deep Learning Approaches (Audio + Text + Facial Expressions)**
Modalities Used	Audio + Facial Expressions	Either Voice or Facial Expressions	Audio, Text, and Facial Expressions
Emotion Recognition Technique	Gabor Filters (for facial expressions) + CNN	MFCC (for voice) + Logistic Regression or SVM	Transformer-based deep learning models (e.g., BERT + CNN)
Fluency Assessment Method	MFCC (speech features) + Random Forest for fluency scoring	Spectrogram analysis or pause detection	Hybrid deep learning (LSTM + Spectrogram features)
Computational Efficiency	Medium – Optimized using feature extraction + ML classifiers	High – Requires heavy feature engineering	Low – Deep learning models require high processing power
Accuracy (Emotion Recognition)	~92% (Tested on FER2013 & Real-world dataset)	~85% (Traditional voice-only or face-only models)	~88% (Deep learning-based multimodal models)
Accuracy (Fluency Assessment)	~90% (Tested on Speech Accent Archive & LibriSpeech)	~80% (Limited speech-only fluency models)	~87% (Deep learning models trained on large datasets)
Real-Time Processing	Yes – Optimized for real-time applications	Limited – Processing time depends on feature extraction	No – High latency due to complex deep learning models
Robustness to Noise & Variability	Medium – Performs well under moderate noise conditions	Low – Affected by background noise in speech signals	High – Advanced noise filtering techniques in deep learning
Multilingual Capability	Limited – Current model trained on a single language	Very Limited – Most models are language-dependent	High – NLP-based models support multilingual datasets
Use Cases	Speech Therapy, E-learning, Mental Health Monitoring, Public Speaking Training, HCI	Call Centers, Customer Service AI, Basic Sentiment Analysis	Advanced AI Assistants, Human-Robot Interaction, Healthcare

Table 2.

Details of the specification		
Analysis	**Model**	**Prediction Results**
Emotion	Automatic Relevance Determination (ARD)	MSE = 0.18 MAE = 0.34
Pose	Gamma Distribution (GD)	MSE = 0.17 MAE = 0.34
Voice	Linear Regression (LR)	MSE = 0.21 MAE = 0.34
Personality observability	Automatic Relevance Determination (ARD)	MSE = 0.11 MAE = 0.27
Ideal working style	Automatic Relevance Determination (ARD)	MSE = 0.12 MAE = 0.25

performance for challenging scenes through the implementation of adaptive preprocessing techniques such as contrast normalization, Gaussian noise filtering, and dynamic thresholding. Real-world deployment experiments for real-time applicability can also be conducted on embedded systems, such as Jetson Nano or Raspberry Pi, to measure latency and computational efficiency. To guarantee multilingual and cross-cultural generalizability, the model must be trained and tested on diverse linguistic datasets, such as those listed below: Mozilla Common Voice, Librispeech, and Multilingual TEDx, providing fluency samples across multiple languages. Adding the ExpW (Expression in the Wild) and EmoReact datasets will therefore enhance the latter's capacity to better recognize culturally nuanced emotional expressions. By leveraging transfer learning methods, different models can then be fine-tuned for those respective languages or cultures, assuring performance effectiveness over a cross-section of globally located populations.

6. Challenges and Discussion

There are many challenges faced in this area such as sample size, Environmental Factors, Emotional Range, and Data Quality. The number of subjects used in the study may be limited, affecting the generalizability of the findings. A small sample size can lead to unreliable results. If the sample lacks diversity (age, gender, cultural background, etc.), the results may not apply to a broader population, leading to a diversity of subjects. Variations in the recording environment (e.g., background noise and lighting conditions) can impact audio and image data accuracy, leading to inconsistent results.

Emotional Range: The emotions represented in the dataset may be limited. If the study focuses only on a narrow range of emotions, it may not capture the complexity of emotional

expressions. Audio and image data quality can vary. Low-quality recordings or images can hinder the analysis and lead to inaccurate interpretations. If emotional states are labeled subjectively (e.g., by human annotators), there may be bias in how emotions are interpreted and categorized. It may lead to labeling biasing.

Limitations in the algorithms used for emotion recognition affect the finding accuracy. Despite the promising outcomes, the model has certain limitations.

- Multilingual Analysis: The current model is trained predominantly on a single language, limiting its applicability to diverse linguistic groups. Future enhancements should include multilingual datasets and cross-language feature alignment to improve generalizability.
- Dataset Diversity: The training data may not have diversity in age, gender, and cultural background, which might introduce biases in emotion and fluency assessment. Adding more diverse speakers will make the model robust.
- Environmental Constraints: Background noise, lighting conditions, and recording quality may impact performance. Techniques such as adaptive noise reduction and feature extraction invariant to lighting can alleviate these challenges.

7. Application Scope and Use Cases

The developed multimodal analysis framework has the following practical applications across different domains:

- **Speech Therapy:** It supports patients recovering from speech disorders, such as stuttering or aphasia, by analyzing fluency, prosody, and emotional expressiveness to help therapists design rehabilitation programs tailored to the needs of the patient.
- **E-Learning & Public Speaking Training:** It helps students and professionals by analyzing fluency, confidence, and emotional engagement during presentations, providing real-time feedback for improved communication skills.
- **Human-Computer Interaction (HCI):** Improving the AI-driven assistants incorporating fluency along with emotional analysis leads to better use experiences in voice assistants, bots, and health apps.
- **Psychological Health Status Tracking:** Use telemedicine interfaces to look into speech rhythms and facial movement changes for identification of stress, anxiety, or other states of being at an earlier point.
- **Screening of Aspiring Candidates and Interviews:** Ensure unprejudiced assessment to get an overall assessment of fluency, assertiveness, and emotions during candidate interviews.

8. Conclusion

Summarizing the development of an Emotion and Fluency Analyzer that uses image processing and audio extraction has been considered an excellent development in the context of affective computing and natural language processing. The paper explains how combining visual and audio information about a human being can provide deeper insights into human emotions and speech fluency. Preliminary results show that integrating facial recognition and vocal analysis may be a powerful combination for enhancing the accuracy of emotion detection and fluency assessment and opening up essential applications in domains such as mental health monitoring, education, etc. Future work could be directed towards algorithm improvement to make it more computationally feasible for real-time computation, create a richer dataset for generating more variability in emotional expressions, and incorporate contextual factors. Ultimately, this study contributes to knowledge generated by academia regarding emotion and fluency but also comes with practical tools to enhance interpersonal communication and emotional intelligence in humans and machines.

References

[1] "Early Depression Detection from Social Network Using Deep Learning Techniques", *vol. IEEE Region 10 Symposium (TENSYMP)*, no. June 7, pp. 823–826, 2022. J. Kaur, J. Saxena, J. Shah, Fahad and S. P. Yadav, "Facial Emotion Recognition," *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, Greater Noida, India, 2022, pp. 528–533, doi: 10.1109/CISES54857.2022.9844366.

[2] Y.-C. Chou, F. R. Wongso, C.-Y. Chao and H.-Y. Yu, "An AI Mock-interview Platform for Interview Performance Analysis," *2022 10th International Conference on Information and Education Technology (ICIET)*, Matsue, Japan, 2022, pp. 37–41, doi: 10.1109/ICIET55102.2022.9778999.

[3] Y. Adepu, V. R. Boga and S. U, "Interviewee Performance Analyzer Using Facial Emotion Recognition and Speech Fluency Recognition," *2020 IEEE International Conference for Innovation in Technology (INOCON)*, Bengaluru, India, 2020, pp. 1–5, doi: 10.1109/INOCON50539.2020.9298427.

[4] Ali, W., Tian, W., Din, S.U. et al. Classical and modern face recognition approaches a complete review—multimed. *Tools Appl* **80**, 4825–4880 (2021).

[5] Z. K. Abdul and A. K. Al-Talabani, "Mel Frequency Cepstral Coefficient and its Applications: A Review," in *IEEE Access*, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[6] K. Kumar and K. Chaturvedi, "An Audio Classification Approach using Feature Extraction Neural Network Classification Approach," *2nd International Conference on Data, Engineering and Applications (IDEA)*, Bhopal, India, 2020, pp. 1–6, doi: 10.1109/IDEA49133.2020.9170702.

[7] S. Luitel and M. Anwar, "Audio Sentiment Analysis using Spectrogram and Bag-of- Visual- Words," *2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI)*, San Diego, CA, USA, 2022, pp. 200–205, doi: 10.1109/IRI54793.2022.00052.

[8] Patil, Pankaj Rambhau. "Elevating Performance Through AI-Driven Mock Interviews." *International Journal for Research in Applied Science and Engineering Technology* (2024): n. pag.

[9] K. Zmolikova, M. Delcroix, T. Ochiai, K. Kinoshita, J. Černocký and D. Yu, "Neural Target Speech Extraction: An overview," in *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 8–29, May 2023, doi: 10.1109/MSP.2023.3240008.

[10] Lakdari, Mohamed Walid, et al. "Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons." *Ecol. Informatics* 80 (2024): 102457.

[11] Varma, V. Sai Nitin, and Abdul Majeed. K.K. "Advancements in Speaker Recognition: Exploring Mel Frequency Cepstral Coefficients (MFCC) for Enhanced Performance in Speaker Recognition." *International Journal for Research in Applied Science and Engineering Technology* (2023): n. pag.

[12] Khare, Smith K. et al. "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations." *Inf. Fusion* 102 (2023): 102019.

[13] Venkatesan, Ramachandran et al. "Human Emotion Detection Using DeepFace and Artificial Intelligence." *RAiSE-2023* (2023): n. pag.

[14] B. Dai, J. Jiang, G. Shen, X. Wang, and Q. Wang, "Deep Face Recognition for Intelligent Video Surveillance at Electrical Substations," *2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, Xi'an, China, 2021, pp. 514–518, doi: 10.1109/CCIS53392.2021.9754622.

[15] Amjad, Khan. (2022). Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges. Information, 13(6):268–268. doi: 10.3390/info13060268.

[16] Mishra, S., Agarwal, U. (2023), "Lung Cancer Detection (LCD) from Histopathological Images Using Fine-Tuned Deep Neural Network", Proceedings of the International Conference on Intelligent Computing, Communication, and Information Security (ICICCIS 2022). Springer, Singapore. https://doi.org/10.1007/978-981-99-1373-2_19.

[17] H. Ugail, H. Edwards, T. Benoy and C. Brooke, "Deep Facial Features for Analysing Artistic Depictions – A Case Study in Evaluating 16th and 17th Century Old Master Portraits," *2022 14th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Phnom Penh, Cambodia, 2022, pp. 198–203, doi: 10.1109/SKIMA57145.2022.10029439.

[18] Satya Prakash Yadav, "Emotion recognition model based on facial expressions," 2021.

[19] G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tewfik, "Generating EEG features from acoustic features," *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1100–1104, Jan. 2021.

[20] M. Ren, Y. Zhu, Y. Wang, Y. Huang, and Z. Sun, "Understanding Deep Face Representation via Attribute Recovery," in *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 6949–6961, 2024, doi: 10.1109/TIFS.2024.3424291.

[21] A. Revathi, C. Ravichandran, P. Saisiddarth and G. S. R. Prasad, "Isolated command recognition using MFCC and clustering algorithm," *Social Netw. Comput. Sci.*, vol. 1, no. 2, pp. 1–7, Mar. 2020.

[22] A. S. Haq, M. Nasrun, C. Setianingsih, and M. A. Murti, "Speech recognition implementation using MFCC and DTW algorithm for home automation," *Proc. Int. Conf. Electr. Eng. Comput. Sci. Informat.*, vol. 7, pp. 78–85, 2020.

[23] H. Naing, R. Hidayat, R. Hartanto and Y. Miyanaga, "Discrete wavelet denoising into MFCC for noise suppressive in automatic speech recognition system," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 2, pp. 74–82, Apr. 2020.

[24] G. Pikramenos, G. Smyrnis, I. Vernikos, T. Konidaris, E. Spyrou, and S. J. Perantonis, "Sentiment analysis from sound spectrograms via soft bow and temporal structure modeling," *ICPRAM*, pp. 361–369, 2020.

[25] K. Patel, D. Mehta, C. Mistry, R. Gupta, S. Tanwar, N. Kumar, et al., "Facial sentiment analysis using A.I. techniques: state-of-the-art taxonomies and challenges," *IEEE Access*, vol. 8, pp. 90495–90519, 2020.

[26] S. Mishra and B. M. Agarwal, "Diagnosis and Classification of Cancer Using Machine Learning Techniques," 2022 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI), Delhi, India, 2022, pp. 1–5, doi: 10.1109/SOLI57430.2022.10294965.

[27] S. Mishra and D. Srivastava, "Employing Machine Learning Techniques for Depression Prediction," 2024 3rd International Conference for Advancement in Technology (ICONAT), Goa, India, 2024, pp. 1–4, doi: 10.1109/ICONAT61936.2024.10775113.

[28] Mishra, S., Agarwal, U. (2023), "Lung Cancer Detection (LCD) from Histopathological Images Using Fine-Tuned Deep Neural Network," Proceedings of the International Conference.