iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75678

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

An AI-Driven System for Evaluating Emotional Responses and Confidence in Mock Interviews

Gyara Amani
[1]*Research Scholar (MTech in Embedded Systems), Electronics & Communication Engineering Department, JNTUH University College of Engineering, Science and Technology, Hyderabad, Telangana, India*

Abstract: ***Emotional intelligence is very significant in the current competitive job market as it determines the success of candidates during the interview process. The presented paper is a proposal of an AI-based system that directly addresses the needs of assessing the emotional response and the level of confidence of the candidates during a mock interview. The application has two main modules, which include facial emotion detecting and audio emotion classification. To detect emotions on the face, the system uses the YOLO model that involves processing video feeds by detecting important emotions like anger, happiness, sad, fear and surprise. The emotion classification audio module employs the use of advanced models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RCNN), and a combination of CNNBlock and ConformerBlock to determine the emotional tone of the voice of the candidate, which can be anger, fear, and happiness among others. The webapp is developed with a backend written in flask and frontend written in HTML, CSS, and JavaScript to give users a very user-friendly interface to upload video and audio files. Not just does it predict the emotional state but it also gives a confidence score to each prediction which the candidate can find useful. The combination of these technologies will assist the candidates to control their emotions better and understand themselves, which will eventually make them perform better in a stressful interview situation. This study supports the possibility of AI in the future of mock interview assessment and emotional intelligence training.***
Keywords: Artificial Intelligence, Emotion Recognition, Confidence Evaluation, Mock Interviews, Facial Emotion Detection, Audio Emotion Classification, YOLO, Convolutional Neural Networks, LSTM, Hybrid Models.

I. INTRODUCTION

Human resource management has changed so much with the introduction of artificial intelligence (AI) in different industrial sectors overall. Interviewing candidates is one of the most important requirements of the recruitment procedure. The conventional interview testing methods tend to concentrate on the verbal answer and technical skill and disregards the emotional intelligence and non-verbal bodily features, which are highly important in assessing the suitability of a job applicant. Emotional reactions and the degree of confidence tend to be one of the key factors of how well a candidate performs during stress, but they are not always easy to measure and assess. The proposed project fills this gap by creating an AI-based system to assess the emotional reactions and confidence levels of the candidates during fake interviews and using facial expression analysis and audio emotion recognition.

The system contains two significant features: facial emotion recognition and audio emotion recognition to evaluate the emotional state and confidence of the candidate. Facial emotion detection uses the state-of-the-art deep learning algorithm known as the YOLO model to process video streams and identify facial expressions that are associated with specific emotions like anger, happiness, sadness, surprise, fear and neutrality. This method can be used to determine the emotional condition of a candidate during the interview and gives an idea of how an individual would behave under various circumstances e.g. feeling stressed, angry or excited.

Besides the facial emotion recognition, the system also includes the audio emotion classification that examines the tone of the speech of the candidate. Advanced neural networks like Consssvolutional neural networks (CNN), Recurrent neural networks (RCNN), and a combination of CNNBlock with ConformerBlock were utilized in this component. The system recognizes anger, fear, happiness, and sadness by analyzing speech patterns, tone and pitch, creating a list of emotions that is an all-inclusive emotional analysis that plays off facial expression analysis.

The ability to give confidence ratings to both facial and audio emotion prediction is one of the peculiarities of this system. Such scores represent the level of confidence about the emotional classification, and they give the candidates a measurable output on their emotional reactions and their general confidence. The confidence score will also aid in showing areas that candidates can improve upon to provide practical self-awareness and development.

The web application is constructed based on Flask (back-end) and HTML, CSS, JavaScript (front-end) as its two components to offer an interactive platform to the candidates to interact with the system. Users are able to post video and audio recording of their mock interviews and the system will process such recordings in real time and provide feedback of emotional responses and confidence levels. The feedback, in their turn, makes the candidates realize their emotional control and performance at the interviews, which may be useful to personal development and improving interview skills.

This system aims to fill this gap in the conventional interview reviews to provide a more comprehensive evaluation of the candidates based on emotional intelligence and technical skills. The introduction of AI into the assessment of mock interviews can be viewed as a significant milestone of the recruitment process as it will allow both job applicants and human resource managers to gain a better understanding of the emotional implications of interviews.

II. RELATED WORK

The recent development of artificial intelligence has tremendously improved accuracy and efficiency of emotion recognition systems in the visual and audio as well as multimodal field of operation. Facial emotion detection has moved to real-time detection with the introduction of optimized YOLO-based architectures. A YOLO v8 model had been shown to achieve good-speed facial expression recognition thus could be used in dynamic applications where immediate feedback is necessary like in the mock interviews [1].The advancement in the audio-visual fusion has also made it possible to have models to pick up emotional cues better by paying joint attention to complementary modalities using recursive attention mechanisms [2].

Emotion classification based on speech is still continuing to develop with hybrid deep learning systems. Speech emotion recognition systems that are powered by CNN demonstrated that they perform well in detecting the emotion tone using audio cues [3]. Multi-modal graph neural network based transformer architectures have also been used to complement conversation-level emotion detection by capturing long range dependencies [4].Further improvements of multimodal reasoning made possible by instruction-optimized large language-vision models introduce further emotional insights and contextual clues [5].

The perception of emotion in adaptive systems has been enhanced by self-context-aware structures that are capable of adapting to current human interactions on a real-time basis [6]. Recognition accuracy has been significantly improved with multimodal fusions techniques which include audio and video streams with temporal modeling [7]. Hybrid feature extraction methods that comprise combining various acoustic representation and spectral representations with CNNs have also been shown to be effective in enhancing speech emotion classification performance [8]. Studies on the exploitation of the complementary nature of speech, language, and vision modalities point out the fact that the integrated representations provide stronger recognition systems [9].

Pre-training methods that are self-supervised, like emotion2vec, have enhanced learning of features, as they are able to learn emotional properties without the need to have a large amount of labeled data [10]. The use of generative emotional audio to accomplish data augmentation has also resolved the limitations of data sets, which enhances the strength of numerous speech samples [11]. The relevance of the synchronized processing of multimodal cues has been confirmed by other researchers in recursive joint attention in fusion-based regression models [12]. The context-sensitive emotion perception in human-robot interaction is an on-going development based on adaptive frameworks in realistic settings [13].

More advances on multimodal emotion recognition include transformer and graph-based fusion schemes [14] and temporal audiovisual modeling models that are aimed at providing strong cross-modal integration [15].

Together, these works signify the increasing usefulness of the deep learning, multimodal fusion, and real-time processing methods, which contributes to the evolution of a combined system to assess emotional reactions and confidence level regarding mock interview situations

III. DATA SET

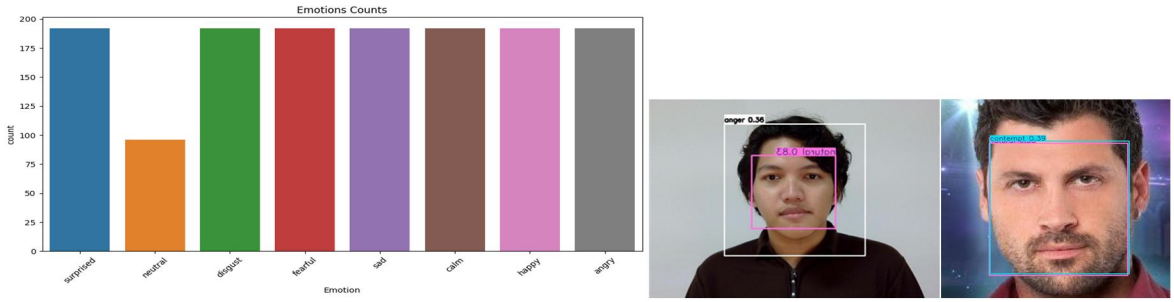


Figure 1 Class distribution for audio data

IV. PROPOSED SYSTEM

The AI-powered Mock Interview Evaluator will set out to determine the emotions of the candidates and the levels of confidence during the mock interview process by examining the facial expressions and audio provided by the candidates. In the elaboration of the proposed methodology, two major parts are that of facial emotion detection and audio emotion classification in development of this system. All the components use the latest deep learning models to deliver an all-encompassing and credible assessment of the emotional state of the candidate. It is developed with Flask in the back and based on HTML, CSS, and JavaScript in terms of user-friendliness.

A. System Overview

Facial Emotion Detection: This module takes in video data, and it is used to detect facial expressions attributed to certain emotions (e.g., happiness, sadness, anger, surprise, etc.).

Audio Emotion Classification: This module is used to classify the emotional tones in the voice of the candidate which include anger, fear, happiness and sadness.

The two modules play a role in the ultimate test by offering knowledge on the emotional condition of the candidate, as well as the level of confidence which he demonstrates in the interview. The system is a combination of the two modalities to provide a more efficient evaluation, because the combination of both modalities makes the evaluation take advantage of the weaknesses of the individual modalities.

B. Facial Emotion Recognition based on YOLO models.

During an interview, facial emotion recognition plays an important role in interpreting the non-verbal factors. YOLO (You Only Look Once) models are used in this project in real-time facial emotion detection. YOLO is an object detection model that is among the fastest and efficient and operates by subdividing the image into a grid and classifying each cell in the grid into a bounding box of the objects in the image. To recognize emotions, the YOLO model detects faces and identifies the locations of important facial features, including mouth, eyebrows, and eyes. Analysis of the sensed facial features is then done to ascertain the mood of the individual

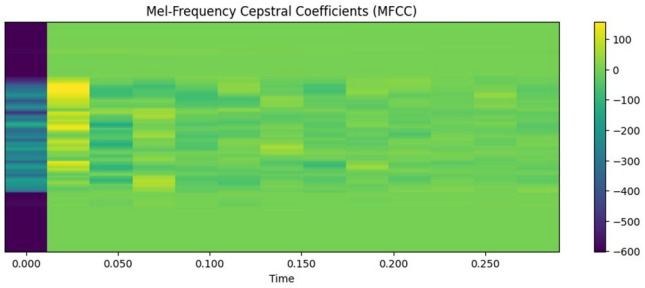
The system makes use of an already trained YOLO model that has been fine-tuned using a specialized collection of facial expressions. Face images with different emotions are labeled and included in the dataset and this allows the YOLO model to classify emotions such as anger, surprise, happiness, sadness and neutral. Through the effectiveness of YOLO, the system is able to handle video frames in real-time, which will give feedback in real-time on the emotion status, when participating in a simulation interview.

Facial emotion recognition could be effectively used on a variety of issues, including occlusions (e.g. glasses or beards), lighting, and even differences between individuals because of YOLO-based facial recognition. Moreover, the model has the capacity to identify more than one face within a single frame, resulting in scalability, in which case, the model can be applicable in group interview scenarios.

C. Emotion Classification in Audio Emotion Hybrid Models.

Simultaneously, the audio emotion classification is carried out on the basis of Convolutional Neural Networks (CNN), Recurrent Neural Networks (RCNN), and hybrid CNNBlock + ConformerBlock. Speech has a considerable amount of emotional material, and it can be examined through the prosodic characteristics of the speech like the pitch, tone and tempo.

The audio information is first processed through the removal of Mel-Frequency Cepstral Coefficients (MFCCs) which are very popular in speech recognition exercises to represent the frequency content of audio signals. These qualities are subsequently inputted into the hybrid model which is aimed at extracting the temporal and spectral contents of the audio signals.



The audio features are learned in CNN layers of the spatial patterns that encode spectral anomalies or voice irregularity that is associated with particular emotions.

The sequential and temporal dependencies among the speech signal are captured with the help of RCNN layers, and the model can recognize the change of emotions in a conversation over time.

CNNBlock + ConformerBlock: The hybrid structure is CVNN and Conformer architectures, that are especially effective when one needs to understand both local feature patterns (CNNs) and long-range temporal dependencies (Conformers) and make the model more effective at classifying notes of subtle emotional cues in speech.

The system is trained with labeled data, which includes emotional speech samples and labels, that is, each audio sample is labeled with a single emotion category, such as anger, happy, sad, and fear. Model output is a prediction of the emotion portrayed in the audio clip with a confidence score which reflects the confidence that the model has in its prediction.

D. Multimodal Fusion of Better Emotion Recognition.

In order to improve the overall output and consistency of emotional recognition system, the audio emotion classification and facial emotion detection outputs are fused. Multimodal fusion is an approach that uses the combination of the predictions of both models using the advantage of complementary ability of facial and audio cues.

The fusion reaction entails two major steps:

1) Feature Level Fusion: The attributes obtained using the facial and the audio data are fused into a single feature. This enables the model to engage both the sight and the audio of the learning process so that the weakness of one of the modalities gets balanced by the other. As an example, in the event that the facial model catches the neutral expressions and the audio model catches a positive tone, fusion mechanism might bias on the positive classification.

2) Decision-Level Fusion: After both the models have given their emotional predictions, then they are merged to a final decision. The predictions can be combined using a weighted voting system or as an ensemble of a neural network, to guarantee that the overall emotion label can be as close to the actual label as possible. The confidence scores of the two models are also summed up to give a general level of confidence when it comes to classifying emotions.Such a multimodal system is much more accurate and robust, which enables a more reliable classification of emotions even in problematic cases.

E. Evaluation and Feedback of Confidence.

A significant attribute of the suggested system will be the consideration of confidence in emotion predictions. The facial emotion detection model as well as the audio emotion classification model will give a confidence score showing how sure the model is of what it is predicting. These confidence scores play an important role in offering real feedback to the candidates.

As an example, in case the system identifies that there is high degree of certainty in the emotional reaction of a candidate, the system will give a high recommendation to work on some of their behaviors, e.g. to relax when a stressful question is raised. On the other hand, the feedback can indicate that more practice or review of specific emotional cues are needed when the score of confidence is low.

The system presents the detailed feedback to the candidate, not only showing what emotions have been identified, but also giving the recommendations on the controlling of emotions and the enhancement of the performance at the interview. This feedback can be seen on the user interface in real-time, where the candidates can make changes when having mock interviews.

F. Web Interface and Web Application Development.

It is a web application with Flask as a backend application. Flask is a lightweight web framework that allows developing the application in an efficient way and being scaled. The server takes the uploaded audio and video files then executes the emotion detectors and provides an evaluation.

The system interface is developed in HTML, CSS and JavaScript and the frontend makes the user interface more interactive and easier to use. Applicants have the opportunity to provide their mock interview recordings and emotional response analysis, as well as get feedback on their performance in real-time. The interface is supposed to be easy but informative, and users will have the means to monitor and enhance their emotional reactions and the level of confidence.

V. RESULTS & DISCUSSIONS

A. Results for Audio Classification

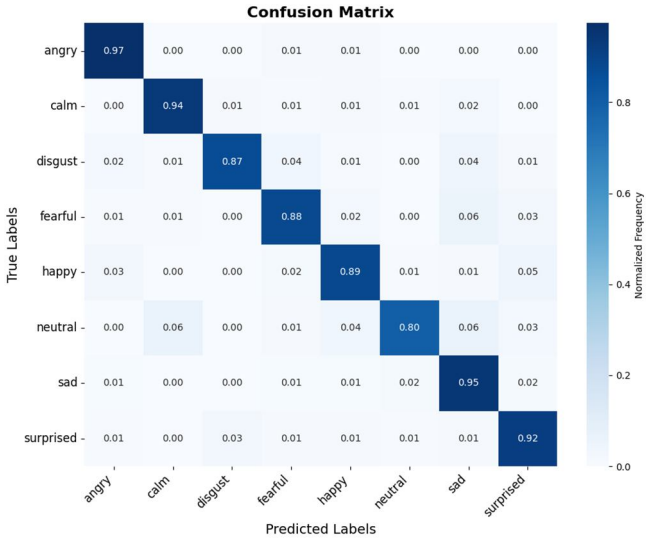


Figure 2 HYRBID MODEL CNN+CONFORMER

This confusion table indicates the results of an emotion classification model on nine categories. The majority of the emotions have high prediction accuracies of above 87 per cent and that of angry is 97 per cent. Misclassifications are also low and this implies high model reliability.

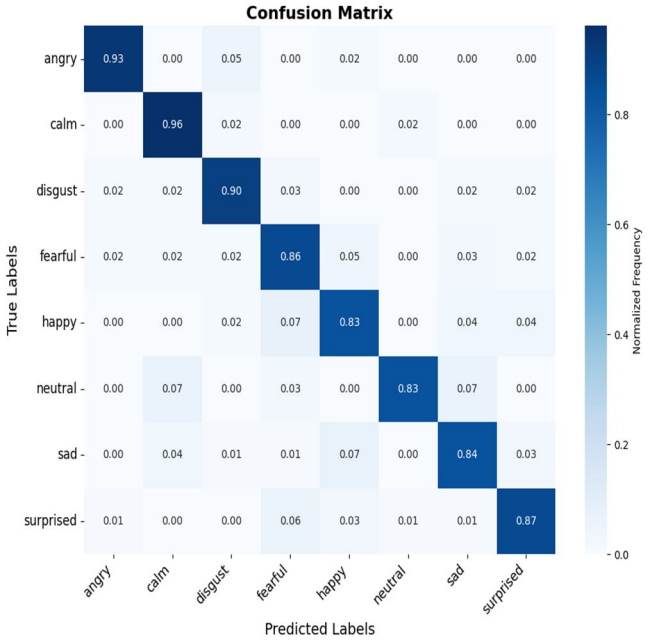


Figure 3 RCNN MODEL

This confusion matrix indicates the model level of accuracy in predicting eight emotions. Majority of the categories such as calm (96%), disgust (90%), etc are categorized correctly and others such as happy, neutral and sad are around 83 percent. There are very low misclassifications that are well evenly distributed.

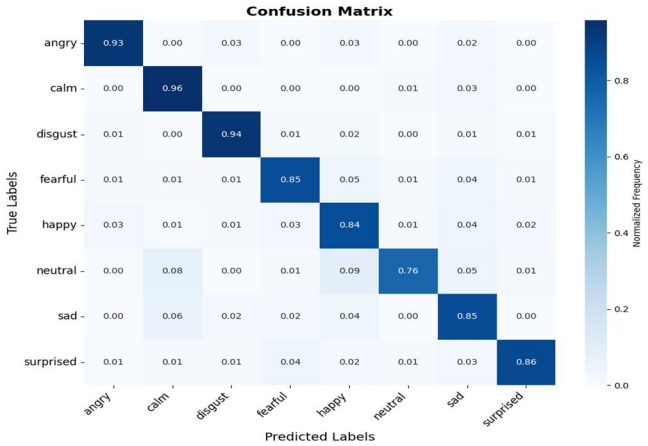


Figure 4 CNN MODEL

The performance of this confusion matrix was very high in the classification of emotions as calm (96%), disgust (94%), and angry (93%) were the most accurate predicted emotions. Other emotions are equally misclassified showing balanced model behavior.

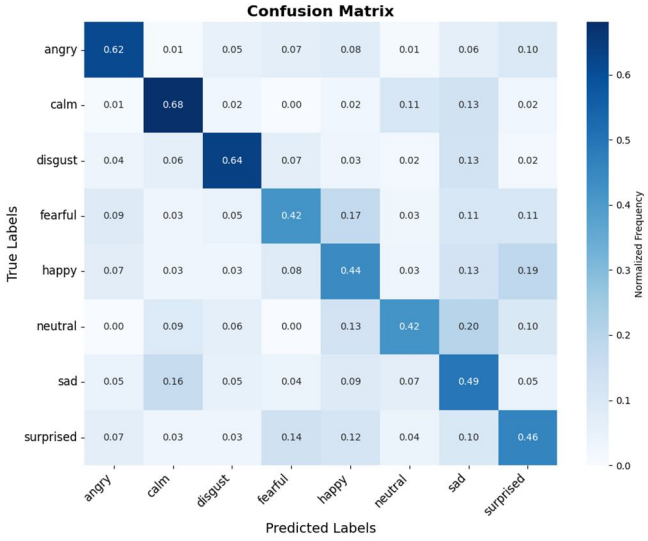


Figure 5 LSTM MODEL

This confusion table demonstrates mediocre classification performance, and the most accurately predicted categories are calm (68) and disgust (64) categories. Such emotions as fearful, neutral, and surprised are very confusing, particularly, with the confusion of predictions between sad and happy.Facial emotion classification results.

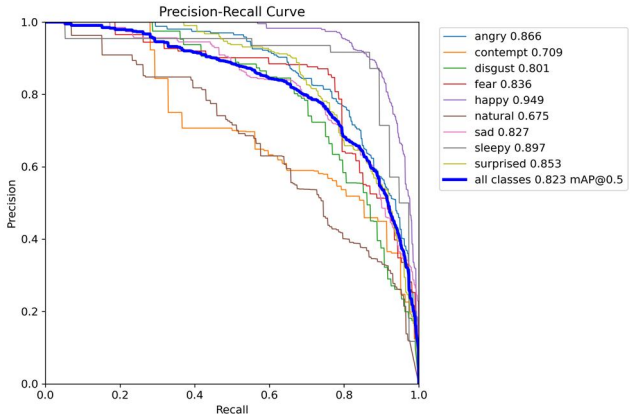


Figure 6 YOLOV9 MODEL

This multi-class performance curve has a high level of precision-recall, and the overall mAP of 0.5 is 0.823. The most accurate predictors are emotions such as sleepy (0.891), angry (0.866), and surprised (0.853), and the same cannot be said about the natural and contempt.

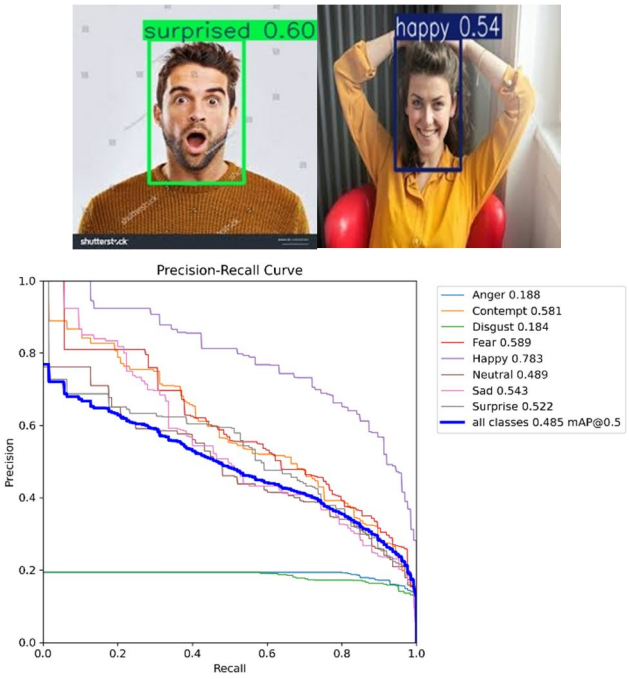


Figure 7 YOLOV9 MODEL

The graph indicates the level of identified different emotions by a model in terms of precision and recall.The colored lines depict the emotion classes and Happy is best (0.783) and Disgust is worst (0.184).The thick blue line illustrates the performance of the overall use of all emotions, and the average precision (mAP) is 0.485 at IoU 0.5.The higher curves imply higher balance in precision-recall of that emotion.

VI. CONCLUSION

In the current paper, the design of an AI-based system to be used in judging emotional feedback and confidence during fake interviews was discussed. The suggested system will combine two essential aspects: facial emotion recognition and audio emotion classification with the help of the latest deep learning models to give a complex evaluation of the emotional conditions of a candidate during an interview. Using the models of YOLO-based facial emotion recognition and audio emotion classification based on the Hybrid CNN + ConformerBlock, the system can analyze visual and auditory signals as such with high accuracy.

The combination of these two modalities greatly improves the predictability of emotions and it can be said that the emotional intelligence of a candidate is rated more efficiently. The confidence-sensitive feedback system also becomes an even stronger mechanism since it can offer real-time, practical feedback to the candidates to improve on their emotional control and interview performance. This multimodal system is helpful because it is able to record subtle facial expressions and tones in speaking and provides a full picture of how the candidate behaves during the interview.

This study proves that AI can transform the conventional recruitment process since it is no longer based on verbal reactions but also on emotional intelligence. The combination of the real-time assessment of emotional responses in the system with the detailed feedback offered can become useful not only to the interviewees but also the interviewers to improve the preparation in the interview and self-awareness. Further development of the models, extending the emotion categories, and implementing the system to the variety of real-life situations might be considered as a direction of the future work to further confirm the validity of the system.

VII. FUTURE SCOPE

Although such an AI-powered tool of mock interview evaluation shows encouraging outcomes, it can be improved and developed in a number of different ways. The following improvements can be considered in the future:

Increased Categories of Emotions: The existing system is able to identify a few emotions. Further effort might be done on expanding the categories of emotions to cover more complex emotions, like confusion, frustration, or excitement, to offer more in-depth emotional analysis in interviews.

Cross-Cultural Validation: Is the expression of emotions in some cultures different? It may also be possible to increase the system to capture cross-cultural differences in facial expressions and speech to enhance the robustness of the model and make it more applicable.

Increased Accuracy using More Data: The system already works well on the existing dataset although with a bigger and more diverse dataset one might also increase the accuracy and generalizability of the models that are being trained.

Combination with Virtual Interview Platforms: To provide the system with more accessibility, it can be combined with virtual interview platforms that are used by companies. This would enable real-time emotion and confidence analysis in real job interview situations.

Better Performance through Hybrid Models: The better performance of the system could be achieved through exploring more hybrid models, including deep learning with the traditional machine learning methods, which would be more effective in addressing the more difficult emotional situations.

REFERENCES

[1] F. Ullah, S. M. Sarwar, and A. Xiong, “Optimizing Real-Time Emotion Recognition: A YOLO v.8 Deep Learning Solution for Facial Expression Analysis,” Proceedings of the IEEE International Conference on Computer and Communications, ICCC, no. 2024, pp. 150–156, 2024, doi: 10.1109/ICCC62609.2024.10942205.

[2] R. G. Praveen, E. Granger, and P. Cardinal, “Recursive Joint Attention for Audio-Visual Fusion in Regression based Emotion Recognition,” Apr. 2023, Accessed: Nov. 15, 2025. [Online]. Available: https://arxiv.org/pdf/2304.07958

[3] S. Kour, P. Sharma, A. M. Zargar, A. Sonania, T. Hassan, and Nijamuddin, “Emotion Recognition from Speech Signals Using Hybrid CNN Model,” Proceedings - 3rd International Conference on Advancement in Computation and Computer Technologies, InCACCT 2025, pp. 666–670, 2025, doi: 10.1109/INCACCT65424.2025.11011474.

[4] H. Jin, T. Yang, L. Yan, C. Wang, and X. Song, “Multimodal Emotion Recognition in Conversations Using Transformer and Graph Neural Networks,” Applied Sciences 2025, Vol. 15, Page 11971, vol. 15, no. 22, p. 11971, Nov. 2025, doi: 10.3390/APP152211971.

[5] Z. Cheng et al., “Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning”.

[6] Z. Lin, F. Cruz, and E. B. Sandoval, “Self context-aware emotion perception on human-robot interaction,” Australasian Conference on Robotics and Automation, ACRA, Jan. 2024, Accessed: Nov. 15, 2025. [Online]. Available: https://arxiv.org/pdf/2401.10946

[7] J. Salas-Cáceres, J. Lorenzo-Navarro, D. Freire-Obregón, and M. Castrillón-Santana, “Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics,” Multimedia Tools and Applications 2024 84:23, vol. 84, no. 23, pp. 27327–27343, Sep. 2024, doi: 10.1007/S11042-024-20227-6.

[8] R. Rani and M. K. Ramaiya, “Enhancing Speech Emotion Recognition with Multi-Modal Hybrid Features and CNN,” International Journal of Electronics and Communication Engineering, vol. Volume 12, no. 7, pp. 35–46, Jul. 2025, doi: 10.14445/23488549/IJECE-V12I7P104.

[9] T. Thebaud et al., “Multimodal Emotion Recognition Harnessing the Complementarity of Speech, Language, and Vision,” ACM International Conference Proceeding Series, pp. 684–689, Nov. 2024, doi: 10.1145/3678957.3689332;PAGE:STRING:ARTICLE/CHAPTER.

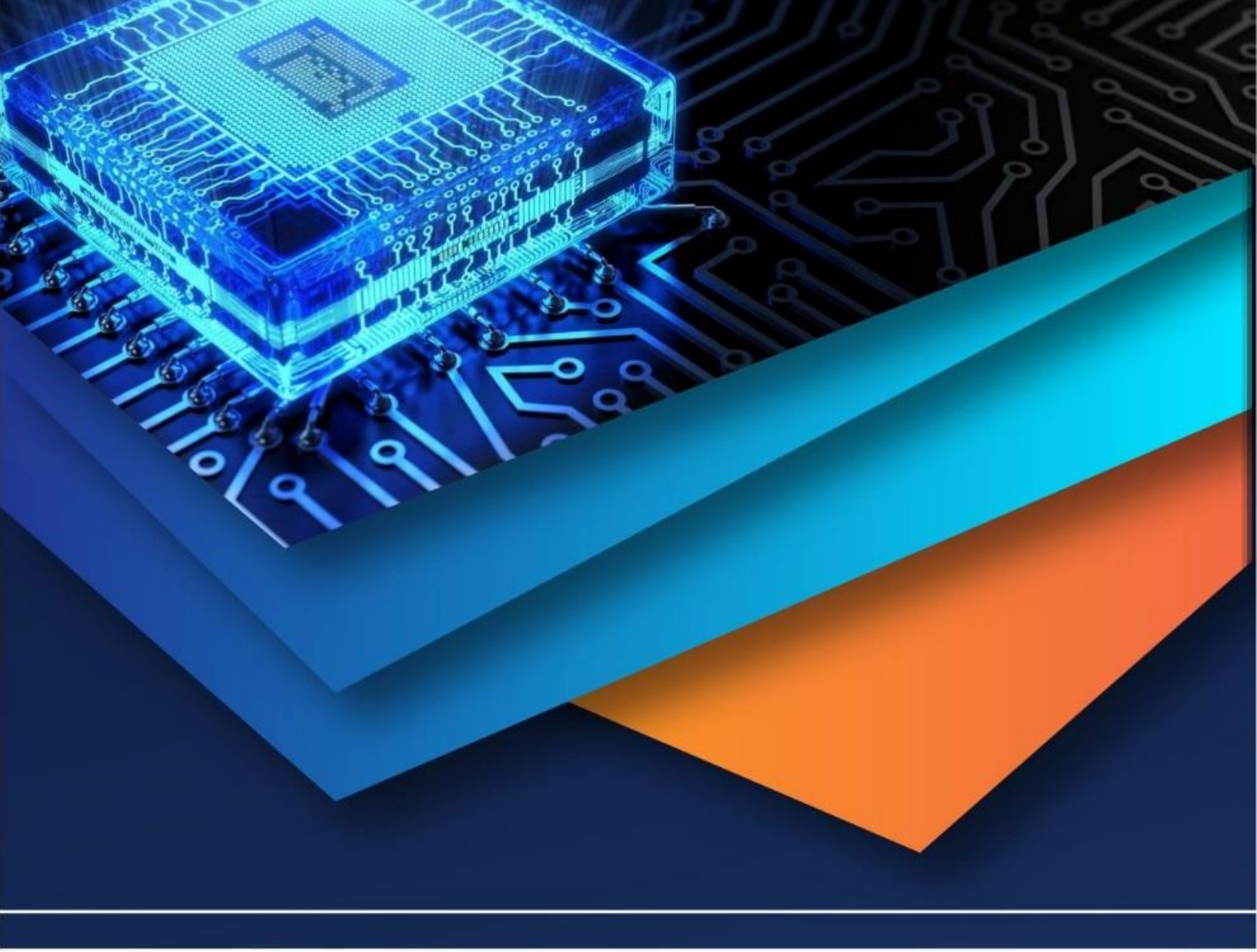
[10] Z. Ma et al., “emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation,” Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 15747–15760, Dec. 2023, doi: 10.18653/v1/2024.findings-acl.931.

[11] S. Latif, A. Shahid, and J. Qadir, “Generative Emotional AI for Speech Emotion Recognition: The Case for Synthetic Emotional Speech Augmentation,” Applied Acoustics, vol. 210, Jan. 2023, doi: 10.1016/j.apacoust.2023.109425.

[12] R. Gnana, E. Granger, and P. Cardinal, “RECURSIVE JOINT ATTENTION FOR AUDIO-VISUAL FUSION IN REGRESSION BASED EMOTION RECOGNITION”, Accessed: Nov. 15, 2025. [Online]. Available: https://github.com/

[13] Z. Lin, F. Cruz, and E. B. Sandoval, “Self context-aware emotion perception on human-robot interaction,” 2023.

[14] J. Salas-Cáceres, J. Lorenzo-Navarro, D. Freire-Obregón, and M. Castrillón-Santana, “Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics,” Multimedia Tools and Applications 2024 84:23, vol. 84, no. 23, pp. 27327–27343, Sep. 2024, doi: 10.1007/S11042-024-20227-6.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY