Signals and Syntax: Deep Learning for Emotional Intelligence and Text-Based Linguistic Proficiency

[1]A N Jyothsna and [2]Pamela Vinitha Eric
[1,2]Presidency School of Computer Science and Engineering, Presidency University, Bengaluru, Karnataka, India.
[1]forjyothsna@gmail.com, [2]pamelavinitha.eric@presidencyuniversity.in

Correspondence should be addressed to Pamela Vinitha Eric: pamelavinitha.eric@presidencyuniversity.in

Article Info
Journal of Machine and Computing (https://anapub.co.ke/journals/jmc/jmc.html)
Doi : https://doi.org/10.53759/7669/jmc202505214
Received 12 July 2025; Revised from 27 August 2025; Accepted 19 September 2025.
Available online 05 October 2025.

Abstract— In the real-world communication scenario, understanding Emotional Intelligence (EI) of humans plays a pivotal role and this strengths to a greater extent when the modalities are combined i.e video & audio. This paper investigates the human emotions which are crucial in communication using Video, Audio and Video & Audio (V-A). A unique dataset of impulsive Emotional Intelligence videos was collected. Action Units (AU's) were extracted from the videos and from audios – spectral descriptors, MFCCs and prosodic features. To train the model we propose a 1D-CNN (One-Dimensional-Convolutional Neural Network). Multimodal late-fusion signals were trained on 1D-CNN architecture which achieved significantly 83.33% accuracy. This demonstrates that by integrating multimodal cues predictions are stable and diminish weakness of unimodal. This study emphasis on efficient multimodal design of EI recognition frameworks for assessment of real-world communication. We also experimented on textual domain. From the audios we extracted the text and the text has been fine-tuned for multiclass classification with BERT-embeddings. We trained the features with Multilayer Perceptron (MLP) where the model has achieved 91% accuracy on the test set.

Keywords— Emotion Intelligence, Action Units, Deep Learning, MLP, BERT.

I. INTRODUCTION

In human communications Emotional Intelligence emerges as a critical qualification which brings effectiveness in interpersonal, social interaction and in decision making. The traditional ways of examining the EI depends majorly on self-ratings, questionnaires or by expert evaluations. These approaches may fail when EI assessment is carried out by compromising on the dynamic cues that are expressed. With the Machine Learning (ML) emerging now-a-days and with its data-driven and automated approaches, these visual cues can be captured in a valid manner. The systems like multimodal – that combinedly takes vocal cues, facial movements and then the linguistic content offers a rich platform for EI assessment. The advances in Deep Learning (DL) with convolutional architectures have proved successful that captures facial action units (AU's) and speech features. In the frameworks which involves spontaneous emotional expressions from actual speakers, these architectures are under explored.

Linguistic Proficiency (LP), is the ability to use a particular language precisely. While humans interact, the language plays a vital role to make the listener to understand better when the speaker is communicating. This minimizes the ambiguity, can express complex thoughts with ease and also ensures accurate interchange of the context or ideas. The use of LP is beyond the simple grammar, where it makes the professionals confident and clearly, they can articulate their ideas and can even navigate to diverse cultural cues. Analyzing linguistic proficiency is critical in screening for interviews. Combining LP with Emotional Intelligence as performed in this study, provides sophisticated systems, beside measuring what is said, it also measures how confidently (text is also from confidence videos in this study) the message has been delivered. [1] defines LP as the potential to use the language precisely and correctly in various settings. [2] their work distinguishes linguistic competence and communicative competence.

The paper is organized as follows: Section II is about literature review. Section III tells about model design and has detailing about collection of data and methodology. Section IV is about pre-processing of the data and feature extraction has been discussed. Results, resulted from the implementation is provided in the Section V. Conclusion is in Section VI.

II. LITERATURE REVIEW

'Emotional Intelligence' was developed originally in 1970's and 80's by three psychologists namely John Mayer, Howard Gardner and Peter Salovey in their writings and work. Emotional Intelligence is highly suitable for individual development and growth of the organization as its main principles gives a way to assess and understand management styles, behavior of people, interpersonal skills, attitudes, in HR planning, recruitment and selection, job profiling, customer relations and service, management development and much more.

In conversations which happen naturally, conveys more than a meaning of words spoken. This speech may also communicate speaker's emotional state and their personality [3,4]. The speech signal properties which are called as paralanguage which is used consciously or unconsciously communicates emotion. These features cover the utterance contents and connotation change thus include speech rate, pitch, loudness, and voice quality and include extra vocal behaviors like laughter, sighs, and gasps. Hence paralinguistic speech properties play a vital role in human conversation which affect a listener's utterance perceives. [5], work speaks about the trending use of popular evolving features of MFCCs. [6] used features which are hand-crafted and were fed into deep neural networks. To hierarchical DNN classifier voice-quality, prosody and spectral features were used to train the model and achieved accuracy of 81.2 % on RAVDESS dataset. [6] proved that, vocal information is only one of the contributors to predict emotions. The emotion prediction result improves if other facial features were incorporated. And they concluded that by including textual features, the results were enhanced for the speech emotion recognizer.

In the communication process, individual thoughts and emotions negatively or positively are contemplated. With the high level of communication skills, the emotion of the individuals can be handled successfully. [7] demonstrated that the emotional intelligence is positively correlated with communication skills. [8] identified an important relationship between communication and emotional intelligence which led to academic success. They derived that the use of positive emotions has a valuable effect on communication skills. This work proves the relationship between effective communication and emotional intelligence, and this emotional intelligence in turn leads to effective communication skills.

Initially acoustic facial features were used to identify behavioral features like stress and emotions etc. [9] in this research they have utilized the face points like cheeks, eyebrows, chin and forehead to detect the emotion and using Gaussian Mixture Model (GMM) they trained the model. [10] extracted 68 key points in human faces for emotion detection and these features are trained on Support Vector Machine, Decision Tree and Random Forest. Rao and [11] have considered mouth and eye features for detecting the emotion using Artificial Neural Network (ANN). [12] used hand and head movements with facial features to detect an emotion.

From various areas, emotion recognition leverages techniques. This process utilizes different input values like speech [13], facial expressions [14]. However, facial expressions are studied extensively. Majority of the literature on emotion detection focused mainly on face which is the main source, at the price of various other modalities that are available. Group of emotions such as disgust, anger, fear, neutral, sadness, happy and surprise according to Paul Ekman are cross-culturally and universally communication with the expressions in face [22].

The most important evolution in the landmarks context is Action Units (AUs). [15], AUs were extracted and fed these features into an auto-encoder of seven-layer, and achieved high rate of accuracy in emotion recognizing. All the above are the trending tools which are used to extract facial morphology and for expression detection among video frames. With the evolution of Convolutional Neural Network (CNN) models, recognition of emotion results has improved drastically. [16] has shown that the CNNs which were trained for the purpose of emotion recognition has learnt the features which strongly align with Facial Action Units which is proposed by Ekman [22].

Effective communication is a collaborative process that exists between Linguistic Proficiency (LP) and Emotional Intelligence (EI). While LP gives foundation for expression, EI commands the deliberate application of language in social contexts navigation. [17] propose, competence in communication requires basic knowledge of sociolinguistic variations that are emotionally inherent. Further, the categorization ability into positive and negative emotions – a main component of daily emotion by [18] and text-based EI rely on linguistic proficiency as described by [19]. The synergy between EI and LP is well organized in [20] which showed that the verbal ability will be the core part of emotion understanding. In this paper, their meta-analysis validates that for the effective communication the integration of EI strategies with LP tools, allows the individuals to propagate both positive emotions and negative emotions [18] with precision.

In this study the emotions are classified into positive emotions(like-happy/joy) and negative emotions (like-anger/fear/sad). This classification is in-line with Ekman's basic emotion model [22] and also [21] affect framework of positive affects (happy/joy) and negative affects (anger/fear/sad). As mentioned by [18], this binary classification is a crucial step for understanding individual's emotions which can be seen in their daily life.

III. MODEL DESIGN

Data Collection

For evaluating the emotional intelligence of speakers, data collection was done with a designed protocol, as there is no data available publicly for this task. Consent was taken from all the participants and video recording procedure was explained. Recordings were conducted in a room to avoid any disturbances. Laptop with front-camera was placed in-front of the participants and instructions were given to respond to set of questions and to keep the recording active until all the responses were completed. Participants were encouraged to take pause and answer in the case of uncertainty. No feedback was given to the participants after the recordings. For emotions to express, the questions were divided into two categories i.e positive emotion and negative emotion. As the positive emotion are directly linked to happy, joy [23], the questions were also framed accordingly so that the participants can express the emotions positively. Same thing is applied for the negative emotions [22] as well. To name a few – "No exams and you are promoted to next semester, what is your reaction" a positive emotion question and – "You are home alone at night, and suddenly you hear unfamiliar noises, how do you feel" a negative emotion question. Five positive questions and five negative emotion questions were framed. 138 videos were collected representing various age groups. To maintain consistency all the participants were asked the same set of questions. Questions were framed in such a way that the participants can express positive emotions like happy/joy with the questions like "*No exams and you are promoted to next semester, what is your reaction*". These types of questions helped the participants to exhibit the emotions clearly. Negative emotion questions like "*You are home alone at night, and suddenly you hear unfamiliar noises, how do you feel* "these kinds of questions allowed the participants to show the negative emotion like fear etc.

IV. METHODOLOGY

The proposed architecture (shown down) is a novel specialization of the single dimensional CNN, which is customized to process static 48 vector dimension of the fused features of multimodality. The major goal is in the innovation which lies in the application of unconventional Conv1D with kernel size as 3 to the features which allows the network to do cross-feature interaction discovery which means learning the relations that are predictive between local groups of adjacent features instead of across time. This is achieved through 2 blocks of convolution that is stabilized by BatchNorm1d after each Conv1D layer. Most crucially, this architecture includes a dual-stage, aggressive regularization technique using high Drop-out rates (0.4) in both fully connected and convolutional stages.

The CNN1D architecture is designed exclusively to maximize the generalization across the multimodal feature vector space. Our design incorporated Batch normalization immediately followed by each convolution. This step is critical for balancing feature vector distributions and to maintain consistency scale over all the convolution layers, importantly when fused features of contrasting origins (Multimodal).

Meta Data

Bagheri et al. [22] used Action Units (AU's) for emotion recognition. In this research participants from various age groups were taken for the data collection. Gender balance has been taken care. The specific purpose of the recordings was explained to the participants. All the participants were having normal vision and hearing problems. Even though few participants were wearing glasses that did not hinder the recordings. The recordings were taken in a closed room and structured questions were asked to each individual participant. The video duration on an average range from 6 seconds to 25 seconds.

We structured our analysis by classifying the questions into positive and negative emotions. Positive emotions like happy, joy and negative emotions like sad, anger, surprise were [22]. Action Units (AU's) from face as been considered as visual cues. When the positive questions were asked, speakers were able to show their emotion (happy/joy) while answering which helped us to capture the visual cues of the participants. When negative emotion questions were asked, speakers could express their negative emotions such as sad, fear, anger etc. which were captured as visual cues for negative emotions. After the data collection, all the videos were labelled accordingly and a random check has been carried out to ensure the labelling accuracy.

Randomly the videos were evaluated on the full dataset to check whether they are correctly grouped into positive emotion and negative emotion. 4 positive emotion questions and 4 negative emotion questions were asked to each participant. In the preprocessing procedure, few videos were removed as the speakers were not responding according to the questions that were asked to them. The set-up was made in such a way that it resembles a real-world scenario.

V. DATA PRE-PROCESSING

Feature Extraction

Manually the data was analyzed to check the labels as positive emotion and negative emotion. This step is carried out after the data collection is finished. Action Units (AU's) were extracted from all the videos. The research on facial expression recognition, AU's which are defined by the Facial Action Coding System (FACS), provides a very important framework which qualifies the facial movements linked to emotions [22]. These AU's gives various facial recognition units for happy, sad, fear etc. This precise way of measuring of facial movements facilitates the most difficult expression identification like sad, happy or anger. With the help of AU's emotions have been classified which makes it possible to interpret various human behaviors from visual data with good result.

The dataset extracted from the video tracks of emotional intelligence, has a high-dimensional feature matrix, which comprises of a VideoName linked to a related emotion label. This set has a various range of mid-level spectral features and low-level descriptors. Mel-Frequency Cepstral Coefficients (MFCCs) the spectral features, Spectral Bandwidth, Spectral Centroid, Spectral Rolloff, Tonnetz and chroma features. All these compact multiple feature combination helps in selective representation of speech acoustics, which in turn are optimized for model training and to target the fine-tuned emotion recognition.

The train dataset comprises 138 samples categorized into positive emotion and negative emotion. The test samples are 30 and again divided into positive and negative emotions. **Fig 1** shows Proposed Architecture CNN1D.

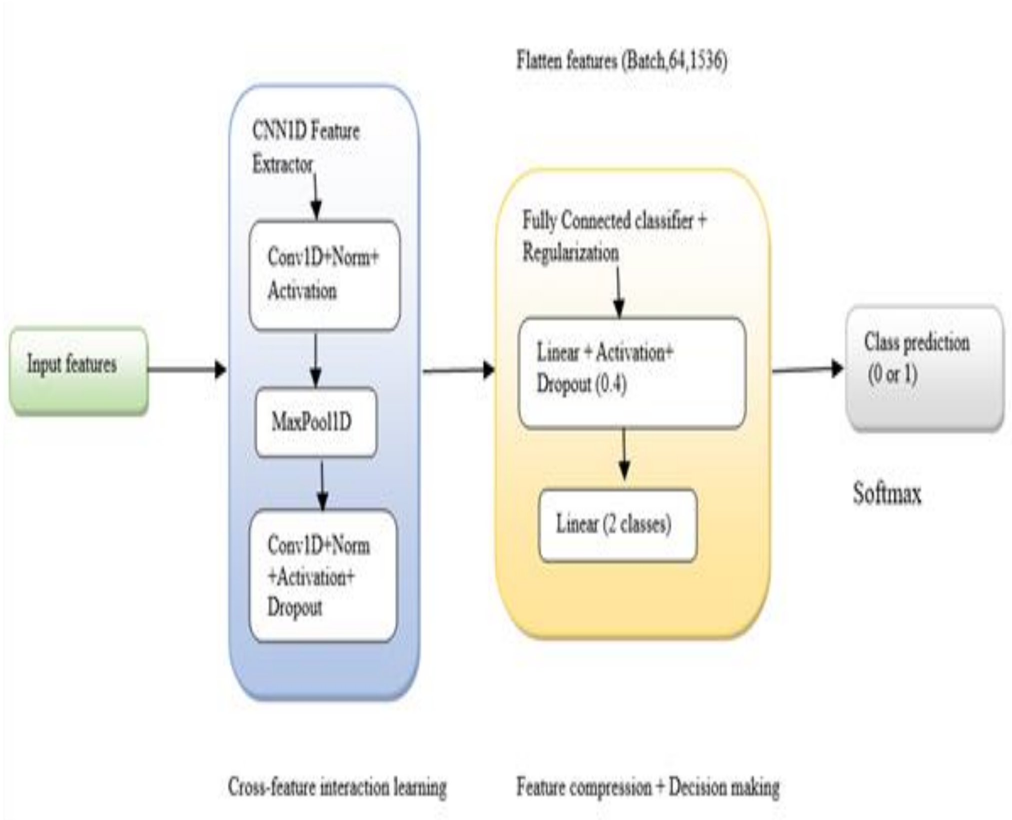

Fig 1. Proposed Architecture CNN1D.

VI. MODEL BUILDING

The train dataset comprises 138 samples categorized into positive emotion and negative emotion. The test samples are 30 and again divided into positive and negative emotions.

Video Data Analysis

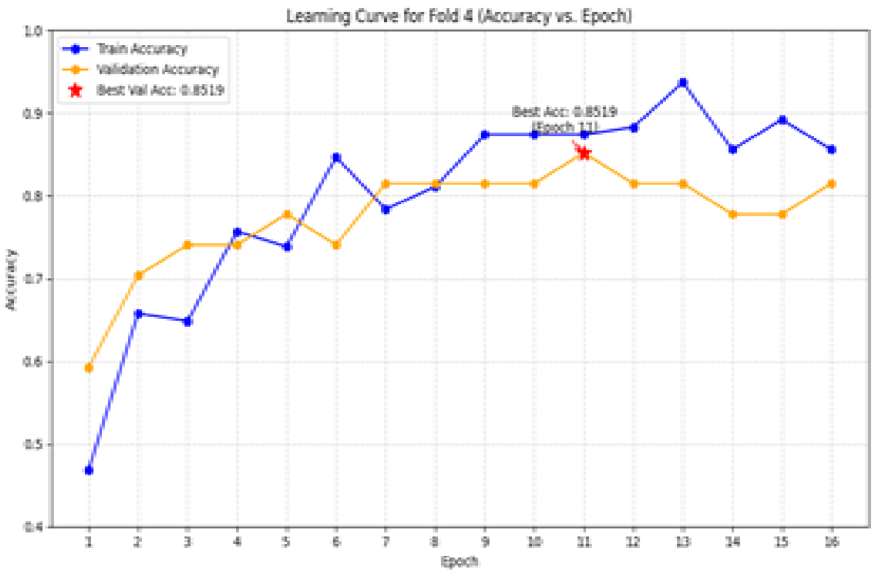

Fig 2. The Graph Indicates Fold 4 Train & Validation Accuracy.

Table 1. The Various Accuracies on Different Folds

Fold	Best Validation Accuracy	Best Training Accuracy	Early Stop Epoch
1	0.6071	0.7091	8
2	0.7143	0.8273	12
3	0.8214	0.7273	11
4	**0.8519**	**0.8739**	**11**
5	0.6667	0.5315	6
Average	**0.7323**	**0.7338**	**9.6**

The model overall performance is ensemble test accuracy of 83.333 and F1-score is 83.331 across both the classes, which demonstrates unbiased and robust capabilities of classification on the test unseen data. More importantly, with K-fold cross validation the model potentially achieved the best accuracy of 85 in fold 4, which is the success of early stopping mechanism that selected optimal weights are the 11th epoch. **Fig 2** shows The Graph Indicates Fold 4 Train & Validation Accuracy. **Table 1** shows The Various Accuracies on Different Folds.

Audio Data Analysis
Audio has been extracted from the above videos for audio analysis. Preprocessing of the audio files has been carried out where all the audio files were denoised.

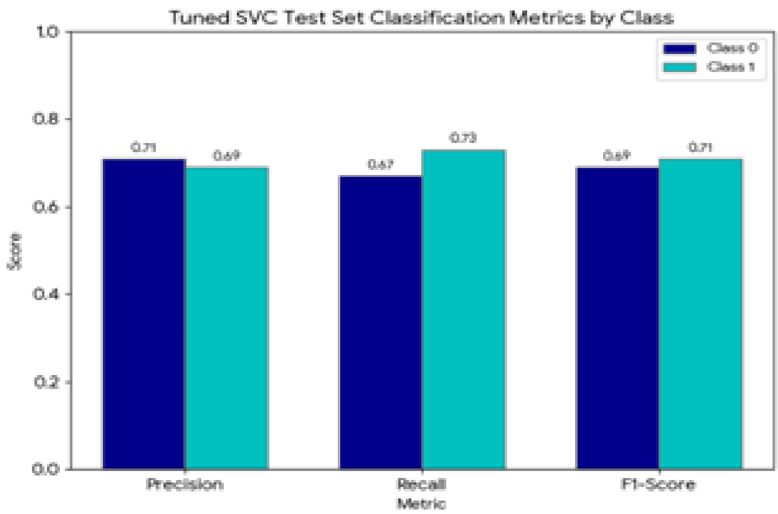

Fig 3. Various Metrics By Class.

For the audio model training, we have utilized SVC (Support Vector Classifier), which yielded good result, especially the contrast low performing Deep Learning approach. The model achieved the best cross-validation score of 58%. This transition was required as 1D+CNN architecture which was used for video only analysis was resulting in very low performance giving only 53% accuracy on test set, which strongly shows that for this kind of dataset and feature set, DL is detrimental. This limitation was overcome by SVC giving a final test accuracy of 70%. The overall test accuracy score is 70%, which has balanced F1-score with 69% and 71%. (shown in the above graph). **Fig 3** shows Various Metrics By Class.

Video & Audio Analysis
The multimodality approach resulted in an outstanding performance on the independent test dataset with a robust Ensemble test accuracy of 83%. With the 5-fold cross-validation, the model performance was moderate but the final ensemble method leveraged successfully the robustness across these folds to get a good score of prediction, which proved the effectiveness of the multimodal fusion strategy.

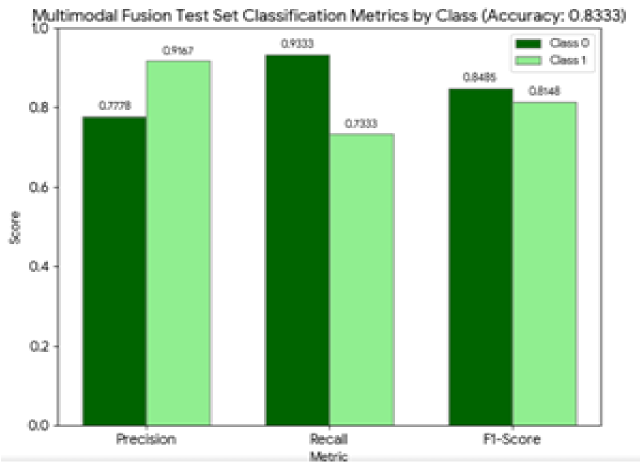

Fig 4. Various Metrics.

The graph shows the precision for class 1 which is 0.91% with best recall score for class 0 with 0.933. **Fig 4** shows Various Metrics.

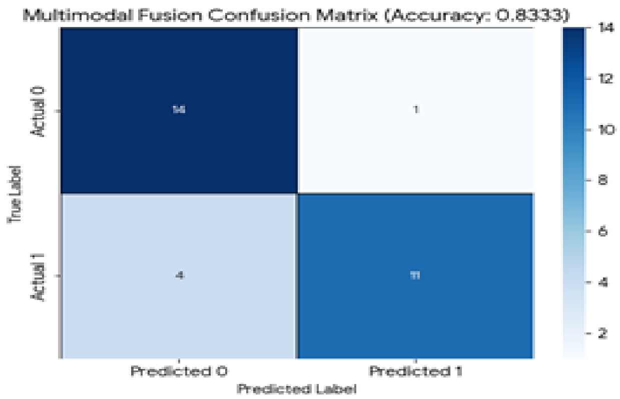

Fig 5. Confusion Matrix.

The above confusion matrix shows that the overall accuracy 83%. The matrix shows that out of 30 samples, 25 were correctly predicted and a total 5 were False Negatives. **Fig 5** shows Confusion Matrix.

Linguistic Proficiency

The transcribed text has been extracted from the confidence and emotional intelligence videos. Total 229 rows of transcribed text has been extracted. Same thing is used for training the model.

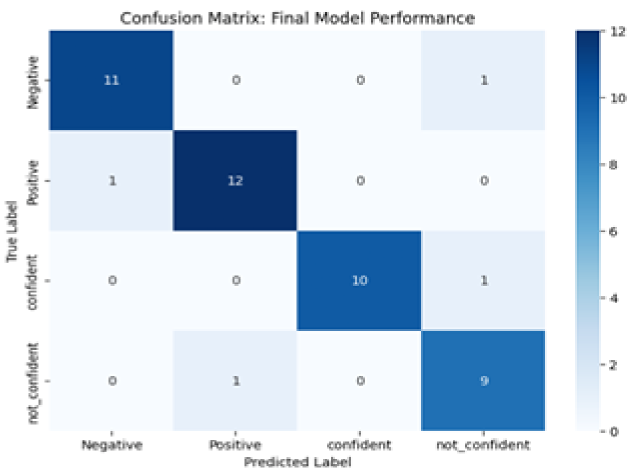

Fig 6. Multi-class Confusion Matrix.

The above matrix is a visual representation of the model which displays true against predicted. The matrix shows clearly the accuracy of 91.3%. This conveys that the model has diagnosed the 42 as correct predictions and only 4 errors out of 46 rows of test set. And showed balanced and robust performance on the challenging categories Positive and Negative. **Fig 6** shows Multi-class Confusion Matrix.

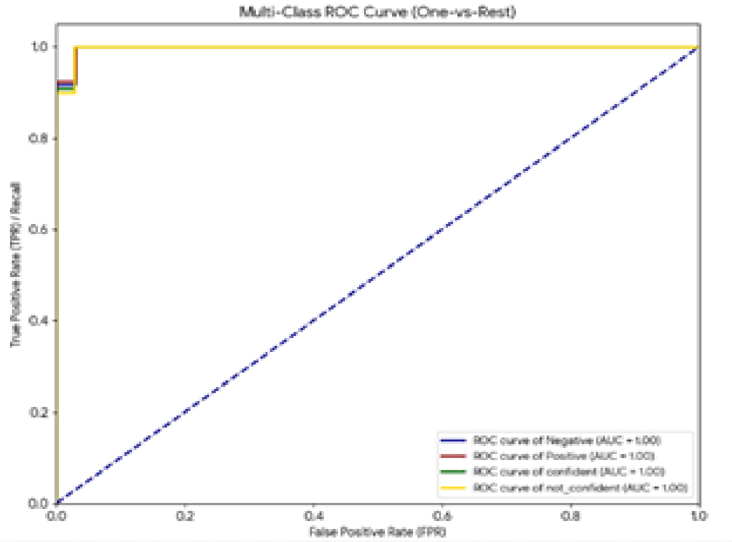

Fig 7. ROC-AUC Curve.

The AUC of 99.8% depicts the BERT fine-tuned model have exceptionally high discriminatory strength. It constantly ranked true positive more than true negatives. This draws an insight that the features that were extracted are separable highly and robust for these kind of multi-class classification problems. **Fig 7** shows ROC-AUC Curve.

VII. CONCLUSION

This research demonstrated successfully the potentiality of multimodal approach for the Emotional Intelligence classification task, concentrating mainly on the classification of positive and negative emotions. With video and audio data integrating through CNN1D architecture (proposed), the system could capture the facial spatial-temporal cues and temporal cues of vocal modality. The multimodal approach using our proposed CNN1D architecture obtained a robust accuracy of 83.33% for the kind of dataset which we used to train the model. The findings also suggest, by integrating both visual cues with audio feature have provided a holistic and robust representation of speaker's Emotional Intelligence.

The proposed architecture has achieved 83.33% with multimodal (Video & Audio) and unimodal (Video only) configurations highlights the stability of the architecture. We have also successfully demonstrated classification framework to identify the speaker's states by combining the Confidence cues and Emotional Intelligence cues via linguistic approach. The fine-tuned BERT architecture which used the transcribed text from the audio files, illustrated that the deep contextual embeddings were able to capture the precise nuances of LP which are required for this kind of multi-class classification. The future work will be focused to optimize fusion layer to prioritize the specific modalities which is based on noise levels of the environment which ensures the system remains strong for the real-world communication scenarios.

CRediT Author Statement

The authors confirm contribution to the paper as follows:
Conceptualization: A N Jyothsna and Pamela Vinitha Eric; **Methodology:** Pamela Vinitha Eric; **Data Curation:** Pamela Vinitha Eric; **Writing- Original Draft Preparation:** A N Jyothsna and Pamela Vinitha Eric; **Investigation:** A N Jyothsna and Pamela Vinitha Eric; **Supervision:** Pamela Vinitha Eric; **Writing- Reviewing and Editing:** A N Jyothsna and Pamela Vinitha Eric; All authors reviewed the results and approved the final version of the manuscript.

Data Availability

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Conflicts of Interests

The authors declare no conflict of interest

Funding

No funding agency is associated with this research.

Competing Interests
There are no competing interests.

Consent to Publish
All the authors gave permission to Consent to publish.

REFERENCES

[1]. Cloud, N., Genesee, F., & Hamayan, E. (2000). Teaching English language learners: A focus on content and language. Heinle & Heinle.
[2]. M. CANALE and M. SWAIN, “THEORETICAL BASES OF COMMUNICATIVE APPROACHES TO SECOND LANGUAGE TEACHING AND TESTING,” Applied Linguistics, vol. I, no. 1, pp. 1–47, Jan. 1980, doi: 10.1093/applin/i.1.1.
[3]. C.-H. Wu, J.-C. Lin, and W.-L. Wei, “Survey on audiovisual emotion recognition: databases, features, and data fusion strategies,” APSIPA Transactions on Signal and Information Processing, vol. 3, no. 1, 2014, doi: 10.1017/atsip.2014.11.
[4]. S. Chanda, K. Fitwe, G. Deshpande, B. W. Schuller, and S. Patel, “A Deep Audiovisual Approach for Human Confidence Classification,” Frontiers in Computer Science, vol. 3, Oct. 2021, doi: 10.3389/fcomp.2021.674533.
[5]. Bhavan, A., Chauhan, P., & Shah, R. R. (2019). Recognizing emotions from speech using MFCC and deep learning. 2019 IEEE Third International Conference on Multimedia Computing, Networking and Applications (MCNA), 7–12. https://doi.org/10.1109/MCNA.2019.8938161
[6]. Y. B. Singh and S. Goel, “An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning,” Multimedia Tools and Applications, vol. 80, no. 9, pp. 14001–14018, Jan. 2021, doi: 10.1007/s11042-020-10399-2.
[7]. Yousefi, F. (2006). The relationship between emotional intelligence and communication skills in university students. Journal of Iranian Psychologists, 3(9), 5-13. https://www.sid.ir/en/journal/ViewPaper.aspx?id=10312
[8]. Taşlıyan, M., Hırlak, B., & Harbalıoğlu, M. (2017). The relationship between emotional intelligence, communicatıon skills and academic achievement: An application on university students. Assam Uluslararası Hakemli Dergi, 2(3), 45-58. https://dergipark.org.tr/tr/pub/assam/issue/32282/358459
[9]. A. Metallinou, S. Lee, and S. Narayanan, “Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice,” 2008 Tenth IEEE International Symposium on Multimedia, pp. 250–257, Dec. 2008, doi: 10.1109/ism.2008.40.
[10]. B. T. Nguyen, M. H. Trinh, T. V. Phan, and H. D. Nguyen, “An efficient real-time emotion detection using camera and facial landmarks,” 2017 Seventh International Conference on Information Science and Technology (ICIST), pp. 251–255, Apr. 2017, doi: 10.1109/icist.2017.7926765.
[11]. K. S. Rao and S. G. Koolagudi, “Recognition of emotions from video using acoustic and facial features,” Signal, Image and Video Processing, vol. 9, no. 5, pp. 1029–1045, Jul. 2013, doi: 10.1007/s11760-013-0522-6.
[12]. Tripathi, S., Tripathi, S., and Beigi, H. (2018). Multi-Modal Emotion Recognition on Iemocap Dataset Using Deep Learning. arXiv
[13]. D. Kamińska and A. Pelikant, “Recognition of Human Emotion from a Speech Signal Based on Plutchik’s Model,” International Journal of Electronics and Telecommunications, vol. 58, no. 2, pp. 165–170, Jun. 2012, doi: 10.2478/v10177-012-0024-4.
[14]. F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, “Vocal-based emotion recognition using random forests and decision tree,” International Journal of Speech Technology, vol. 20, no. 2, pp. 239–246, Feb. 2017, doi: 10.1007/s10772-017-9396-2.
[15]. E. Bagheri, A. Bagheri, P. G. Esteban, and B. Vanderborgth, “A Novel Model for Emotion Detection from Facial Muscles Activity,” Robot 2019: Fourth Iberian Robotics Conference, pp. 237–249, Nov. 2019, doi: 10.1007/978-3-030-36150-1_20.
[16]. P. Khorrami, T. L. Paine, and T. S. Huang, “Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition?,” 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 19–27, Dec. 2015, doi: 10.1109/iccvw.2015.12.
[17]. I. Mueller et al., “If it Ain’t Broke, Don’t Fix it: Positive Versus Negative Emotion Regulation in Daily Life and Depressive Symptoms,” Journal of Affective Disorders, vol. 348, pp. 398–408, Mar. 2024, doi: 10.1016/j.jad.2023.12.037.
[18]. K. A. Lindquist, J. K. MacCormack, and H. Shablack, “The role of language in emotion: predictions from psychological constructionism,” Frontiers in Psychology, vol. 6, Apr. 2015, doi: 10.3389/fpsyg.2015.00444.
[19]. C. MacCann, Y. Jiang, L. E. R. Brown, K. S. Double, M. Bucich, and A. Minbashian, “Emotional intelligence predicts academic performance: A meta-analysis.,” Psychological Bulletin, vol. 146, no. 2, pp. 150–186, Feb. 2020, doi: 10.1037/bul0000219.
[20]. D. Watson, L. A. Clark, and A. Tellegen, “Development and validation of brief measures of positive and negative affect: The PANAS scales.,” Journal of Personality and Social Psychology, vol. 54, no. 6, pp. 1063–1070, 1988, doi: 10.1037/0022-3514.54.6.1063.
[21]. E. Bagheri, A. Bagheri, P. G. Esteban, and B. Vanderborgth, “A Novel Model for Emotion Detection from Facial Muscles Activity,” Robot 2019: Fourth Iberian Robotics Conference, pp. 237–249, Nov. 2019, doi: 10.1007/978-3-030-36150-1_20.
[22]. P. Ekman and W. V. Friesen, “Facial Action Coding System,” PsycTESTS Dataset. American Psychological Association (APA), 1978. doi: 10.1037/t27734-000.