Date of publication SEP-30, 2025, date of current version MMM-dd, 2025.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.3.XXXX

The Sound and Sight of Confidence: An Audiovisual ML Approach

JYOTHSNA AN[1], PAMELA VINITHA ERIC[1], SMITHA RAO M.S[2]
[1]Presidency University, Bengaluru, Karnataka, India
[2]Vidyashilp University, Bengaluru, Karnataka, India

Corresponding author: Jyothsna AN (e-mail: forjyothsna@gmail.com).

ABSTRACT Research shows that confidence, which is crucial in conversations, has gained importance in various fields. Researchers have proved that the speaker's confidence can be gazed from face and as well as from their speech. This paper introduces a study on confidence level detection of speakers using multimodal AI approach which combines both video and audio modalities. With the development in the field of technology, capturing these cues has improved significantly. Obtaining the cues both from video and audio is crucial in the multimodality approach. We extracted features such as head pose, gaze direction as video features and spectral and prosodic features as audio features. With careful evaluation, we have achieved a notable accuracy of 85% with Gradient Boosting Machines along with AOC of 0.98 which emphasis on multimodality approach on fused test set. The findings highlight the importance of integrating visual and auditory cues to improve the accuracy of confidence level detection systems, with potential applications in education, public speaking, and virtual communication platforms.

KEYWORDS Multimodality; video-audio fusion; Machine Learning; Gradient Boosting; Early Fusion.

I. INTRODUCTION

Communication skills are crucial for success in today's professional landscape. Communication involves the exchange of information between individuals, typically through verbal and non-verbal methods, with verbal communication being the most common. Signage and symbols are usually used to convey the information. Paper [1] defines communication as the concurrent sharing and meaning creation to achieve goals and actively plays a role in effective communication. Confidence is crucial for achieving goals and plays a key role in effective communication. It provides significant personal and professional benefits, often leading to a more positive outlook on life. Factors like experience, social status, cultural background, and appearance can all impact confidence. Ultimately, confidence is vital for an individual's overall development. Effective communication is vital for success, shaping strong personal and professional relationships. It involves exchanging messages through verbal and non-verbal cues to achieve shared understanding.

A perceived natural phenomenon refers to modality, such as audio and speech, videos, and images. Unimodal systems serve as the foundational elements for constructing a multimodal system, making their strong performance, essential for developing an intelligent multimodal system. Paper [2] demonstrated that multimodality (fusion methods) has improved performance over unimodal systems. Paper [3] also elaborates about how sequential fusion methods performed better than unimodal. Based on this, multimodal communication is considered as all forms of signs or semiotics which occur in communication and include voice, text, images, body gestures, body language, video etc. As a result in this paper, we explore the multimodal fusion of unimodal information.

Organization of the paper: Section 2 talks about related work which concentrated mainly on the video and audio cues. Section 3 is about model design wherein detail discussion about data collection and methodology has been elaborately discussed. Section 4 talks about data pre-processing details of video and audio in which detail explanation data pre-processing and feature extraction has been discussed. Results of the implementation are presented in section 5. Section 6 is conclusion of the work.

II. RELATED WORK

With the abundance of data available today, machine learning methods have evolved significantly, leveraging extensive datasets to their full potential. Researchers have emphasized the need to enhance communication through non-verbal which should comprise facial expressions, visual features, body gestures and arm movements. Gestures, facial expressions,

tone of voice convey attitudes, emotions, and feelings, whereas verbal communication conveys exact words [4, 5]. Non-verbal signals, including gestures, eye movements, facial movements, postures, vocal behavior, have a greater impact on human interaction than verbal signals [6, 7]. According to Brunswick's Lens Model [8], non-verbal messages play a crucial role in communication alongside verbal messages. Furthermore, non-verbal cues provide more information than verbal signals, making them central to effective communication. While communicating with others there lies a positive relation and confidence in a topic being discussed [9]. Behavioral parameters indicating reliability and self-esteem are key measures of confidence [10]. A study on self-confidence explored its relationship with various personality traits and psychometric properties [11]. Confidence is defined as an assertive feeling of one's ability, grounded in security and realism, reflecting inner strength and knowledge, but it is not a sense of superiority over others.

Speakers frequently make visual cues which reflect their level of confidence [12, 13]. Interlocutors can detect a speaker's level of confidence based on their presentation. Studies suggest that observers automatically decode visual cues to assess a speaker's confidence [14], an impact observed in contexts like courtroom witness assessments [15, 16] or job interviews [17, 18]. Interlocutors use these non-verbal cues to evaluate the speaker's confidence, credibility, believability, trustworthiness [19-22]. In spontaneous communication, speakers show visual signals that reflect their cognitive processes, such as the retrieval of lexical and semantic information from memory to convey concepts through language [23]. These visual cues are used by speakers to pragmatically reinforce their message and signal reliability to others [14]. For instance, frequent gestures of the speakers just before speaking to aid in word retrieval from memory [24, 25]. With changes in eye gaze and facial expressions, research suggests that there is an indication of lexical retrieval [26, 27]. Depending on this, a conclusion can be drawn that low confidence speakers show averted gaze in contrast with high confidence speakers. This difference in the behavior of gaze serves as a reliable indicator of the level of confidence in the speaker [26, 28, 29].

As observed in [30], tone is more indicative of trustworthiness than pauses or stammering in audio communication. Paper [31] noted that in verbal communication, a person's confidence can be positively identified by their voice tone. Additionally, listeners perceived that the speakers with intonation falling at the end of the sentences show high confidence in contrast to those whose intonation is raised [32].

Various experiments have proved that speakers with confidence communicate with a louder tone in contrast to speakers with low confidence [33-36] research has shown that there exists a link between perceived confidence and vocal loudness. The results revealed that speakers with confident voices naturally spoke faster, with fewer pauses and louder. Research [33] also confirms that confidence individuals speak in higher pitch in accordance with acoustic analysis and listeners could judge these speakers with high confidence. This paper with 34 samples has achieved an UAR of 49% with deep learning architecture and multimodal late fusion techniques in detecting confidence of speakers in an interview conversation and to classify the level of confidence into high, medium, and low [32]. The literature review indicates that, there is a requirement for video-audio models to classify the levels of confidence. As a result, we designed a model which classifies audio and video into high confidence or low confidence based on facial movements and human voice.

III. MODEL DESIGN

Basic workflow of the system is shown in Figure 1.

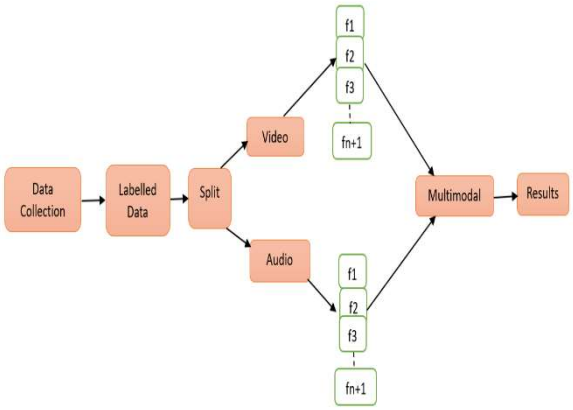

Figure 1. Basic Workflow of the system

A. DATA COLLECTION

To evaluate speaker confidence, data was collected through a systematically designed protocol, addressing the absence of publicly available annotated datasets for this specific task. Informed consent was obtained from all participants, and the recording procedure was clearly communicated prior to data collection. Video recordings were conducted in a controlled indoor environment to ensure consistency and minimize background noise or visual distractions. Participants were seated on a bench facing a laptop equipped with a front-facing camera. They were instructed to respond to a structured set of questions and to keep the recording active until all responses were completed. In cases of uncertainty, participants were encouraged to pause and reflect before answering. No evaluative feedback was provided during or after the session. The question set included two prompts designed to elicit low-confidence responses, requiring cognitive effort and uncertainty management, followed by two prompts intended to elicit high-confidence responses. The latter were straightforward, enabling participants to respond assertively while maintaining steady eye contact with the camera.

Video data from 65 participants, representing diverse age groups, was collected for this study. To ensure consistency across sessions, all participants were asked the same set of questions. A specialized questionnaire was developed to support confidence classification, comprising two distinct categories: low-confidence and high-confidence prompts. The low-confidence questions—such as *"What are Africa's Big Five animals?"*—were designed to elicit spontaneous, cognitively demanding responses, thereby enabling the capture of nuanced verbal and non-verbal expressions under uncertainty. In contrast, high-confidence questions—such as *"How many languages do you know, what are they, and which is your favorite?"*—focused on familiar, self-referential topics, allowing participants to respond with greater ease and assurance. These confident responses provided clear and reliable facial cues, facilitating more accurate analysis of speaker confidence.

B. METHODOLOGY

Model building and evaluation pipeline is shown in Figure 2.

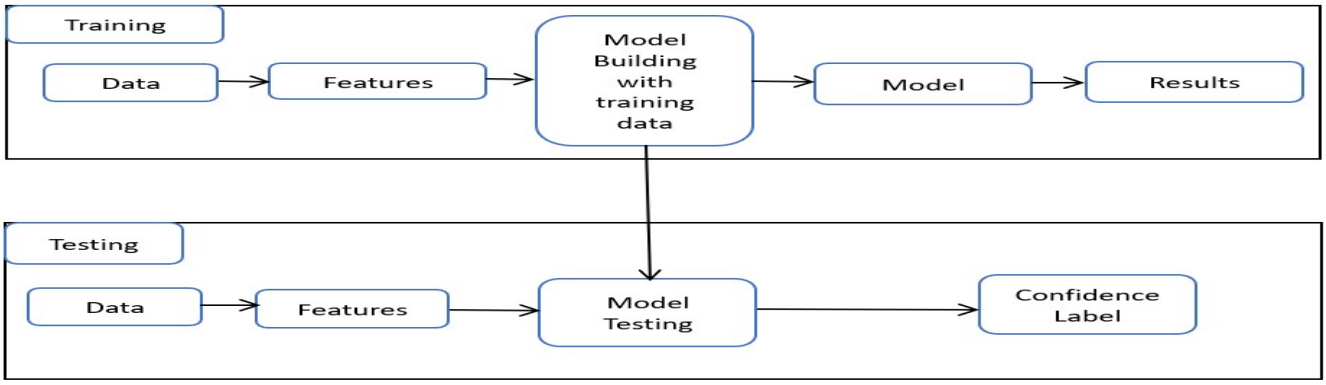

Figure 2. Model Building and Evaluation Pipeline

B.1. METADATA

According to [37], individuals tend to rely more heavily on facial expressions than vocal cues, as facial expressions are generally more effective in conveying emotions during direct face-to-face interactions. This study involved participants from diverse age groups, with a balanced representation of both male and female individuals. To ensure natural responses, participants were not informed of the specific purpose of the experiment during data collection. All participants had normal hearing and vision; while some wore eyeglasses, these did not obstruct visibility of key facial regions such as the eyes or eyebrows. Each participant entered the recording room individually and responded to a set of structured questions. The duration of the video recordings varied, ranging from 6 seconds to 30 seconds on average, with some extended responses lasting up to 30 minutes.

To strength our analysis, we classified the questions into low confidence and high confidence. This labelling allowed us to anticipate the video cues from participants. For low confidence questions, we observed visual cues such as eye-gaze, eye movement, eye-gaze (both vertical and horizontal), and head pose changes. In contrast, the high confidence questions were asked, speakers answered with confidence while maintaining steady eye contact with the camera, enabling us to capture their eye movements accurately. To ensure the accuracy of our labelling, we conducted random validation on the entire dataset to verify that the videos were correctly categorized as either low confidence or high confidence. Participants were asked between 2 to 4 questions. During the preprocessing step, some videos were excluded because they were either too short or the participants did not respond appropriately to the questions. The data collection took place within the setting of a real-time interview.

For the audio part, we extracted audio from the recorded videos, and relevant features were analyzed. The authors have used multimodal fusion for automating diagnosis and clinical outcome and proved that multimodal fusion has outperformed over unimodal [38].

In our research, we have applied multimodality approach by combining video and audio cues to effectively determine speaker's confidence.

IV. DATA PRE-PROCESSING

A. FEATURE EXTRACTION

After the data collection procedure, data was manually analyzed to verify the labels as either low confidence or high confidence. Gaze angles, including vertical and horizontal gaze values were extracted from videos. Positive values indicate the speaker is looking to the right, while negative values suggest a focus on the left, often signaling that the speaker is thinking while answering. If a speaker consistently looks to the left or right for an extended period, it is considered an indication of low confidence, as their gaze is not fixed on the camera. Conversely, when a speaker maintains eye contact with the camera while answering, with minimal gaze shifts to either side, it suggests high confidence. Additionally, statistical features related to gaze movements, such as standard deviation, minimum/maximum, mean, range, skewness, zero-crossing rate, first derivative mean standard, and kurtosis were calculated to provide further insight. Euler angles – pitch, yaw and roll are generally used to determine the head pose estimation, which defines head rotation in 3D environment. The above features are used to carry our analysis further. Additionally, statistical features related to gaze movements, such as standard deviation, minimum/maximum, mean, range, skewness, zero-crossing rate, first derivative mean standard, and kurtosis were calculated to enhance analysis.

In our analysis, we extracted a comprehensive set of audio features from the video recordings to enhance the understanding of speaker's confidence. Spectral features like Mel-Frequency Cepstral Coefficients (MFCCs) were utilized to capture the power spectrum, while the spectral centroid, bandwidth, contrast, and roll-off provided insights into the energy distribution within the audio signals. Temporal features, including root mean square (RMS) energy and zero-crossing rate (ZCR) were analyzed to assess signal variability and loudness. Prosodic features, such as pitch, formants, and intensity, were also extracted to understand the nuances of speech delivery. Additionally, harmonic features like the harmonic-to-noise ratio (HNR) and chroma features offered

information on the tonal quality and pitch classes present in the audio. This diversified set of features gave a robust foundation for evaluating the confidence levels of the speakers on the audio recordings.

In the analysis of audio data, certain features are crucial for accurately assessing characteristics such as speaker confidence. These features are extracted from the audio signals to capture relevant patterns that reflect the underlying attributes of the speaker. Following are the audio features that were extracted: MFCCs represent the short-term power spectrum of sound, which gives a compact representation of the spectral properties of the audio signal. High pitches or varying pitches may suggest uncertainty, while steady pitches are often associated with confidence. Positions and movements of formants can provide insights into the speaker's articulation and vocal quality. Zero-Crossing Rate- ZCR feature is often used to detect the presence of voice activity or distinguish between different types of sounds. Spectral Features help to differentiate between different vocal qualities. The extraction of these features allows us to model and analyze the nuances of the speaker's voice, providing a basis for distinguishing between different levels of confidence. By focusing on these features, our analysis can capture both the acoustic and perceptual aspects of the audio data. Approximately 288 features were extracted for the analysis for both video and audio.

V. MODEL BUILDING

The training dataset consists of 273 samples distinguished between low confidence and high confidence. After early fusion we got test samples as 137 which were distributed equally into low confidence and high confidence. The dataset has samples which were distributed between low and high confidence.

A. VIDEO RESULT ANALYSIS

Below graph shows the performance of several machine learning models. All exhibit moderate accuracy levels. Additionally, the validation for each model shows strong results. A 5-fold cross-validation was applied to the training data to ensure these outcomes. The visualization clearly highlights the varying generalization performance of various models.

The Naive Bayes classifier demonstrated a surge in performance with a strong 92% on the test set, suggesting effective generalization for this specific test distribution. Random Forest with 80% on the test data. Naïve Bayes a strong performer, indicates that it effectively handles the data used in this scenario, outperforming several other classifiers. This suggests that Naive Bayes, despite its simplicity, can be highly effective when its assumptions about the data hold true, making it a robust choice for this specific classification task.

All the models' accuracy has been shown, and all the models have demonstrated good accuracy scores on both train and test as well. The Naive Bayes classifier demonstrated high accuracy and successfully distinguished between classes with an AUC of 0.90.

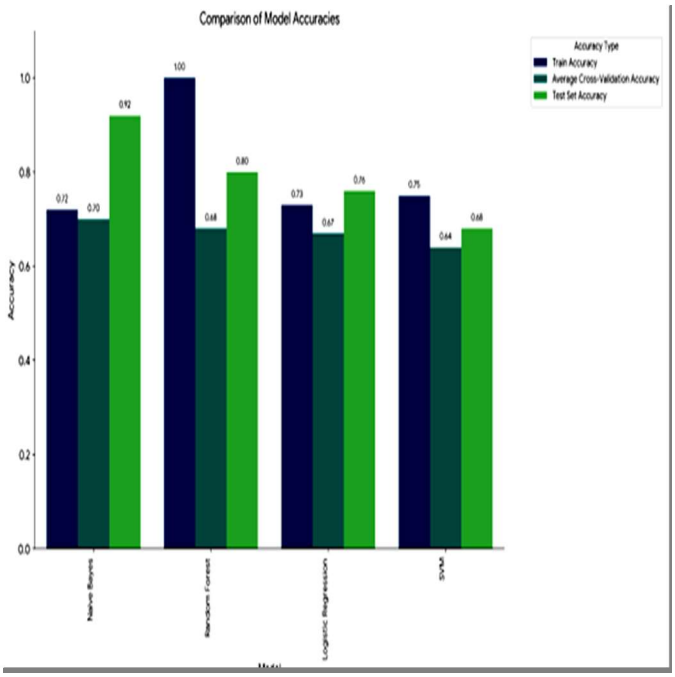

Figure 3. Train, cross validation & Test accuracy

B. AUDIO RESULT ANALYSIS

Train & test accuracy is shown in Figure 4.

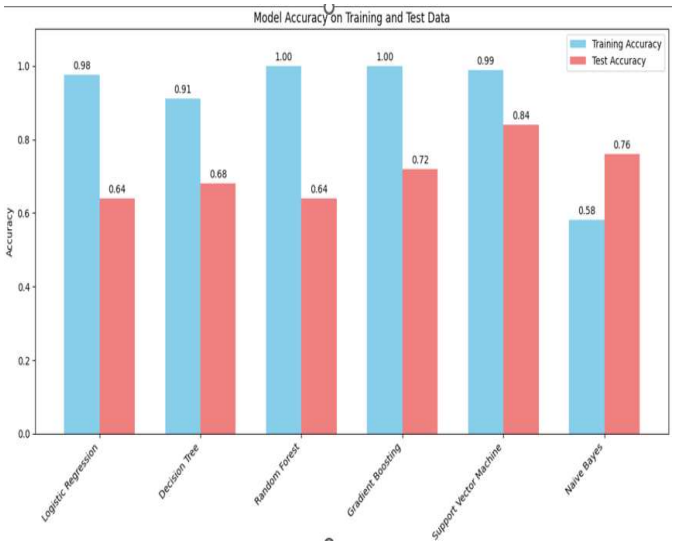

Figure 4. Train & Test accuracy

The bar chart represents the train and test accuracies of various ML models for audio classification. Support Vector Machine has achieved the accuracy of 84% followed by Gradient Boosting, Naïve Bayes.

C. VIDEO-AUDIO FUSION MODEL RESULTS

Test Accuracy with CV of all the ML models is shown in Figure 5.

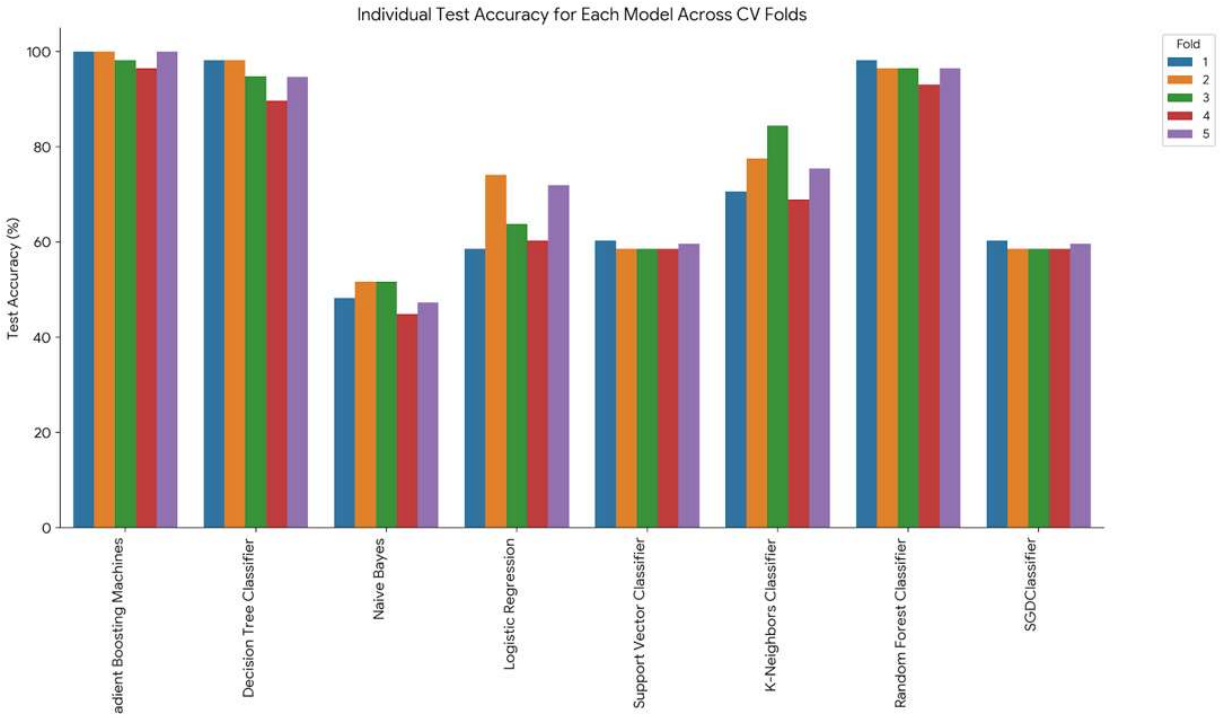

Figure 5. Test Accuracy with CV of all the ML models

Top Performers: The Gradient Boosting Machines and Random Forest Classifier consistently achieve the highest accuracies, often hitting 95% or higher across all folds. This reinforces their strong overall performance seen in the mean accuracies. Models like Support Vector Classifier and SGDClassifier show very little variation across folds, while others like Logistic Regression and K-Neighbours Classifier have more fluctuation. Models like Naive Bayes consistently show much lower accuracies. It is shown in the above graph.
The following fusion methos is applied in the analysis of the data. Combining video and audio cues, we are generating a single vector and then fed into the model.

$$f_{()} = f_{(vm)} + f_{(am)}$$

Figure 6. Formula used in the analysis

To achieve this, early fusion technique, where we are combining the various features extracted from each modality (video & audio) independently and combined into a single feature vector. This feature vector is fed into the ML models for further analysis.

As mentioned above the top performer is Gradient Boosting Machines, the same algorithm is used for further model training.

Classification Report on Fused Test Set:

	precision	recall	f1-score	support
confident	0.97	0.64	0.77	55
not_confident	0.80	0.99	0.89	82
accuracy			0.85	137
macro avg	0.89	0.81	0.83	137
weighted avg	0.87	0.85	0.84	137

Figure 7. Classification Report of the best model – Gradient Boost

The above classification report is the result of the best model trained based on the mean test accuracy with cross validation. Gradient Boost has outperformed over all the models with 85% accuracy.

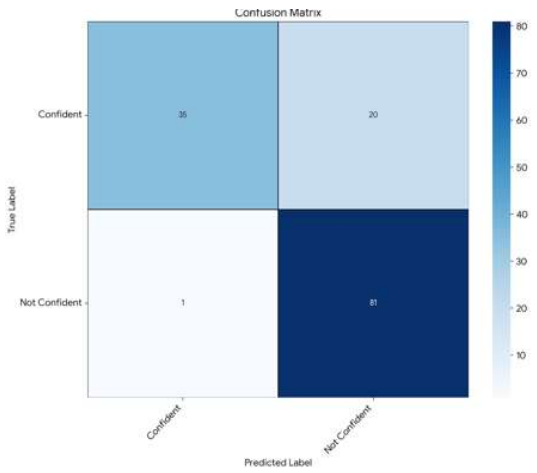

Figure 8. Confusion matrix on test data of Gradient Boost

In the Fig. 8, the confusion matrix depicts that the model is particularly strong at identifying instances that are 'Not Confident' with very few misclassifications. While it performs well on 'Confident' instances, it does have a notable number of false negatives (20), meaning it sometimes fails to recognize a 'Confident' instance, classifying it as 'Not Confident'.

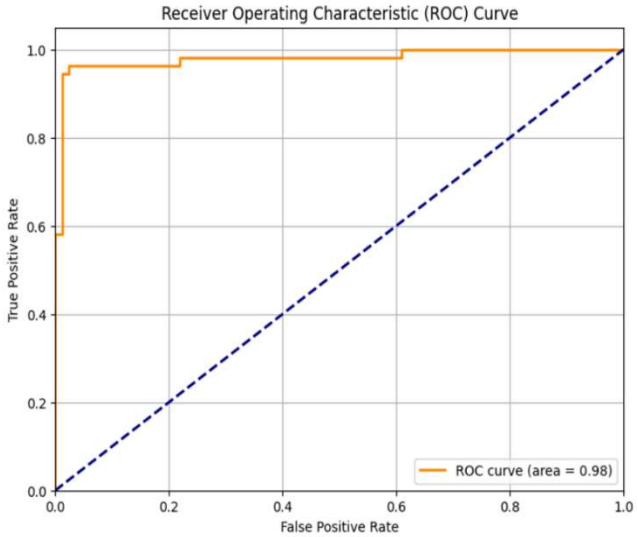

Figure 9. ROC of Gradient Boost

AUC of 0.98 has significantly showed strong discrimination capacity within the target classes. With this score model has depicted the strong ability in classifying the target classes into low confidence and high confidence.

VI. CONCLUSION

Our research developed successfully a novel AI with multimodal framework to detect speaker confidence, leveraging strategy of fusion by incorporating video and audio features. This comprehensive approach allowed our machine learning models to capture the subtle, cross-modal indicators of self-assuredness and delivery style. This work provides a transparent and highly effective means of analysing speaker behaviour, moving beyond purely textual or generalized affective approaches. The choice of using traditional ML algorithms is rational, considering our sample size and efficiency of computation. Looking ahead, while our current framework delivers strong results, future research will explore the integration of deep learning architectures. This will aim to potentially enhance prediction accuracy further by automatically learning hierarchical representations from raw multimodal data, while critically maintaining a focus on developing methods for model transparency and interpretability to ensure actionable insights. The established framework holds direct implications for advancing intelligent systems in communication training, public speaking programs, and virtual interaction platforms.

Among the various classifiers evaluated, the Gradient Boosting Machines model emerged as the most robust and accurate performer. It demonstrated exceptional generalization capabilities, achieving a final accuracy of 85% on an independent, unseen fused test set. Crucially, its discriminatory power was further validated by an outstanding Area Under the Receiver Operating Characteristic Curve (AUC) of 0.98, indicating its strong ability to differentiate between confident and non-confident states. This performance underscores the suitability of ensemble methods like Gradient Boosting for complex, feature-rich multimodal datasets, effectively learning intricate interactions between diverse signals while maintaining a degree of interpretability

References

[1] W. J. Seiler, & M. L. Beall, Communication: Making connections, 6th. Ed., Boston: Allyn & Bacon, 2005.

[2] R. N. Rodrigues, et al., "Robustness of multimodal biometric fusion methods against spoof attacks," *Journal of Visual Language and Computing*, vol. 20, issue 3, pp. 169-179, 2009, https://doi.org/10.1016/j.jvlc.2009.01.010.

[3] U. Gawande and Y. Golhar, "Biometric security system: a rigorous review of unimodal and multimodal biometrics techniques," *Int. J. Biometrics*, vol. 10, no. 2, pp. 142–175, 2018. https://doi.org/10.1504/IJBM.2018.10012749.

[4] D. Conrad and R. Newberry, "24 business communication skills: attitudes of human resource managers versus business educators," *Am. Commun. J.*, vol. 13, pp. 4–23, 2011.

[5] G. Mohammadi and A. Vinciarelli, "Towards a technology of nonverbal communication: vocal behavior in social and affective phenomena," in *Affective Computing and Interaction: Psychological, Cognitive, and Neuroscientific Perspectives*, D. Gokcay and G. Yildirim, Eds. IGI Global, 2012. https://doi.org/10.4018/978-1-61692-892-6.ch007.

[6] M. Cai and J. Tanaka, "Go together: providing nonverbal awareness cues to enhance co-located sensation in remote communication," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, p. 19, 2019. https://doi.org/10.1186/s13673-019-0180-y.

[7] J. B. Walther, "Theories of computer-mediated communication and interpersonal relations," in *The Handbook of Interpersonal Communication*, M. L. Knapp and J. A. Daly, Eds., Thousand Oaks, CA: SAGE, 2011, pp. 443–479.

[8] S. Nestler and M. D. Back, "Applications and extensions of the lens model to understand interpersonal judgments at zero acquaintance," *Curr. Dir. Psychol. Sci.*, vol. 22, pp. 374–379, 2013. https://doi.org/10.1177/0963721413486148.

[9] E. Cech, B. Rubineau, S. Silbey, and C. Seron, "Professional role confidence and gendered persistence in engineering," *Am. Sociol. Rev.*, vol. 76, pp. 641–666, 2011, https://doi.org/10.1177/0003122411420815.

[10] P. D. Bennett and G. D. Harrell, "The role of confidence in understanding and predicting buyers' attitudes and purchase intentions," *J. Consum. Res.*, vol. 2, pp. 110–117, 1975, https://doi.org/10.1086/208622.

[11] F. Meyniel, M. Sigman, and Z. F. Mainen, "Confidence as Bayesian probability: From neural origins to behavior," *Neuron*, vol. 88, pp. 78–92, 2015. https://doi.org/10.1016/j.neuron.2015.09.039.

[12] T. O. Nelson, and L. Narens, "Metamemory: a theoretical framework and some new findings," in The Psychology of Learning and Motivation, ed G. H. Bower (San Diego, CA: Academic Press), 125–173, 1990. https://doi.org/10.1016/S0079-7421(08)60053-5.

[13] A. Boduroglu, A. I. Tekcan, and A. Kapucu, "The relationship between executive functions, episodic feeling-of-knowing and confidence judgements," *J. Cogn. Psychol.*, vol. 26, pp. 333–345, 2014. https://doi.org/10.1080/20445911.2014.891596.

[14] W. G. Moons, J. R. Spoor, A. E. Kalomiris, and M. K. Rizk, "Certainty broadcasts risk preferences: verbal and nonverbal cues to risk-taking," *J. Nonverbal Behav.*, vol. 37, pp. 79–89, 2013. https://doi.org/10.1007/s10919-013-0146-0.

[15] R. J. Cramer, S. L. Brodsky, and J. Decoster, "Expert witness confidence and juror personality: their impact on credibility and persuasion in the courtroom," *J. Am. Acad. Psychiatry Law*, vol. 37, pp. 63–74, 2009.

[16] R. J. Cramer, C. T. Parrott, B. O. Gardner, C. H. Stroud, M. T. Boccaccini, and M. P. Griffin, "An exploratory study of meta-factors of expert witness persuasion," *J. Individ. Differ.*, vol. 35, pp. 1–11, 2014. https://doi.org/10.1027/1614-0001/a000123.

[17] T. DeGroot, J. Gooty, "Can nonverbal cues be used to make meaningful personality attributions in employment interviews?" *J. Bus. Psychol.*, vol. 24, pp. 179–192, 2009. https://doi.org/10.1007/s10869-009-9098-0.

[18] T. DeGroot, and S. J. Motowidlo, "Why visual and vocal interview cues can affect interviewers' judgments and predict job performance," *J. Appl. Psychol.*, vol. 84, pp. 986–993, 1999. https://doi.org/10.1037/0021-9010.84.6.986.

[19] J. Brosy, A. Bangerter, and E. Mayor, "Disfluent responses to job interview questions and what they entail," *Discourse Process*, vol. 53, pp. 371–391, 2016. https://doi.org/10.1080/0163853X.2016.1150769.

[20] S. A. J. Birch, N. Akmal, and K. L. Frampton, "Two-year-olds are vigilant of others' non-verbal cues to credibility," *Dev. Sci.*, vol. 13, pp. 363–369, 2010. https://doi.org/10.1111/j.1467-7687.2009.00906.x.

[21] F. Meyniel, M. Sigman, and Z. F. Mainen, "Confidence as Bayesian probability: From neural origins to behavior," *Neuron*, vol. 88, pp. 78–92, 2015, https://doi.org/10.1016/j.neuron.2015.09.039.
[22] T. O. Nelson and L. Narens, "Metamemory: a theoretical framework and some new findings," in *The Psychology of Learning and Motivation*, G. H. Bower, Ed., San Diego, CA: Academic Press, 1990, pp. 125–173, https://doi.org/10.1016/S0079-7421(08)60053-5.
[23] A. Boduroglu, A. I. Tekcan, and A. Kapucu, "The relationship between executive functions, episodic feeling-of-knowing and confidence judgements," *J. Cogn. Psychol.*, vol. 26, pp. 333–345, 2014, https://doi.org/10.1080/20445911.2014.891596.
[24] W. G. Moons, J. R. Spoor, A. E. Kalomiris, and M. K. Rizk, "Certainty broadcasts risk preferences: verbal and nonverbal cues to risk-taking," *J. Nonverbal Behav.*, vol. 37, pp. 79–89, 2013, https://doi.org/10.1007/s10919-013-0146-0.
[25] R. J. Cramer, S. L. Brodsky, and J. Decoster, "Expert witness confidence and juror personality: their impact on credibility and persuasion in the courtroom," *J. Am. Acad. Psychiatry Law*, vol. 37, pp. 63–74, 2009.
[26] R. J. Cramer, C. T. Parrott, B. O. Gardner, C. H. Stroud, M. T. Boccaccini, and M. P. Griffin, "An exploratory study of meta-factors of expert witness persuasion," *J. Individ. Differ.*, vol. 35, pp. 1–11, 2014, https://doi.org/10.1027/1614-0001/a000123.
[27] T. DeGroot and J. Gooty, "Can nonverbal cues be used to make meaningful personality attributions in employment interviews?" *J. Bus. Psychol.*, vol. 24, pp. 179–192, 2009, https://doi.org/10.1007/s10869-009-9098-0.
[28] T. DeGroot and S. J. Motowidlo, "Why visual and vocal interview cues can affect interviewers' judgments and predict job performance," *J. Appl. Psychol.*, vol. 84, pp. 986–993, 1999, https://doi.org/10.1037/0021-9010.84.6.986.
[29] J. Brosy, A. Bangerter, and E. Mayor, "Disfluent responses to job interview questions and what they entail," *Discourse Process.*, vol. 53, pp. 371–391, 2016, https://doi.org/10.1080/0163853X.2016.1150769.
[30] S. A. J. Birch, N. Akmal, and K. L. Frampton, "Two-year-olds are vigilant of others' non-verbal cues to credibility," *Dev. Sci.*, vol. 13, pp. 363–369, 2010, https://doi.org/10.1111/j.1467-7687.2009.00906.x.
[31] X. Jiang and M. D. Pell, "On how the brain decodes vocal cues about speaker confidence," *Cortex*, vol. 66, pp. 9–34, 2015, https://doi.org/10.1016/j.cortex.2015.02.002.
[32] X. Jiang, R. Sanford, and M. D. Pell, "Neural systems for evaluating speaker (un)believability," *Hum. Brain Mapp.*, vol. 38, pp. 3732–3749, 2017, https://doi.org/10.1002/hbm.23630.
[33] Y. Mori and M. D. Pell, "The look of (un)confidence: Visual markers for inferring speaker confidence in speech," *Front. Commun.*, vol. 4, p. 63, 2019, https://doi.org/10.3389/fcomm.2019.00063.
[34] B. Rimé and L. Schiaratura, "Gesture and speech," in *Studies in Emotion & Social Interaction: Fundamentals of Nonverbal Behavior*, R. S. Feldman and B. Rimé, Eds., New York, NY: Cambridge University Press, 1991, pp. 239–281.
[35] R. M. Krauss, Y. Chen, and P. Chawla, "Nonverbal behavior and nonverbal communication: what do conversational hand gestures tell us?," in *Advances in Experimental Social Psychology*, M. Zanna, Ed., San Diego, CA: Academic Press, 1996, pp. 389–450, https://doi.org/10.1016/S0065-2601(08)60241-5.
[36] M. H. Goodwin and C. Goodwin, "Gesture and coparticipation in the activity of searching for a word," *Semiotica*, vol. 62, pp. 51–76, 1986, https://doi.org/10.1515/semi.1986.62.1-2.51.
[37] J. Bavelas and J. Gerwing, "Conversational hand gestures and facial displays in face-to-face dialogue," in *Social Communication*, K. Fiedler, Ed., New York, NY: Psychology Press, 2007, pp. 283–308.
[38] P. Ekman and W. V. Friesen, *Manual for the Facial Action Coding System*, Palo Alto, CA: Consulting Psychologists Press, 1978, https://doi.org/10.1037/t27734-000.
[39] E. Krahmer and M. Swerts, "How children and adults produce and perceive uncertainty in audiovisual speech," *Lang. Speech*, vol. 48, pp. 29–53, 2005, https://doi.org/10.1177/00238309050480010201.
[40] J. A. Hall, "Voice tone and persuasion," *J. Personal. Soc. Psychol.*, vol. 38, pp. 924–934, 1980, https://doi.org/10.1037/0022-3514.38.6.924.
[41] N. D. Cook, *Tone of Voice and Mind: The Connections between Intonation, Emotion, Cognition, and Consciousness*, John Benjamins Publishing, 2002. https://doi.org/10.1075/aicr.47.
[42] S. Chanda, K. Fitwe, G. Deshpande, B. W. Schuller, and S. Patel, "A deep audiovisual approach for human confidence classification," *Front. Comput. Sci.*, vol. 3, p. 674533, 2021, https://doi.org/10.3389/fcomp.2021.674533.
[43] X. Jiang and M. D. Pell, "The sound of confidence and doubt," *Speech Commun.*, vol. 88, pp. 106–126, 2017, https://doi.org/10.1016/j.specom.2017.01.011.
[44] C. E. Kimble and S. D. Seidel, "Vocal signs of confidence," *J. Nonverbal Behav.*, vol. 15, no. 2, pp. 99–105, 1991, https://doi.org/10.1007/BF00998265.
[45] K. R. Scherer, H. London, and J. J. Wolf, "The voice of confidence: Paralinguistic cues and audience evaluation," *J. Res. Personal.*, vol. 7, no. 1, pp. 31–44, 1973, https://doi.org/10.1016/0092-6566(73)90030-5.
[46] A. B. Van Zant and J. Berger, "How the voice persuades," *J. Personal. Soc. Psychol.*, vol. 118, no. 4, pp. 661–682, 2020, https://doi.org/10.1037/pspi0000193.
[47] A. Irvine, P. Drew, and R. Sainsbury, "Am I not answering your questions properly? Clarification, adequacy and responsiveness in semi-structured telephone and face-to-face interviews," *Qual. Res.*, vol. 13, pp. 87–106, 2013, https://doi.org/10.1177/1468794112439086.
[48] S. C. Huang, A. Pareek, R. Zamanian et al., "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection," *Sci. Rep.*, vol. 10, p. 22147, 2020, https://doi.org/10.1038/s41598-020-78888-w.

Ms. Jyothsna AN, ***Research scholar, Presidency University, Bengaluru, Karnataka, India. Area of interests: Artificial Intelligence, Machine Learning, Multimodal Learning.***

Dr. Pamela Vinitha Eric, ***Professor, School of Computer Science & Engineering, Presidency University, Karnataka, India. Research Interests: Bioinformatics, Data compression, Network Security, Cryptography.***

Dr. Smitha Rao M. S ***Program Chair, – Computer science and Engineering (Data Science), Vidyashilp University, Bengaluru, Karnataka, India. Research interests: Deep Learning Modelling for Business applications, NLU, Multi Modal learning, Machine Learning, Data Visualization Storytelling.***