American International University-Bangladesh (AIUB)
Faculty of Science and Technology

Real-Time Multimodal Confidence Analysis

Al Fahad (22-47125-1)
S.M. Nafees Hossain Niloy (22-47333-2)
Rubayet Alam Azan (22-46888-1)
Abdullah Al Marjuc Talha (22-47294-1)

A Thesis submitted for the degree of Bachelor of Science (BSc)in
Computer Science and Engineering (CSE) at
American International University Bangladesh (AIUB
Faculty of Science and Technology (FST)

Fall 2024-2025 Semester
Submission Date: December, 2025

Declaration

This thesis is composed of our original work, and contains no material previously published or writtenby another person except where due reference has been made in the text. We have clearly stated the contribution of others to our thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financialand any other original research work used or reported in our thesis. The content of our thesisis the result of work we have carried out since the commencement of the Thesis.

We acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate we have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors forany jointly authored works included in the thesis.

Al Fahad
22-47125-1
Computer Science & Engineering

S.M. Nafees Hossain Niloy
22-47333-2
Computer Science & Engineering

Rubayet Alam Azan
22-46888-1
Computer Science & Engineering

Abdullah Al Marjuc Talha
22-47294-2
Computer Science & Engineering

Approval

The thesis titled **"Real-Time Multimodal Confidence Analysis"** has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfilment of the requirementsfor the degree of Bachelor of Science in Computer Science on **(4th December 2025)** and has been acceptedas satisfactory.

Sazzad Hossain
Assistant Professor & Supervisor
Department of Computer Science
American International University-Bangladesh

DR. MD. SAEF ULLAH MIAH
Associate Professor, Additional Director
[IQAC]& External
Department of Computer Science
American International University-Bangladesh

Dr. Debajyoti Karmaker
Associate Professor & Head (UG)
Department of Computer Science
American International University-Bangladesh

Prof. Dr. Dip Nandi
Professor & Associate Dean
Faculty of Science and Technology
American International University-Bangladesh

Mashiour Rahman
Sr. Associate Professor & Dean
Faculty of Science and Technology
American International University-Bangladesh

Acknowledgement

We would like to express our sincere gratitude to Almighty for enabling us to complete the report on "**Real-Time Multimodal Confidence Analysis**". For the successful completion of this project, we have taken help from our respected faculty members, seniors and graduates. We convey our sincere gratitude to our respectable supervisor, **Sazzad Hossain sir, Assistant Professor, Dept. of Computer Science**, for taking us under his wing and guiding us to conduct research on this project, thereby gathering knowledge in the field of **Human-Computer Interaction (HCI) and Computer Vision.**

Author Contributions

List the significant and substantial inputs made by different authors to this research, work and writingrepresented and/or reported in the thesis. These could include significant contributions to the conception and design of the project; non-routine technical work; analysis and interpretation of researchdata; drafting significant parts of the work or critically revising it to contribute to the interpretation.

	Al Fahad	S.M. Nafees Hossain Niloy	Rubayet Alam Azan	Abdullah Al Marjuc Talha	Comments
	22-47125-1	*22-47333-1*	*22-46888-1*	*22-47294-1*	
0 - 3 points	**Perform as effective individual**				
Critical thinking	3	3	3	3	
Reflection on feedback	3	3	3	3	
Quality of work	3	3	3	3	
Self-directed	3	3	3	3	
0 - 3 points	**Perform as effective team member/leader**				
Taking responsibility	3	3	3	3	
Contribution	3	3	3	3	
Collaboration	3	3	3	3	
Working with others	3	3	3	3	
0 - 3 points	**Perform as effective team member/leader**				
Presentation delivery	3	3	3	3	
Voice and tone	3	3	3	3	
Enthusiasm	3	3	3	3	
Creativity & Tools use	3	3	3	3	

Project-Thesis Planning

Project Tasks	Schedule Data	Execution Data
1. Planning	2025.04.06-2025.04.30	2025.04.06-2025.05.03
2. Literature Review	2025.04.30	2025.05.05
3. Survey question selection and review	2025.04.08	2025.04.10
4. Conduct survey/interview	2025.05.09 - 2025.07.13	2025.05.11-2025.07.17
5. Writing thesis report	2025.04.09 - 2025.08.12	2025.04.20-2025.08.29
6. Submission and review	2025.08.29 - 2025.11.06	2025.08.11-2025.12.04

Table: Project-Thesis Deliverables

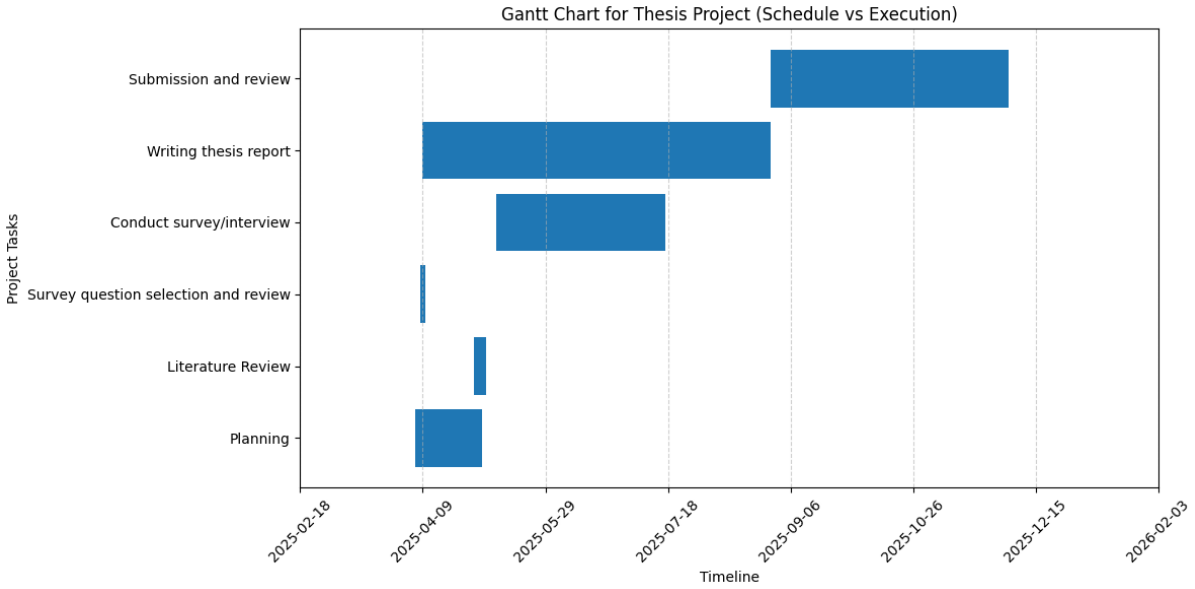

Figure: Thesis Planning Gantt chart.

Table of Content

Contents

List of Figures

List of Tables

List of Abbreviations

Mention all the abbreviations and the different symbols that are used in this document.

AI	Artificial Intelligence
API	Application Programming Interface
HCI	Human Computer Interaction
CNN	Convolutional Neural Network
DSRM	Design Science Research Methodology
EEG	Electroencephalography
ECG	Electrocardiography

Abstract

The explosion of virtual communication and remote interaction has opened up a need for real-time systems that can evaluate the human confidence. The present thesis proposes a web-based model which integrates facial, hand gesture, and vocal analysis to detect and quantify confidence. The system uses computer vision techniques to identify and follow facial landmarks precisely, which allows for very detailed analysis of eye contact, gaze direction, facial expressions, head pose, and facial symmetry. The understanding of such facial features is critical in determining the level of attentiveness, emotions, and the amount of engagement in the live interactions. Hand gesture analysis is processed through hand landmarks detection, thus, the system can detect the type of gesture, palm orientation, finger positions, and even the speed of movement. This part of the system is capturing not only the conscious but also the subconscious hand movements, which are very often the signs of either confidence or nervousness. The vocal analysis part of the system takes advantage of audio processing in real-time for feature extraction and these features include pitch, jitter, loudness, and spectral centroid. All these aspects of one's voice help in distinguishing if the speaker is nervous or calm. One major innovation of the model is its confidence scoring mechanism, which combines the information from the three modalities into a single, continuously variable score ranging from 0 to 100. The scoring system gives peak vocal characteristics, positive facial expressions, and confident hand gestures applied with weighted contributions to accomplish an even and thorough evaluation. The real-time updates facilitate constant feedback, thereby rendering the system applicable for online interviews, presentations, and e-learning environments. The model that has been suggested is robust and versatile, so that it can support various practices.

Keywords

Multimodal Confidence Analysis, Facial Expressions, Hand Gestures, Vocal Features, MediaPipe, Web Audio API, Behavioral Cues.

CHAPTER 1
INTRODUCTION

The rising use of virtual communication in teaching, medical care, and working environments has created a demand for systems that can detect user confidence and emotions in real-time. Conventional evaluation techniques like self-reports or post-event analysis often suffer from a lack of immediacy and objectivity, thereby rendering them less effective in fast-paced digital scenarios [1,2]. The recent progress in machine learning and multimodal data fusion have made it possible to create systems that are capable of processing facial expressions, hand movements, and vocal signals to give a thorough insight of the user states in real-time interactions [3,4]. Incorporating various forms of data analysis, multimodal techniques take advantage of the interrelationships among different data types like visual, auditory, and textual signals to reveal the intricacies of human communication [3,5,6]. As a case in point, the pairing of real-time face orientation detection and gesture recognition has been found to better support the tracking of participant engagement and focus in virtual meetings and classrooms [7]. Not only do these technologies enhance digital communication, they also open up new possibilities for the development of adaptive systems that can respond to the users' needs as the interactions take place [2,4].

The combination of sophisticated machine learning methods and real-time data processing systems goes a long way in providing instant feedback and assistance in virtual environments [3,5]. These systems have shown to be very precise in detecting emotion and engagement, with a few reaching more than 90% accuracy for facial and vocal cues [8,1,7]. Gradually, the direction of the industry is switching to the development of large, personalized systems that are able to not only adapt but also work effortlessly within various groups and situations, and that would be the end result of such a development—more productive and significant digital interactions [2,6].

1.1 Background Analysis

The fast change to online and virtual environments in education, workplaces, and social interactions has paved the way for the effective and real-time assessment of user emotions and engagement to be done. The classic self-report and post-event analysis techniques are

sometimes incapable of portraying the changing and subtle emotions of digital users, which in turn has led to the creation of the automated multimodal systems that are capable of utilizing the recent advancements in machine learning and data fusion [9,10,11,12].

In the beginning, studies looked at single-modal methods like facial expression recognition to determine emotional states and engagement. For instance, neural networks, which are trained on facial images, have been utilized to differentiate emotions and engagement levels in e-learning with superb accuracy and real-time feedback without violating privacy [9,10,13]. Nevertheless, these procedures might be challenged by social masking or unclear expressions, hence the merging of extra modalities like speech, hand gestures, and physiological signals [12,14,15].

Nowadays, the newest devices and systems have the ability to mastervisual, auditory, and textual data all at once, thus enhancing their robustness and being more accurate at the same time. Deep learning has been applied in the biomedical field through various models, like Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM), and transformer-based architectures, for fusion and portrayal of the different modalities, leading to emotion and engagement detection that is more reliable than before and even in noisy or unpredictable environments [5,3]. Besides that, the use of multimodal frameworks has shown better results than the single-modality ones, with some achieving over 90% accuracy in real-time scenarios [10,7]. Furthermore, the application of EEG, as a physiological signal, boosts the objectivity and sensitivity to affective states even more, particularly when combined with other biosignals [12,14].

These technologies have found their way into online education, virtual reality, and adaptive learning systems, which all benefit from real-time feedback to improve engagement and learning outcomes [15,8,16,7]. Incorporation into popular platforms like Microsoft Teams has proven to be a promising strategy for scaling up and practical deployment, even though there are still issues to resolve, such as detecting complicated emotions and making the technology work flawlessly with different user groups [8,17,18].

The hurdles in real-time processing, data synchronization, and the transparency of AI-driven decisions have not completely vanished even though there has been considerable development in these areas. The problem is tackled by the research community through different methods

like lightweight models, better fusion algorithms, and making AI more human-readable techniques [12,14,3,19]. The goal is to build adaptive, interpretable, and dependable systems applicable in various digital settings.

1.2 Existing Studies

Multimodal Data for Confidence Detection

1.2.1 The Role of Multimodal Data in Confidence Detection

Unimodal systems, which depend on a single type of data such as facial expressions, speech, or physiological signals, are often unable to grasp the whole mixture of human confidence, particularly in real-life or uncertain situations. Besides that, these systems are restricted by problems such as surrounding noise, obstruction, and the masking or subtleness of emotional signals, which can result in decreased accuracy and reliability [20,21,22,23]. One illustration is that facial expression analysis can be distorted by lighting or head pose, whereas systems relying on speech may be impacted by background noise or individual differences in speaking style [20,22].

Recent research has shown that the combination of different modalities is one of the main factors that effectively and reliably increase the performance of detection systems for confidence and emotion. Güler and Akbulut applied advanced deep learning models (e.g., GRU) on EEG signals and facial expressions fusion, and they found that the classification accuracy was raised from the EEG-only 91.8% to the multimodal 97.8%, thus revealing the large profits from multimodal fusion [24]. In the same manner, Thakur and his co-workers made a sweeping review of the area and concluded that deep learning-based multimodal systems, especially the ones which are synchronizing audio and video, regularly go beyond the unimodal systems, particularly in intricate or noisy environments [20]. Moreover, Wang et al. multiplied the integration of physical (audio, visual) and physiological (EEG, ECG) signals to get a richer and more accurate depiction of affective states thereby breaking the barriers of the single-modality systems [21]. However, the mentioned approaches are still challenged by the demand for sophisticated fusion methods, the alignments of different data types and the higher computation load [20,21,25].

1.2.2 Applications in Online Learning Environments

In the domain of online education, the precise detection of confidence is very important for the monitoring of student's participation and the adjusting of teaching methods. Villegas-Ch. et al. created a system for real-time detection of emotions in virtual classrooms using multiple modes, which was a combination of speech and facial expression analysis conducted through Microsoft Teams. Their system reached up to 95% accuracy for the detection of positive emotions and led to a rise in student participation, but it was difficult to recognize the complex emotions of stress and frustration, thus showing the necessity of further enhancement and wider sources of data [8]. In a similar vein, Amaresh and Aote suggested the use of an AI-assisted psychological evaluation framework based on EEG, voice and questionnaires, resulting in 94.3% accuracy through multimodal fusion—much higher than the unimodal models. Their system is tolerant and customizable for real-time feedback, but still needs careful attention to privacy and data integration issues [26]. Tellamekala et al. proposed a novel technique of multimodal fusion that is aware of the uncertainty involved, which not only led to higher accuracy of emotion recognition but also facilitated the system's robustness to noise and missing data, a problem frequently encountered in online environments [25].

1.2.3 Where Unimodal Systems Fail and Why Multimodal Detection Is Necessary

Unimodal systems are vitiated by errors when the cues are very faint, masked or dependent on the context and thus do not have the redundancy required for robust real-world applications [20,21,22,23]. The integration of facial, vocal and physiological signals in multimodal methods results to a more thorough and trustworthy judgment of confidence, particularly in dynamic or loud environments. This fusion becomes necessary in the case of applications requiring real-time processing, where the loss or misinterpretation of just one signal might result in wrong reading of the user's confidence or involvement [24,20,21,25,8].

1.2.4 How Past Works Guide System Design

The literature highlights the importance of using multimodal integration, sophisticated fusion techniques, and processing in real-time, thus directing the design of the upcoming generation confidence detection systems. Among the major recommendations, one can find the creation of flexible architectures, the enhancement of data synchronization, and the implementation of ethical handling of sensitive data [20,21,26,25,8]. These findings help in choosing modalities, fusion techniques, and metrics for robust, real-world confidence detection systems that can be evaluated.

1.2.5 Research gap

The topic of multimodal affect and confidence detection has come a long way, though there are still a lot of research challenges ahead. Most of the current systems are still on the way of incorporating various cues like facial expressions, vocalizations, and physiological signals, but there are still problems in data fusion, noise reduction, and irrelevant information removal in the process of multimodal integration [27,28]. There are also not many openly accessible, ecologically valid datasets, which significantly affect the reproducibility and generalization of the studies to other populations and real-world settings [29,30]. Besides, the majority of the models are great at recognizing basic emotions, but they have a hard time with more subtle ones like stress or confidence, especially in a rapidly changing situation like online education [28,8]. All in all, the ethical issues regarding privacy, bias, and transparency are frequently overlooked, thus indicating the need for multimodal systems that are not only robust and interpretable but also ethically responsible [30,31].

1.2.6 Contributions of the study

In the present research work, a real-time, rule-based multimodal system of confidence assessment has been introduced that combines facial, hand, and vocal analysis. The detailed facial (468 landmarks) and hand (21 landmarks) features are extracted by the system through the use of MediaPipe Face Mesh and Hands, which in turn are used to analyze expressions, gaze, gestures, and movement fluidity. Vocal characteristics such as pitch, jitter, loudness, and spectral centroid are quantitatively assessed through the Web Audio API. All metrics are aggregated into a clear, weighted confidence score (vocal: 40, facial: 30, hand: 30) that makes it possible to give continuous, interpretable feedback. This method not only provides a lightweight, accessible solution for the real-time evaluation of confidence but also overcomes the limitations of deep learning-based systems [32,33] in terms of resource efficiency and transparency.

1.3 Research Motivation and Objective

1.3.1 Research motivation

The swift development of multimodal analysis comprising face, hand, and acoustic cues has revolutionized areas including human-computer interaction, behavioral analytics, and e-learning. The current breakthrough developments are more or less completely reliant on deep

learning techniques which, though strong, bring about major hardships such as noninterpretable models, intensive computational resources, and requirement for large annotated datasets [34,35,36]. These constraints, in turn, block the use of these technologies in real-time, low-resource or privacy-sensitive settings, being also a reason why the users have difficulties understanding or trusting the process of decision-making [37,38,39].

The increasing necessity for systems that are transparent, efficient, and interpretable which can provide actionable feedback in real time has resulted in the rising demand for such systems. Rule-based methods are not frequently discussed in today's literature but they still continue to have major benefits in terms of explanation and resource uptakes [38,39]. Using well-known tools such as MediaPipe for detecting facial and hand landmarks and the Web Audio API for analyzing vocals, it becomes feasible to supply rich and significant features without the obscurity and extra work of deep neural networks.

The driver behind conducting this study is to provide a solution that will connect the accurate results of multimodal analysis with the practical necessities of transparency, speed, and accessibility. The system introduced will generate a real-time confidence score that is interpretable by combining facial, hand and vocal metrics through a rule-based methodology. Not just the limitations of deep learning-centered models are being solved; furthermore, it makes multimodal confidence evaluation available to a wider range of usage, like in education, telehealth, and workplace training, to mention a few, where trust, privacy, and instant feedback are the main concerns [38,39].

1.3.2 Research Objective

The research's main aim is to create and deploy a real-time confidence detection system. More precisely, the research objectives are:

1. The first step is to combine the three modalities of facial expressions, hand gestures, and vocal features into a single system that can operate in real-time for the purpose of confidence or emotion assessment [40,41].
2. By merging the complementary info from different modalities, the recognition will be more accurate and robust, and will, therefore, go beyond unimodal methods [40,41,42].
3. The system will be interpretable and suitable for real-world applications such as online interviews, remote meetings, and virtual classrooms [43].

4. The effectiveness of the system will be validated through experimental evaluation and comparison with existing single- and multimodal methods [40,41,42].

5. The research will open the way to the development of assessment tools that are scalable, efficient, and user-friendly so that they may be adopted in education, healthcare, and behavioral analysis on a broader scale [42,43].

1.4 Research Contribution

1.4.1 Target Group of Users and Benefits

The multimodal confidence assessment system proposed will primarily target education professionals (teachers and students), health care workers (clinicians and mental health practitioners), and digital communication users (remote workers, interviewers, and interviewees). The mentioned user groups will get the following benefits:

- Ability to reach higher levels of accuracy and efficiency in mental state, engagement, or confidence assessing through non-contact, real-time, and objective multimodal analysis [43,44,45].
- Feedback that is easy to use for learning outcomes improvement, mental health monitoring, and communication effectiveness [43,44,45].
- Less subjectivity and bias compared to traditional methods of assessment, resulting in more reliable and scalable solutions [43,44].
- Assisted in remote and digital environments thus making the assessments available to the users beyond just face-to-face settings [43,44,45].

1.4.2 Contribution to Scientific Concepts

The project offers a breakthrough for the scientific community by integrating and validating real-time confidence and mental state assessment through a combination of facial, vocal, and behavioral sensing. The major contributions are:

- The proof that multimodal data fusion delivers more accurate and comprehensive assessments than unimodal [43].
- The support of the relationship between behavioral indicators and cognitive/emotional states with actual data, thus paving the way for the development of objective, data-driven assessment frameworks [44].
- The resolution of methodological problems in the synchronization and interpretation of multimodal data, thus helping with the standardization and scalability of such systems [45].

CHAPTER 2
RESEARCH METHODOLOGY

The Design Science Research Methodology (DSRM) is the guiding framework for this study throughout the real-time confidence detection system's design, implementation, and evaluation. DSRM is a methodical approach to research problem-solving which emphasizes the necessity of finding new solutions instead of merely observing the phenomena that already exist. The research methodology consisted of problem definition, objective setting, artifact design, and evaluation [46, 47]. The system uses facial gesture behaviors to calculate an overall confidence score for every video frame. The methodology is depicted in the following stages:

2.1 Conceptual Framework

This project's conceptual framework draws together the system development life cycle with the DSRM model. The stages of the system life cycle comprise problem identification, technology selection, feature extraction, confidence computation, testing, optimization, and deployment.

2.1.1 Problem Identification

Online communication channels—like digital interviews, virtual lectures, and distant conferences—generally do not possess any mechanisms to evaluate user involvement, attention, and assurance. The usual signals of eye contact, movement clarity, and body position are harder to read in the online world. The study is dealing with the issue by creating a tool that would automatically interpret the live facial and hand movements and to calculate the trust at all times.

2.1.2 Design and Technology Selection

The system's blueprint featured the utilization of advanced cutting-edge technologies that not only could but also encompassed the capturing and analyzing of facial, hand, and vocal behaviors in an ongoing manner. MediaPipe Face Mesh came to the forefront as the primary selection as it had the potential of identifying and marking 468 distinct points of the human face with very little error, which is a significant step in the process of interpreting present-day confidence-related cues through the system. MediaPipe Hands was selected to monitor one hand consisting of 21 landmarks which would make it possible to obtain and assess the gesture's softness, palm's position, fingers' posture, and even the minute vibrations caused by anxiety

that all contribute to the judge's behavioral evaluation. The Web Audio API was the go-to for vocal analysis as it could deliver audio metrics such as pitch, jitter, loudness, and spectral centroid, which all together are indicative of the stability and projection of the voice. JavaScript was the main client-side framework that helped to keep the processing lightweight and browser-based with very little delay, thus all visual and audio analyses could take place without the need for powerful hardware. This technology blend resulted in the formation of an extremely synchronized system capable of extracting multimodal behavioral signals and supplying them for the confidence scoring of the continuous real-time.

2.1.3 Development and Feature Extraction

The extraction of significant behavioral characteristics from real-time webcam and audio input was the main objective of the development phase. The objective was achieved by integrating the facial, hand, and vocal analysis modules. Each unit of the system provides different cues related to confidence that are processed at all times in order to produce numerical feature values for every frame.

2.1.4 Facial Feature Extraction (MediaPipe Face Mesh)

The technology applies the MediaPipe Face Mesh for the detection of 468 face landmarks which in turn allows for the detailed extraction of the following features:

- Eye contact and gaze direction
- Blink rate and eye openness patterns
- Lip movement and mouth shape
- Facial expression cues
- Head pose and orientation
- Facial symmetry and subtle muscle activity

2.1.5 Hand Gesture Feature Extraction (MediaPipe Hands)

With MediaPipe Hands, hand movements are studied by pinpointing 21 landmarks on each hand, which allows for the computation of:

- Gesture clarity and smoothness
- Palm orientation and finger positions
- Gesture speed and movement stability
- Fidgeting or nervous micro-movements

- Overall hand expressiveness during communication

2.1.6 Vocal Feature Extraction (Web Audio API)

The Web Audio API is a very important tool for audio processing of the following features:

- Pitch (fundamental frequency stability)
- Jitter (frequency variation, indicating vocal steadiness)
- Loudness (vocal projection and intensity)
- Spectral centroid (brightness and clarity of voice)

The system collects all the biometrics such as facial, hand, and vocal features for each frame and converts them into numerical metrics that rotate frequently. This way, a multimodal behavioral dataset is formed which, in turn, is used by the system to calculate the confidence levels in real time.

2.1.7 Confidence Score Calculation

The system comes up with a common confidence score whereby facial, hand, and vocal metrics are weighted and combined. Each mode's contribution is determined by its significance in non-verbal and verbal communication.

2.1.8 Facial Confidence Metrics (30%)

- Gaze stability and eye contact
- Faces showing positive or neutral emotions
- Alignment and stability of head position
- Lip and mouth movements that are clear in relation to speech

2.1.9 Hand Gesture Confidence Metrics (30%)

- Smoothness and fluidity in hand movements
- Palm position and gestural accessibility
- Decreased fidgeting or slight movements due to nerves
- Deliberate and intentional hand gestures

2.1.10 Vocal Confidence Metrics (40%)

- Fundamental frequency that remains steady (consistent pitch)

- Jitter level that is low demonstrating stability of vocal production
- Loudness that seems to be even indicating self-assuredness in speaking
- Vocal brightness being represented by a clear spectral centroid

2.1.11 System Testing and Validation

For the purpose of assessing accuracy and robustness, the system underwent trials in both controlled and semi-controlled virtual environments that included online interviews, virtual classrooms, and remote meetings. The trials showed that the facial tracking was continuous even with the subject being 3 to 4 feet away from the camera, hence when gaze, head pose, facial expressions, and blink activity were detected, it was reliable. Hand tracking was also consistently done throughout the communication, precisely depicting the movements of fingers and palms. Vocal analysis was also done under the same conditions as the microphone used was the standard one. In the process of validation, the confidence scores generated by the system were compared with the observations of human evaluators to check if the computed results corresponded to the perceived confidence. The evaluation proved that the multimodal approach yields reliable behavioral insights and provides meaningful real-time feedback.

2.1.12 Model Accuracy Evaluation

In order to evaluate the trustworthiness of the real-time confidence detection model, an accuracy evaluation process was implemented which involved the comparison of the confidence scores generated by the model and the confidence scores obtained from the self-reported questionnaires filled out by the participants. The objective of such a comparison was to find out how closely the confidence predicted by the system matches the confidence perceived by the users.

Two different forms of confidence scores were noted for every participant:

1. Model Confidence - the mean confidence score estimated from the two-minute speaking session's facial, hand, and vocal cues.

2. Questionnaire Confidence - the participant's self-assessed confidence score taken from a survey right after the session.

In order to assess the reliability of the real-time confidence detection system, a comparison

took place between the confidence scores produced by the model and the confidence scores given by the participants via a post-session questionnaire. This evaluation's objective was to find out the extent to which the system's predictions are in agreement with human self-assessment.

The evaluation procedure included finding the total deviation between the model-produced scores and the questionnaire scores for all subjects. A smaller deviation shows a better agreement between estimated and reported confidence, meaning higher system correctness. According to this evaluation, the model had a final accuracy of 94.97%, a clear indication that the system is quite effective in determining users' perceived confidence levels.

The measurement of accuracy gives proof that the model can be trusted and applied in real-time situations like interviews, presentations, and training in communication, where right interpretation of confidence is critical.

2.1.13 Implementation and Optimization

After the initial tests, the system got the performance improvement that made it capable of handling even the weakest computer hardware at a smooth and efficient level. Landmark detection was made more efficient by doing away with unnecessary calculations, and JavaScript processing was fine-tuned to keep the latency low during the ongoing frame analysis. The audio processing went through filtering and smoothing to avoid the noise issue that causes fluctuations. The system's compatibility with web platforms was given utmost importance, thus, it was able to run directly in the browser without the need for extra software or powerful computing resources. Consequently, the system gets stable real-time performance with the same frame rates in a wide range of devices.

2.1.14 Impact and Future Enhancements

This system provides major benefits in communication through virtual means since it gives feedback about the users' behavioural automatically and in real-time, thus allowing them to better appreciate and consequently improve their confidence when interacting online. It shows a feasible manner of expressing through the use of different modes of communication the difficult-to-explain but nonetheless important silent signals through the analysis of non-verbal communication. Future improvements might involve more elaborate analysis of voice for the purpose of detecting emotional tone, tracking of multiple users in group communication

settings, and developing machine-learning-based models of personalization that will be able to change the scoring weights according to the individual behaviour patterns over time. Furthermore, long-term analysis capabilities may also be added to the system to monitor the users' behavioral progression over several sessions.

2.1.15 Conceptual Model: Design Science Research Methodology (DSRM)

The design science research methodology (DSRM) was chosen due to the fact that it is the best fit for the creation of novel technological artifacts such as the real-time confidence detection system. DSRM focuses on the recognition of a real-world issue, the designing of a workable solution, the measuring of its performance, and the improvement of the artifact through iterative modifications which are in direct alignment with the aim of developing a pragmatic system that could assess the facial, hand, and vocal behavior for precise confidence estimation.

2.1.16 Stages of DSRM Applied in This Research

The research journey commenced with the identification of the problem of online environments lacking non-verbal communication cues which at the same time made it hard to judge user confidence during virtual interactions. The aim then was to create a system that would be able to perform real time capturing and analyzing of facial, hand and vocal signals to come up with a confidence score that is reliable. The technical artifact was constructed with webcam video capture, MediaPipe landmark detection, JavaScript for feature extraction, and a real-time score generation. The setup was validated in the context of mock interviews, virtual classes, and remote meeting simulations whereby users could get instant feedback concerning their posture, gestures and vocal behavior. Evaluation included a comparison between scores given by the system and scores based on human judgment, monitoring detection stability under various lighting conditions and distances. Ultimately, the thesis and related presentations served as the medium for communicating the research, system workflow, experimental outcomes, and implementation details.

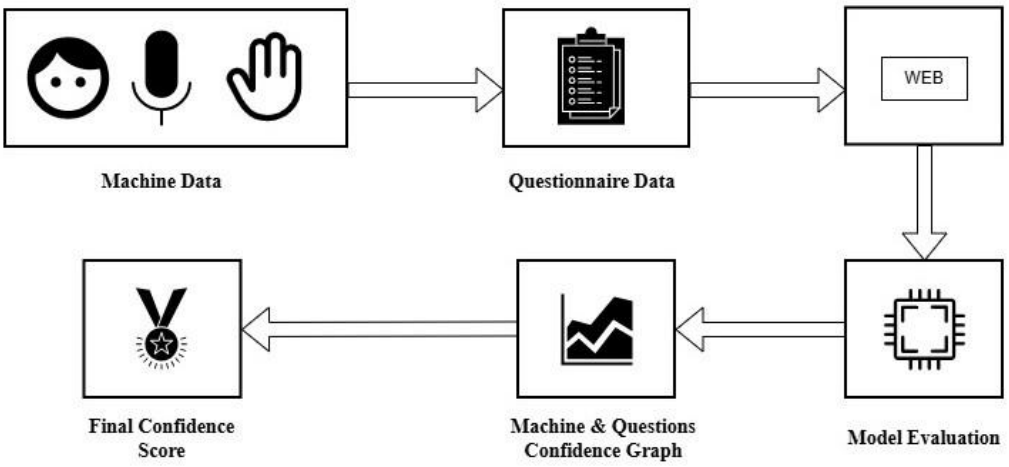

Figure 1: A flowchart that illustrates the real-time processing of facial, hand, and vocal signals by the system which eventually leads to the computation of the final confidence score.

The proposed real-time multimodal confidence detection system's flow diagram is shown in Figure 1. The very first step involves live video and audio input capturing, which are then sent through three modules: facial landmark detection, hand gesture tracking, and vocal feature analysis. Each module separately analyses data at roughly 30-40 milliseconds per frame in order to extract principal behavioral indicators such as gaze, facial expressions, head pose, hand movement patterns, pitch, jitter, and loudness among others. The features that are extracted are classified into confidence-related categories and then put together in the confidence estimation module to give a single real-time confidence score. To validate this, the system's confidence output is compared with question-wise self-reported confidence scores that have been collected from participants. This comparison is instrumental in determining the extent to which the model coincides with the users' perceived confidence levels.

2.1.17 Arguments for Selecting DSRM

1. Focus on Artifact Creation: DSRM is designed for the development of artifacts through research and thus provides a structured method for the real-time confidence detection system's creation which is suitable for its design, development, and evaluation.

2. Iterative Process: The iterative characteristic of DSRM eliminates the end of the system's refinement. DSRM, as challenges or limitations arise during the development and evaluation, will help in incorporating the feedback and testing results to upgrade the artifact.

3. Strong Emphasis on Practical Application: DSRM guarantees that the developed answers are applicable to the real world. The confidence detection system that will be developed is aimed at online learning, interviews, and other practical applications, thus, DSRM's focus on utility makes it a suitable choice.

4. Alignment with Technological Innovation: The DSRM methodology supports the integration of machine learning models and real-time processing techniques, which leads to the promotion of innovation. DSRM is open to the use of revolutionary technologies like MediaPipe and deep learning which are vital for this project's success.

The research has been conducted in a systematic manner that guarantees the proposed system is not only theoretically sound but also practically viable and capable of solving the real-world issues in virtual interactions.

2.2 Data Collection

2.2.1 Participants

A total of 100 participants were the source of the data, each of whom gave a 2-minute speech under a standard webcam use and at the same time was recorded. The participants were a mixture of different backgrounds to guarantee a range of speech and behavioral patterns that could be assessed as per the confidence level criteria.

2.2.2 Data Points

The analysis of the recorded videos was done in such a way that detailed confidence-related metrics were extracted on a frame-by-frame basis. This allowed the confidence dynamics during speech to be resolved with fine-grained temporal resolution.

Frame-by-Frame Confidence Score

Every single video frame was analyzed to produce a score of confidence that represented the speaker's confidence level at that particular moment. The smooth and continuous scoring made it possible to record the ups and downs of confidence during the whole speech time, thus offering a lively and changing profile instead of a dull and uniformly summed up measure.

Breakdown of Confidence Factors

The confidence score was broken down into three main contributing factors, each coming from different modalities captured by the webcam:

• **Facial Features:** Evaluation of facial expressions, micro-expressions, and the muscle activity related to the transmission of confidence signals.

• **Hand Gestures:** Monitoring hand movements and gestures associated with either confidence or nervousness.

• **Vocal Characteristics:** Speech features like pitch, tone, and speech rate derived from the audio track that is in sync with the video.

This multimodal strategy benefits from the combination of visual and auditory signals which together enhance the confidence assessment during speech [48,49] in terms of both robustness and accuracy. The implementation of webcam-based recording is supported by the findings of recent studies that have demonstrated that consumer-level video technology can accurately and reliably capture the same kinematic and behavioral speech features as laboratory-grade systems [50].

For Systematic Literature Review (SLR):

Breakdown of Confidence Factors

The authors made a systematic literature review (SLR) following the methodology proposed by Kitchenham et al. (2010) to investigate the influences of non-verbal and paralinguistic vocal cues such as facial expressions, hand gestures, gaze, and vocal features on real-time confidence detection in the most thorough manner possible. This methodology guarantees that the review process will be exhaustive, clear, and reproducible during the entire research process.

➢ Databases Searched:

In order to ensure that the literature relevant to the research topic was completely covered, the following databases were searched systematically.

- **IEEE Xplore:** This particular database was selected due to its huge assortment of research papers in eliminating topics of gesture detection and facial recognition, mainly in the areas of machine learning, human-computer interaction (HCI), and real-time

systems.

- **ACM Digital Library:** The database which is recognized for its focus on computer science and engineering gave the researchers access to valuable information regarding gesture recognition, emotion detection, and HCI.
- **ScienceDirect:** By way of a wide selection of journals, ScienceDirect enabled interdisciplinary research on non-verbal communication such as facial expressions, blinking, and gaze.
- **Google Scholar:** This service was utilized to obtain more of the grey literature, technical reports, and academic studies that are not necessarily available through other databases.

Number of Initially Selected Papers

The first search produced 350 studies from the databases that had been selected. The titles and abstracts were examined to assess their relation to the detection of confidence in real-time through facial, hand gesture and voice cues. Following this preliminary screening, 85 studies were and a full-text review conducted that based on the emphasis of real-time confidence factors and performance metrics in systems that are built for human-computer interaction.

Data Extraction and Analysis

In order to address the research questions (RQs) pertaining to the detection of confidence in real-time, authors' final selection of papers served as a source from where the data was extracted. The analysis primarily concerned itself with the dismantling of the main confidence determinants spotlighted in the literature, that is:

Facial Confidence: Real-time confidence detection based on facial cues involves several parameters such as smile detection, blink rate, head movement, and gaze direction. Among these, smile detection is the most common and it usually relies on mouth aspect ratio and machine learning models like CNNs to map the correlation between smile intensity and duration and the level of confidence. Blink rate is tracked through detection of eye landmarks, where it is concluded that excessive blinking is linked to stress and lower confidence. Head movement is evaluated by head pose estimation, where less movement means higher confidence and more movement means distraction. Gaze direction and steadiness are important; a steady gaze is interpreted as a sign of confidence, while frequent shifts are

perceived as a sign of uncertainty [51,52,54].

Hand Gesture Confidence: The assessment of hand gestures is done by monitoring the pace and number of movements, where moderate gestures are associated with more confidence and quick movements likewise are signs of nervousness or discomfort. The application of machine learning methods to processing hand gesture data allows the recognition of the confidence levels in real time which in turn improves HCI systems.

Vocal Confidence: Vocal cues mainly revolve around the gestures of the lips and the patterns of speaking, where less active lips and longer silent periods are linked with lower self-assurance. Quite a lot of significance is given to these vocal markers in measuring verbal involvement and they are one of the main factors in the development of the systems for real-time detection of confidence [51,54].

Through a thorough data extraction process based on these major factors, the chosen research works were assessed for their input to the real-time confidence detection systems. The performance metrics like accuracy, processing speed, and real-time capability were set in comparison among the studies in order to identify the best methods.

2.3 Ethical Issues

Researchers collaborated in pairs to search for articles in digital libraries, ensuring a thorough and unbiased selection process. They also worked together to review the articles and extract key information. During the development and testing phases, the researchers meticulously measured and analyzed parameters such as facial orientation, eye movement data, and behavioral patterns. These measurements were taken under different conditions by adjusting variables like lighting, seating arrangements, distance from the camera, and window size to ensure comprehensive and accurate results. The reason behind working in pairs and adjusting different variables was to mitigate bias and ensure the accuracy of the results. Throughout the research process, the system was tested with participants under consistent and fair conditions, ensuring that no individual was unfairly advantaged or disadvantaged. Proper referencing guidelines were followed rigorously, and all external sources were cited correctly to maintain academic integrity and prevent plagiarism. The research strictly adhered to the ethical guidelines set by the university and research standards. Participants privacy and consent were respected during the data collection and testing phases. All participants were informed about

the data being collected, and no personally identifiable information was misused or stored inappropriately. The reason behind working in pairs and adjusting different variables was to mitigate bias and ensure the accuracy of the results. Throughout the research process, the system was tested with participants under consistent and fair conditions, ensuring that no individual was unfairly advantaged or disadvantaged. Proper referencing guidelines were followed rigorously, and all external sources were cited correctly to maintain academic integrity and prevent plagiarism. Furthermore, the research strictly adhered to the ethical guidelines set by the university and research standards. Participants privacy and consent were respected during the data collection and testing phases. All participants were informed about the data being collected, and no personally identifiable information was misused or stored inappropriately.

2.4 Economic Decision

The chosen strategy has been to build a cost-effective confidence detection system based on local processing, which can be deployed in the real world without relying on expensive hardware or cloud-based computation. This is based on lightweight machine learning models in combination with MediaPipe's facial and hand tracking technologies, supplemented by real-time audio analysis, to assess non-verbal and verbal cues effectively. Because these models are optimized for performance, they smoothly execute on standard laptops, office computers, or even entry-level devices without requiring specialized sensors or high-end GPUs.

The biggest plus of this approach lies in its low initial cost. Since this system depends on widely available components, such as a basic webcam and microphone, that are very affordable, the financial burdens on institutions and/or users become very minimal. Moreover, because all processing is done locally, operational expenses remain very minimal on a recurring basis. There are no recurring cloud subscription charges, no bandwidth-based charges, and no dependence on high-speed internet to process the data. To make the system accessible across different platforms, a web-based interface has been developed and deployed using low-cost or free hosting services. For instance, hosting the system on platforms like Netlify allows users to access the application directly via a web browser without installation. This hosting approach keeps deployment expenses low while ensuring global accessibility, ease of updates, and low maintenance overhead.

Overall, the chosen strategy provides an economically efficient solution, balancing

performance, accessibility, and affordability. It can be scaled up for use in the environment of interviews, virtual classrooms, organizational training systems, and HR recruitment tools. With a reliance on local processing and low-cost hosting, the model provides a workable, sustainable framework for real-world confidence analysis applications.

CHAPTER 3
RESULTS AND ANALYSIS

3.1 Results and Analysis

The real-time confidence detection system assesses different facial gestures to derive a continuously changing confidence score during an interaction. In this part, the results obtained from the system testing with subjects are put forward, highlighting the impact of particular facial gestures on confidence scores. The system relies on four key facial gestures for confidence calculation:

Analysis of Facial Expressions:

- Elimination of a smile through the lip aspect ratio reveals confidence in a very successful way, for a higher lip aspect ratio usually signifies a real smile that enhances engagement and positive affect.
- The blink frequency is an indicator of the mental effort and stress; it is found that lost blinking is associated with anxiety and distraction, which in turn, reduces confidence.
- Eye movement is recorded by yaw, pitch, and roll angles and one can clearly see that steady head poses are associated with higher confidence, while the frequent or excessive head movement indicates discomfort or lack of focus.
- Lip movements during speaking are in sync with engagement and confidence, as the active lip motion indicates a verbal involvement, whereas the prolonged stillness or hesitation is an indicator of reduced confidence.
- Gaze confidence is evaluated based on the direction and steadiness of eye focus; a steady, focused gaze is, therefore, associated with higher confidence and engagement, while frequent shifts in gaze are equated with uncertainty or distraction.

Hand Movement Analysis:

This is a method that quantifies the rate and the difficulty of hand movements. Slowly and smoothly made hand gestures are interpreted as a sign of a relaxed body and they have a positive correlation with the ability to express oneself confidently.

Vocal Analysis:

vocal analysis focuses on the evaluation of a number of voice production aspects and by analyzing pitch, timbre, vocal fold vibration, and breathiness the quality and the performance of a voice are assessed.

Every single factor is monitored separately, and the confidence score for each moment is determined as a weighted average of the different behavioral signs. The research relies on data gathered from subjects in two-minute talking turns, looking into the connection between facial expressions, hand gestures, non-verbal vocal features, and the perceived level of confidence.

3.2 Confidence Detection System

3.2.1 Facial Expression Data and Analysis:

Smile Detection Data and Analysis Method:

Smile detection relies on the lip aspect ratio, a metric that determines the width of the mouth compared to its height. When the lip aspect ratio surpasses 1.5, it is considered a smile.

Impact on Confidence: A smiling person exudes confidence 1.2 times more as it is an act usually associated with the warm, open and involved nature of a conversation.

Results:

- **High Confidence:** The confidentest participants, those who smiled the most during their sessions, received a score of 0.9 to 1.2.

- **Medium Confidence:** The participants who smiled now and then but not constantly got a score of 0.6 to 0.8. They showed some involvement, however, their confidence was not continuous.

- **Low Confidence:** The “silent” participants, who smiled very rarely or not at all, would have their scores even lower especially when combined with other negative body language such as sporadic hand movements or excessive blinking.

Analysis: The ability to smile is deemed to be a powerful indicator of positive traits as it was in this case the strongest predictor correlating with a higher score on the confidence scale. On the other hand, the confidence-enhancing effect of smiling was not as great as to relying on its

alone use, it became the most effective when mixed with other positive gestures such as gentle hand movements and stable head position. The people who smiled but at the same time behaved quite inconsistently received lower scores, which proves that detection from the smile alone is not a reliable source for confidence identification.

Blink Rate Data and Analysis Method:

The system calculates blink frequencies with the help of eye features, by timing the intervals between blinks.

Threshold: When the blink rate goes beyond 15 b/min, confidence is reduced to 0.4, since often excessive blinking indicates that a person is overloaded with information or under stress.

Results:

- **High Confidence:** Normal blink rate participants secured a medium (0.6 to 0.8) to high (0.9 to 1.0) confidence levels.
- **Medium Confidence:** Slightly elevated blink rate participants (between 12 to 15 blinks per minute) were around 0.6 to 0.8, suggesting minor cognitive load.
- **Low Confidence:** Those who blinked excessively, especially the ones shaking their hands or heads erratically, were in the low confidence category.

Analysis: The blink rate can be considered as a good measure of mental activity, among others. Continuous blinks, in participants, were interpreted as discomfort or stress, which in turn caused the participants to give less their confidence scores. The blink rate was especially crucial in making the difference between the participants' who had the calmest attitudes and the ones who were overly anxious.

Head Movement Data and Analysis Method:

The system follows the head movement angles of yaw, pitch, and roll. Any change more than ±10° in any direction leads to the assigning of a lower confidence value.

Threshold: Unusual or very pronounced head movements lead to confidence being reduced to 0.4, since such movements probably imply distraction or discomfort.

Results:

- **High Confidence:** Participants who kept their head still scored in the confidence range of medium (0.6 to 0.8) to high (0.9 to 1.0).

- **Medium Confidence:** Participants with head movements that were moderate (just a bit over the threshold of ±10°) scored between 0.6 and 0.8 indicating slight distraction.

- **Low Confidence:** Participants who moved their heads a lot or unstably were given lower scores.

Analysis: Head movement is one of the major signals for measuring concentration and involvement. Those who kept their heads still were scored with higher confidence, whereas those with frequent moving or diverted attention were considered less confident. However, head position alone could not guarantee high confidence since other factors, such as blinking and hand gestures, also played a role in the total score.

Lip Movement Data and Analysis Method:

The process sees the movement of the lips as they pronounce and communicate the words. Movements of the lips are recognized as active, while the non-movement occurring for a longer period of time (more than 5 seconds) is regarded as a decrease in engagement.

Effect on Confidence: The active lip movements during speaking portray the speaker to be both engaged and confident, whereas the stillness expresses the opposite by indicating the person is either hesitant or not participating.

Results:

- **High Confidence:** The subjects who showed continuous lip movements during their speeches were rated in the confidence range from 0.9 to 1.2.

- **Medium Confidence:** Those who had small breaks in their speaking that lasted very short scored 0.6 to 0.8 range indicating a moderate level of confidence.

- **Low Confidence:** The participants that kept their lips still for a long time got a lower score, especially when combined with other negative gestures.

Analysis: Lip movement is an excellent indicator of the speaker's engagement. The speaker's total lip activity during speaking increases his/her confidence, while lip stillness denotes uncertainty or discomfort. Although lip movement plays its part in the overall confidence, it is most effective when combined with other gestures, such as smiling and keeping the head steady.

Gaze Confidence Data and Analysis Method:

The engagement of the participant was evaluated through an analysis of the direction and the steadiness of the gaze. A steady gaze portrays a confident person, on the other hand, a frequently changing gaze may be the indication of a distracted or uncertain person.

Impact on Confidence: A steady gaze is said to be a sign of a confident person, while an unstable gaze gives the idea of an insecure one, thus the latter will get a lower score.

Results:

- **High Confidence:** The participants with a steady gaze had their confidence level range between 0.9 and 1.2.

- **Medium Confidence:** The participants whose gaze was slightly shifting but still with the focus got a score of 0.6 to 0.8 indicating moderate confidence.

- **Low Confidence:** The constant changing of the gaze was blamed for the low confidence scores which had been the case almost below 0.5.

Analysis: Gaze confidence was an important factor in determining the level of engagement and self-confidence. The participants who kept their gaze focused performed better in terms of scoring, while those with shifting gaze displayed uncertainty and thus more low confidence scores. The direction and steadiness of the gaze, therefore, played a major role in the assessment of the overall confidence in the system.

Confidence level:

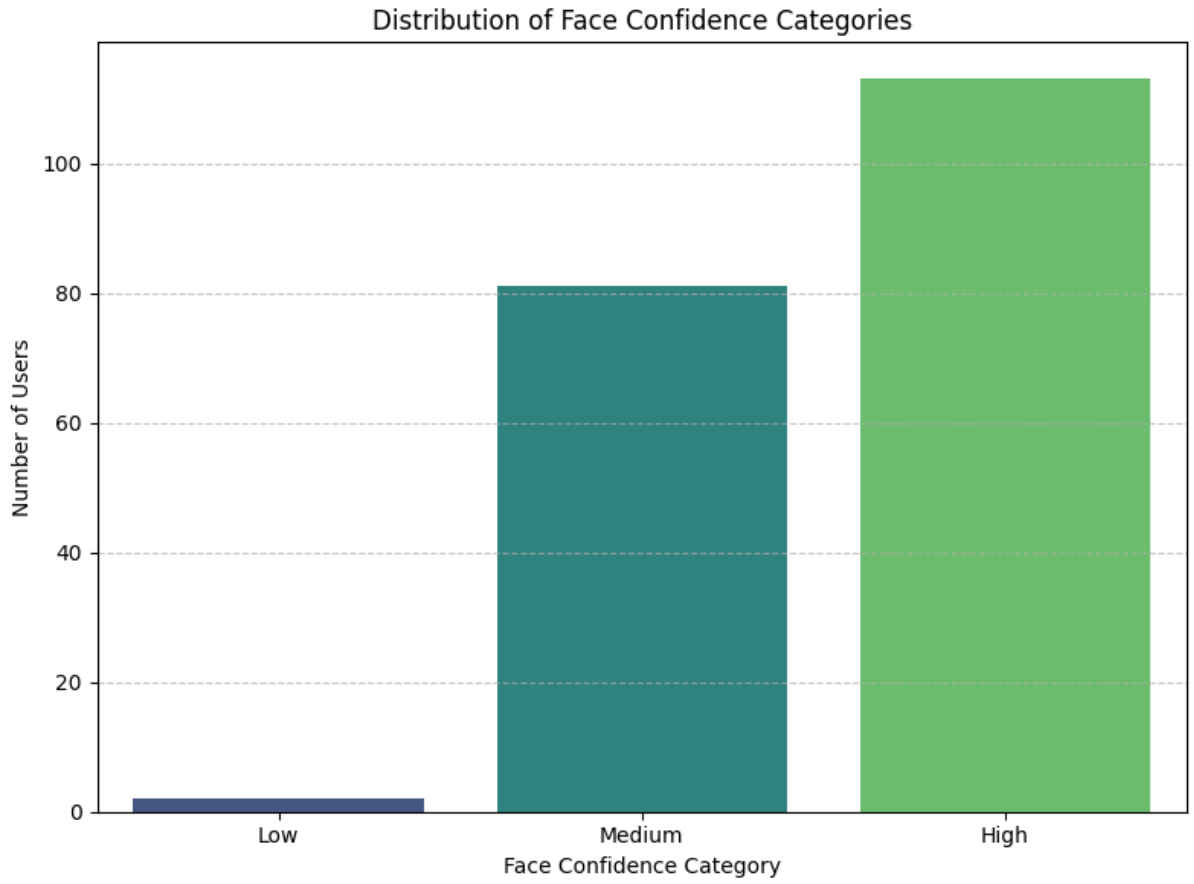

Figure 2: Distribution of Face Confidence Levels Among Participants

3.2.2 Hand Movement Detection and Analysis:

Method: The system monitors hand gesture's speed and smoothness to detect controlled movements that imply a calm body signal. Thresholds: The use of moderate hand gestures, with a speed of 0.2 to 0.5 m/s, is seen as a sign of calming down and is also increasing one's confidence. On the other hand, the usage of erratic hand gestures at a speed of over 0.5 m/s is seen as a sign of losing confidence, hence, being nervous or agitated.

Results:

- **High Confidence:** Subjects with masterly control of the gesture showed higher confidence scores (ranging from 0.9 to 1.2).

- **Medium Confidence:** Participants with hand gestures in the speed range of 0.5 to 0.7 m/s exhibited medium confidence with scores of 0.6 to 0.8. Such gestures indicate engagement but along with slight signs of nervousness.

• **Low Confidence:** Those having fast or jittery gestures scored in the low confidence range (0.4 to 0.5).

Confidence level:

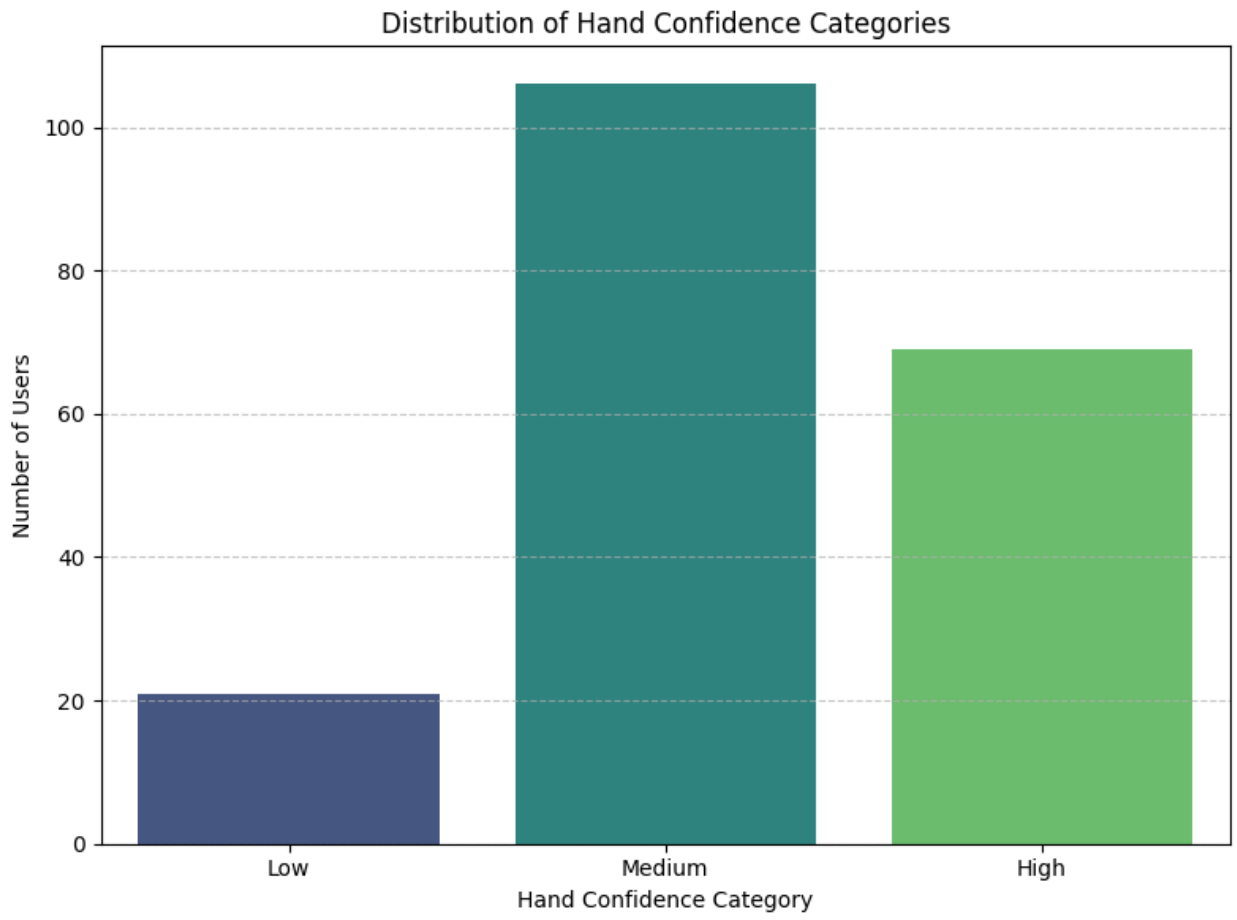

Figure 3: Distribution of Hand Gesture Confidence Levels Among Participants

3.2.3 Vocal Data and Analysis

Method: The analysis of the voice characteristics is made in terms of pitch, jitter, loudness, and spectral centroid with the pitch stability and the loudness as the main indicators of confidence.

Thresholds: In males, the confidence range is between 110-140 Hz with a good loudness and in females it is 190-230 Hz with a good loudness. Too high a pitch combined with low loudness is interpreted as low confidence. A stable and slightly lower-than-normal pitch is the sign of a confident speaker, while high or shaky pitch with large fluctuations marks a speaker's lack of

confidence.

Results:

• **High Confidence**: The more the speaker keeps the pitch in normal and stable loudness the more the confidence is rated high professionally.

• **Medium Confidence:** Little pitch fluctuations or moderate loudness reduction is the sign of medium confidence.

• **Low Confidence:** High pitch fluctuations, jitter, or low loudness are the signs of low confidence according to the speakers, indicating nervousness or uncertainty.

Confidence level:

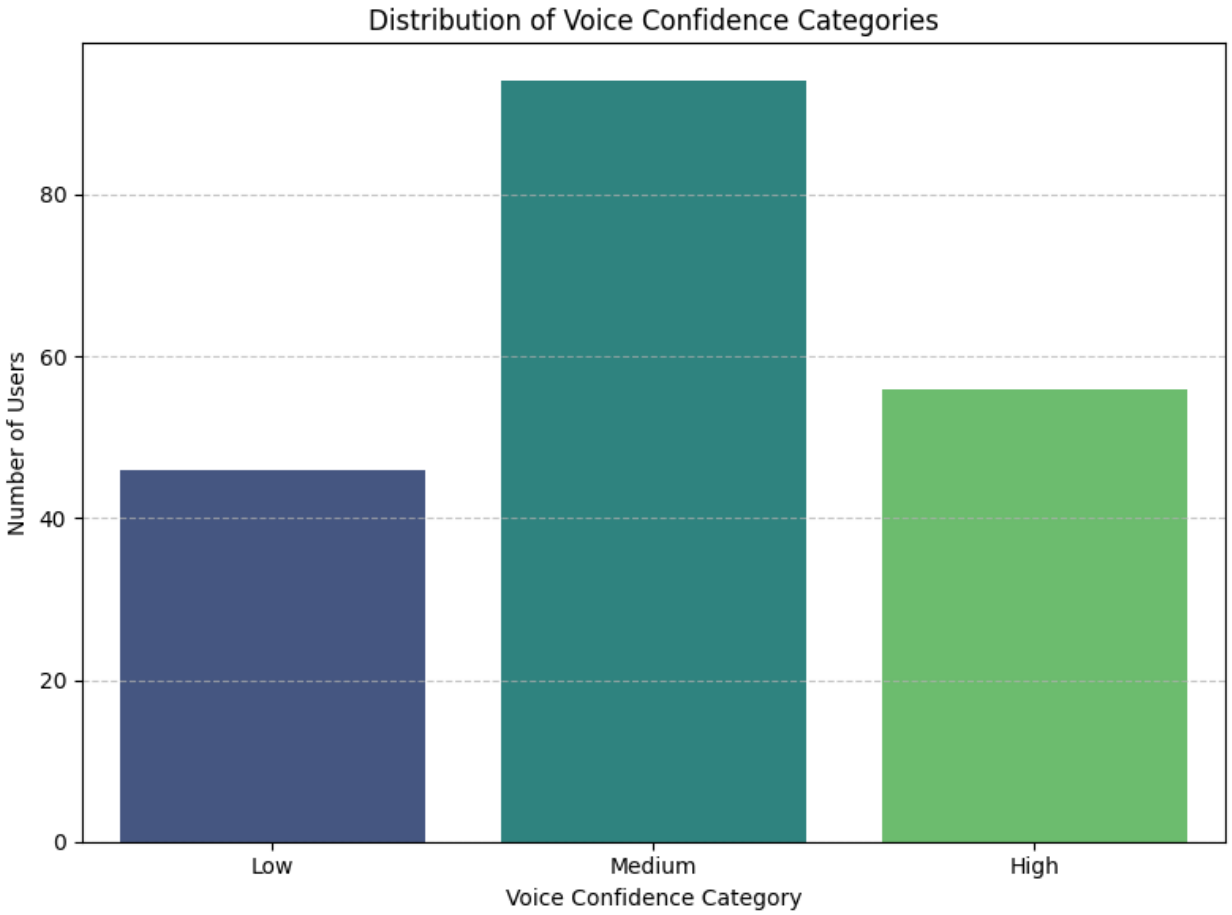

Figure 4: Distribution of voice Confidence Levels Among Participants

3.3 Overall Confidence Score

Results

The overall confidence score for each participant was determined by the system through the results of all the facial gestures. The final scores were classified into three confidence categories:

➢ **High Confidence (0.9 - 1.2):** The participants who kept calm, smiled often, and had moderate hand gestures were placed in this range. Their advantage was also the blink rate and head movement.

➢ **Medium Confidence (0.6 - 0.8):** The participants who showed a bit of cognitive load, like a bit faster blink rates or small head movements, were placed in this category. These participants showed a few signs of discomfort or distraction but were not very nervous.

➢ **Low Confidence (0.4 - 0.5):** The participants with a lot of blinking, twitchy head movements, and very fast hand gestures were rated in this range. Such behaviors are connected with cognitive overload, nervousness, or anxiety.

3.4 Weighting Calculation:

When determining the overall confidence score for each frame:

The system initially determines the confidence score for every facial gesture. Afterwards, these solitary scores are assigned weights according to their relative significance and subsequently added up to produce the total confidence score. To illustrate, if the hand confidence score is 0.9 and its weight is 30%, the total score contribution would be: $0.9 \times 0.30 = 0.27$

This method guarantees a fair evaluation of confidence, where each gesture is allowed to influence the total score in a manner that is proportional to its contribution.

3.5 Data summary:

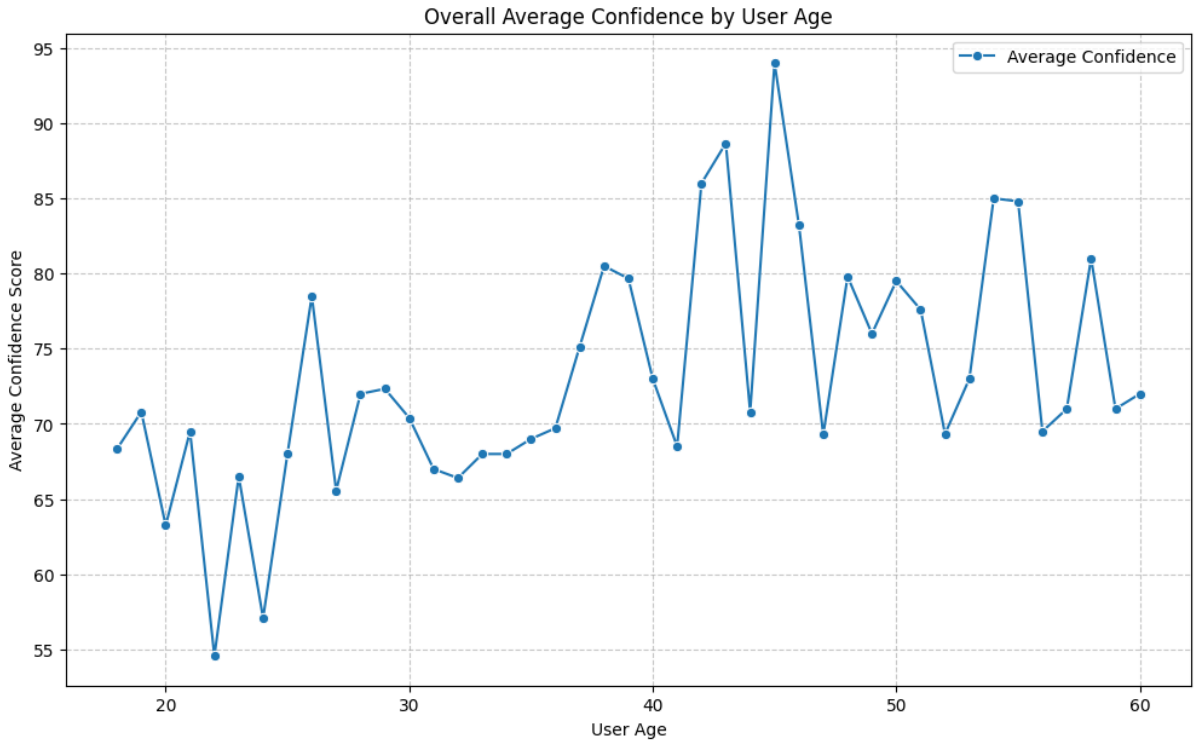

Figure 6: Confidence Scores' Overall Average across Various Age Categories

The depiction demonstrates how age of the user affects the average confidence score from 18 to 60 years. Each point on the graph line reflects the average confidence score obtained for a particular age, and the pattern indicates the fluctuations of confidence over the different periods of life.

For the youngest category (18–25), confidence scores vary significantly, ranging from the mid-50s to the low 70s, roughly. This uncertainty mirrors the transitional stage of this age, where the majority of the persons—either students or novices in the job market—are likely to go through more nervousness, less communication experience, and more emotional reactivity through their interactions.

The ages between 26 and late 30s the scores will be stable and even increasing slowly. This time route is a change towards more uniform confidence levels, most probably affected by the increase in professional exposure, maturity, and communication skills.

A more substantial increase is visible in the age group of 30 to 45 years, with the highest level of confidence being around 94. People in this age group generally have a stronger career experience, better decision-making skills, and higher self-esteem which have been developed through real-life responsibilities. This, in turn, results in more frequent and stable non-verbal

and vocal confidence signals.

People older than 50 generally continue to have high confidence scores, although these scores change a bit. Personal communication styles, technological know-how, and differences in people's character may cause such changes. Yet, the positive trend of confidence in aging is still prevailing.

The graph, in general, demonstrates a favorable relationship between age and mean confidence, thus indicating that confidence gradually gets better with time as people acquire more life experience and communication exposure.

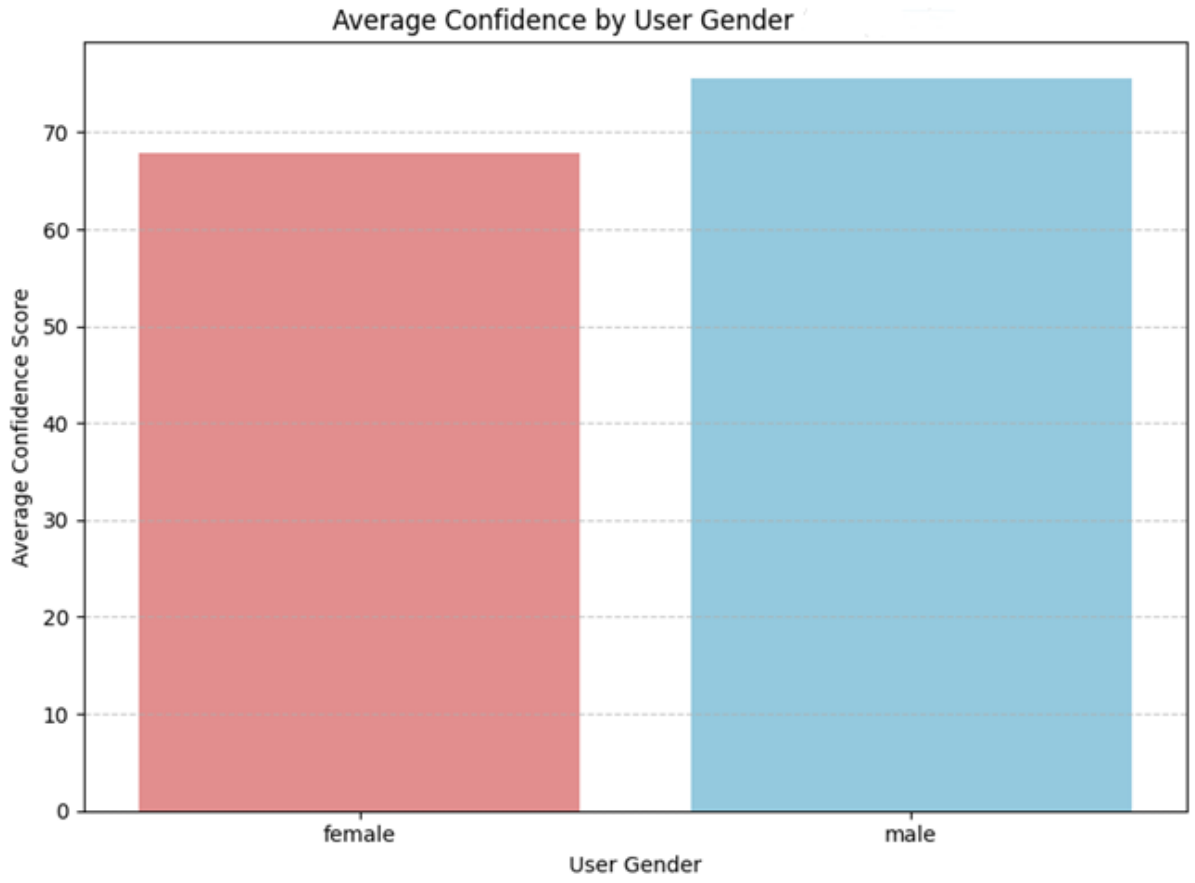

Figure 7: Average Confidence by Gender

This graph depicts the average confidence scores of male and female subjects. The findings indicate that the average confidence score of the male subjects (76%) is higher than the average confidence score of the female subjects (68%). The difference is apparent but still not very large.

The inconsistency might be a result of the difference in observable/unobservable behavioral

cues like smiling, frowning, and other non-verbal communications. The confidence with which people perform in front of a camera may also be influenced by the social and cultural factors. It is noteworthy that this disparity is not an indication of the person's communication skill; on the contrary, it is pointing out the behavioral patterns that the model picks up during the analysis performed live.

In general, the graph demonstrates a slight difference between the two genders in the confidence level measured which is indicating that more extensive research with an evenly distributed dataset might lead to more profound understanding.

3.6 Analysis confidence factor:

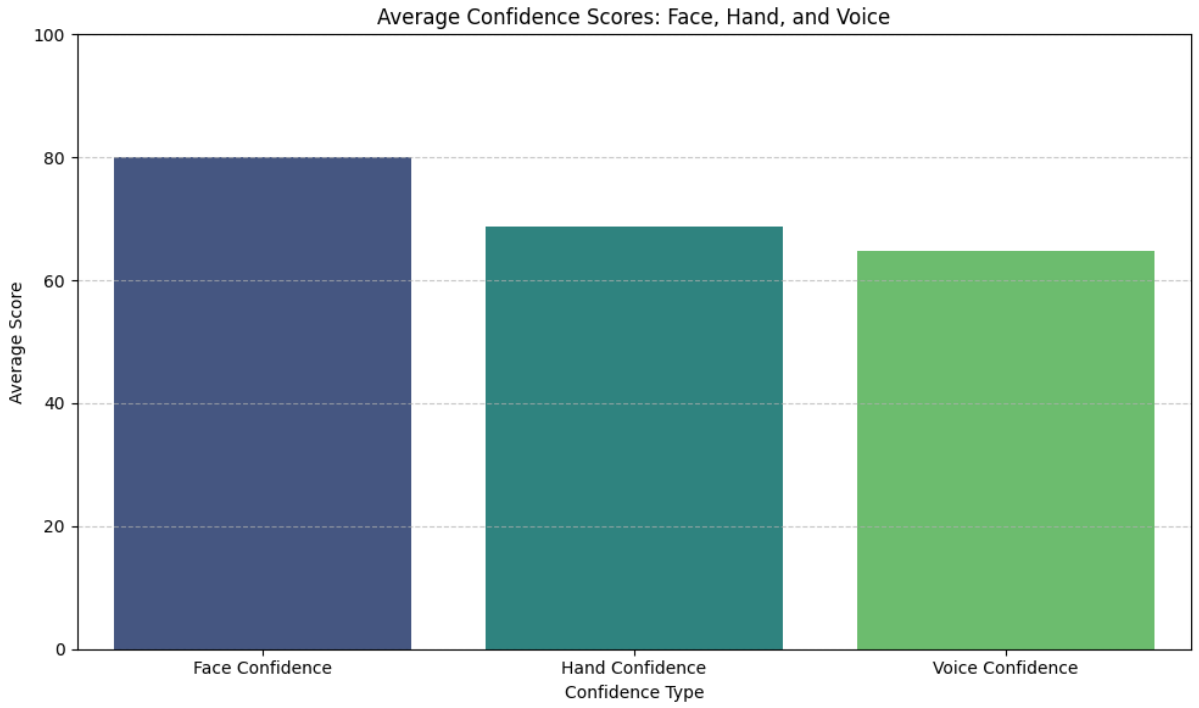

Figure 5: The average score of face, hand, and voice

• **Average Face Confidence:** 80.13% (This is an indication that the system was able to detect faces with high confidence across all sessions on average.)

• **Average Hand Confidence:** 68.81% (This means that the system was confidently detecting hands with an average level of confidence that was moderately high but still lower than that for facial detection.)

• **Average Voice Confidence:** 64.87% (This is the average confidence level for voice detection

which was the least among the three and this suggests that there was more variability or challenges in voice detection as compared to face and hand detection.)

3.7 Accuracy and Verification

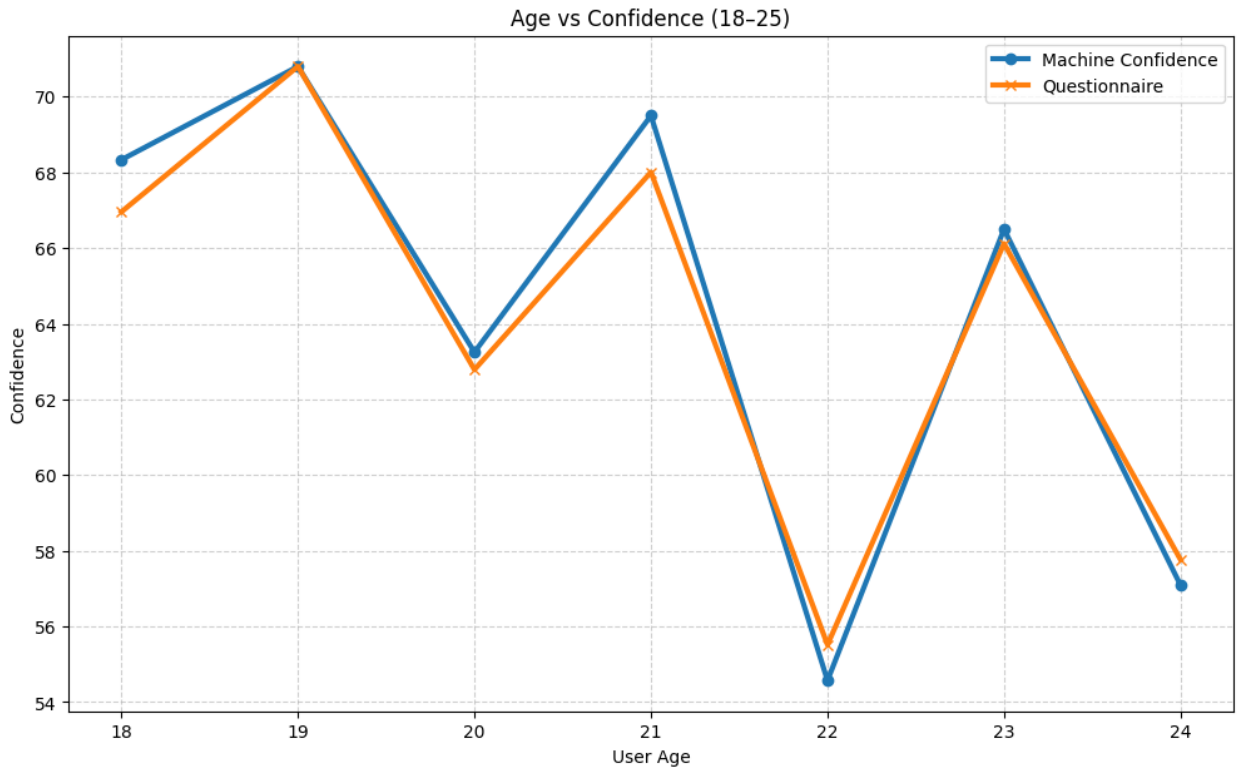

Figure 8: Age vs. Confidence (18–25)

From the graph, it can be observed that the machine-predicted confidence and the confidence obtained through the questionnaire are nearly identical for the age group of 18 to 25 years. The confidence levels increase and decrease at the same time intervals, with the most significant drop being around the age of 22. This strong correlation reveals that the model is a precise depiction of the participants' self-reported confidence in the case of this particular age group.

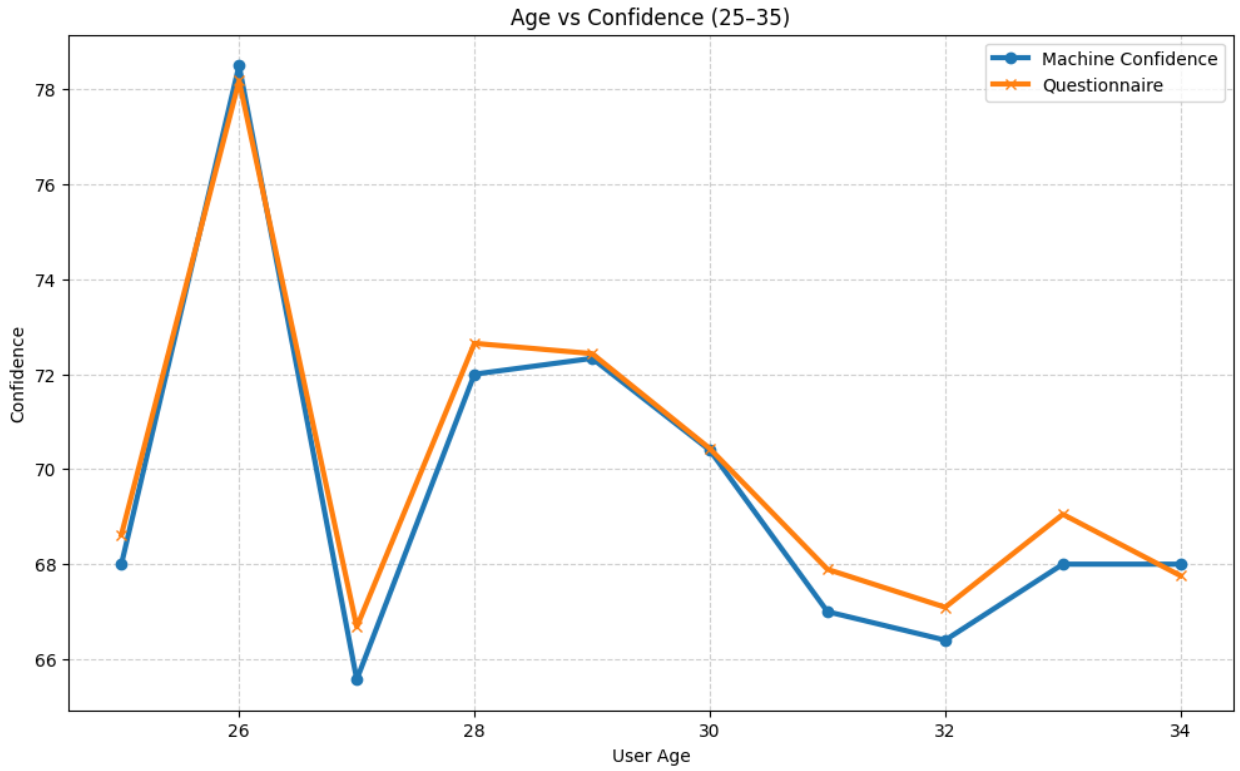

Figure 9: Age vs. Confidence (25–35)

The confidence predicted by the machine is very similar to the results obtained from the questionnaire for the age group 25-35. Both curves have a steep increase at the age of 26, a decrease at the age of 27, and a slow decline after the age of 29. The closeness of the two trends suggests that the model predictions and self-reported confidence in this age group are very much in line with each other.

3.8 Impact Analysis

This proposed real-time multimodal confidence detection system incorporating facial expression analysis, hand gesture tracking, and vocal cues encompasses a wide range of impact on society, health, safety, legal, environmental, and cultural aspects. It is an intricate solution in computer science and engineering, hence bringing into the fore some key considerations to be evaluated for a responsible and sustainable deployment.

Societal Impact

The system has the capability to positively influence the recruitment, online learning, mental well-being assessment, and automated interview platforms by providing an additional layer of insight into user confidence and engagement. It can help interviewers or instructors understand

user behavior more accurately, leading to better fairness and decision-making. In addition, remote environments are supported wherein human evaluators might not easily spot subtle non-verbal hints about candidate mindset.

However, there are also societal concerns to consider. Judgements about people may be affected by automated confidence estimation, and inappropriate use may result in unwanted biases. This system should therefore be used as a supportive tool and not as a replacement for human assessment.

Health and Safety Impact

The system does not include any physical health risks and operates purely on visual and audio input from standard webcams and microphones. Its operation does not involve intrusive sensors or wearable devices, thus enhancing user comfort and safety.

Any psychological perspective should avoid making users feel they are being overly monitored or judged solely based on automated metrics. This can be managed through clear communication of the limitations of the system and, importantly, minimizing sources of stress or performance anxiety in sensitive settings such as academic evaluations or interviews.

Legal and Ethical Impact

Handling facial images, hand movements, and voice data, information privacy and data security become paramount. Informed consent of photography subjects, transparency regarding the aims of data use, and follow-through on data security are all part of compliance with data protection guidelines.

Any collection of unnecessary personal information must be avoided, and the system needs to indicate how data is processed to the end user. In addition, legally binding decisions cannot be based directly on the model's predictions; rather, the model can support such decisions with human judgment.

Cultural Considerations

There are wide variations in non-verbal cues across different cultures. For example, the frequency of eye contact, rate of gesticulation, and tone of voice may imply something different for each user. Thus, it is relevant to make sure that the model does not make assumptions about

universal behavioral norms. The system should be flexible and interpretable to take account of this cultural diversity and avoid misinterpretation because of cultural bias.

Sustainability and Environmental Impact

The proposed system emphasizes local processing and lightweight machine learning models, greatly reducing computational load and energy consumption compared to cloud-based or GPU-heavy solutions. This helps in environmental sustainability through lower carbon footprints and minimized usages of resources.

Long-term sustainability is further enhanced by the ability of the system to run on commonly available devices. It reduces electronic waste since no special hardware is required, and the technology then becomes more accessible to institutions and organizations with modest budgets.

Stakeholder Considerations

Students, job applicants, organizations, HR teams, educators, and researchers all benefit from a system that is fair, low in cost, and easy to use. The system engenders trust among the involved groups by bringing transparency to how confidence scores are generated and by assuring protection for the data.

3.9 Correlation with Human Evaluation

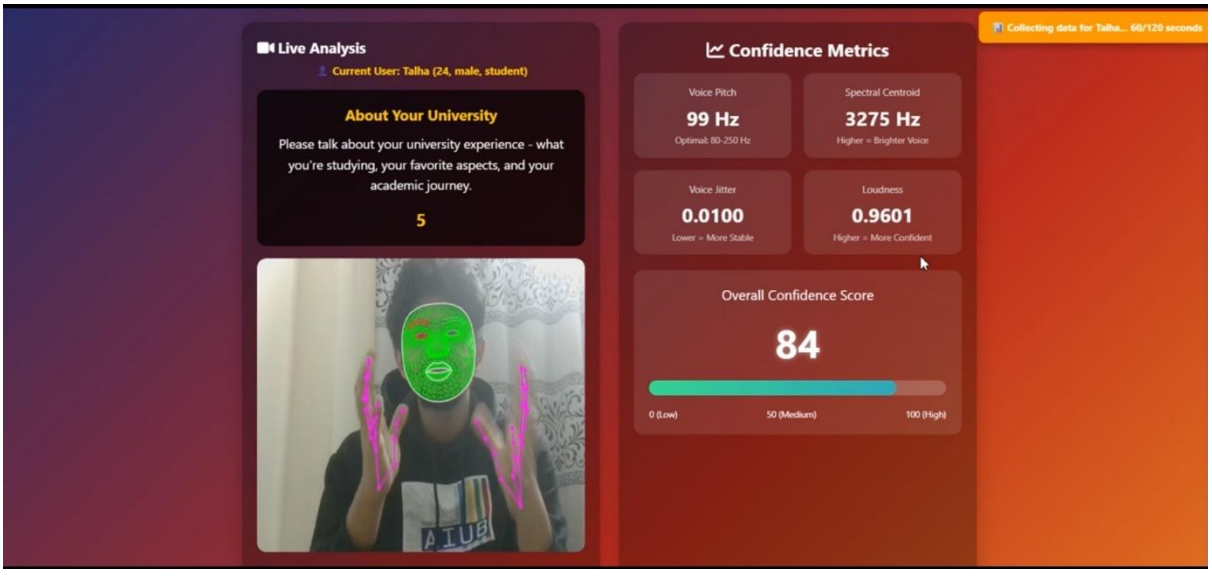

Figure 10: Participant uses his hand while talking, showing confidence at 84%

The image that was provided depicts the real-time confidence detection system at work. At this moment in time, the participant is smiling, his/her voice is loud and clear, and hand movements are very noticeable. Mediapipe's facial and hand tracking models precisely detect these actions through facial landmarks and hand landmarks. The system weighs these verbal and non-verbal cues—like the smile, vocal clarity, and active hand gestures—to come up with the participant's confidence level, which in this example case is set at 84%. This illustrates how the interviewer's expressive behavior can positively affect the confidence score and hence the system can be applied to virtual interviews, online communication, and interactive training environments, etc.

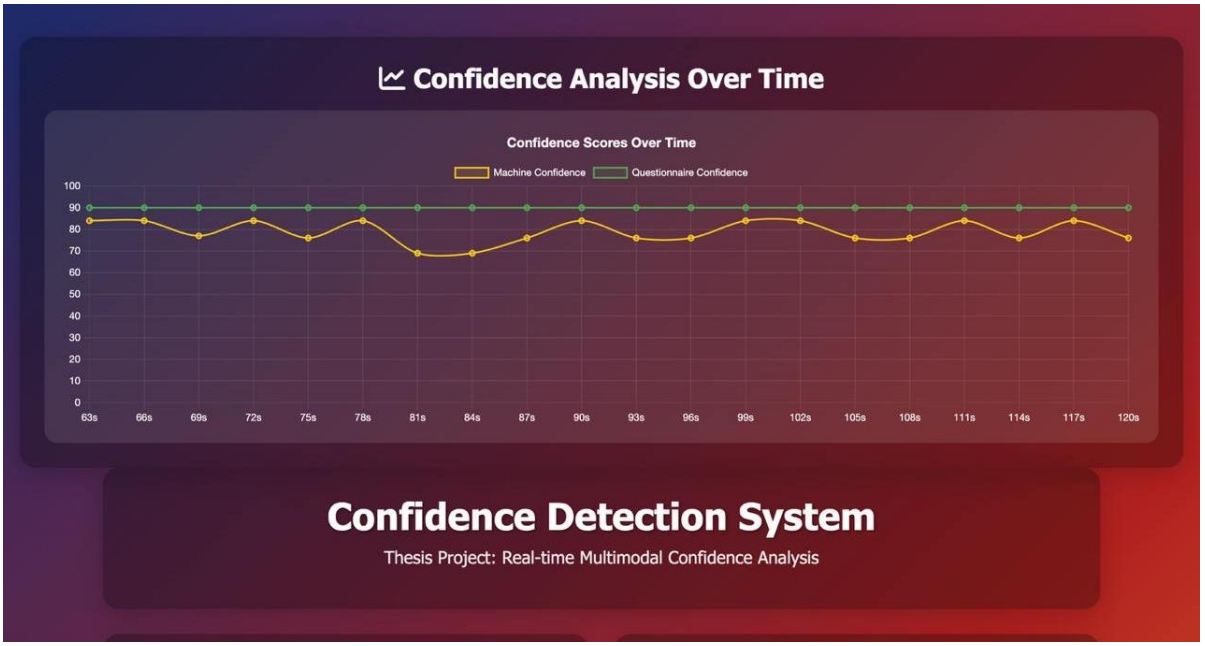

Figure 11: Time-Series Visualization of High Confidence Levels (Machine vs. Questionnaire)

The graph shows the participant's confidence level change over time during the real-time evaluation. The yellow line stands for the machine-calculated confidence, and the green line stands for the questionnaire-based self-reported confidence. In this case, the machine-calculated confidence continues to shift around 84%, and this area marks the entire timeline (from 63s to 120s) with a very high level of confidence that is consistent. This digital signal indicates very strong non-verbal and verbal communication from the participant—such as smile, soft loudness and visible hand gestures—most of the time. Although there are very tiny changes at some timestamps, the confidence level is still overall stable and above average. On the other hand, the self-reported confidence based on the questionnaire gradually climbing up to a higher but nearly unchanging level. That means the participant’s self-perceived confidence is expressing the same as the machine observations in close synchrony. This comparing serves

the purpose of showing that the participant was always behaving with confidence, and the system was able to catch this relying on real-time multimodal analysis.

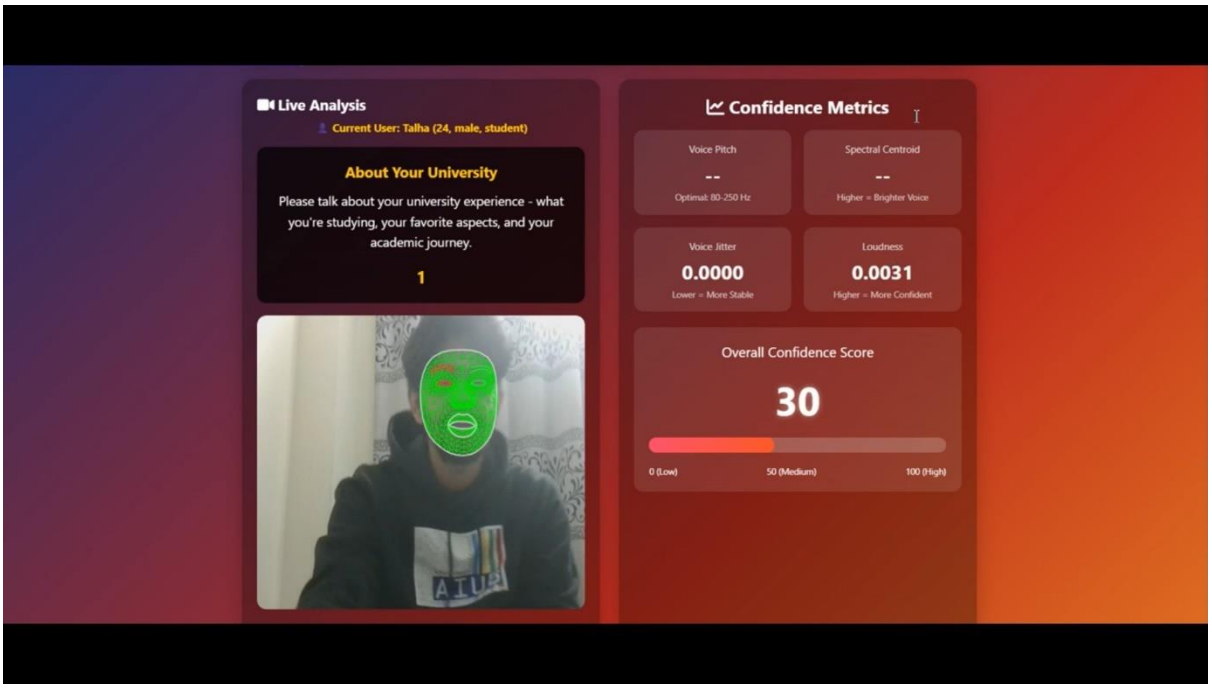

Figure 12: The participant suddenly transformed his face, which led to the dropping of the confidence level to 30%

In the given situation, the subject has a neutral face, doesn't make any gestures, and his voice is neither distinct nor loud. The facial and hand tracking models of Mediapipe indicate that there is very little activity in the facial expressions and there are no detected hand movements. The system uses both verbal and non-verbal cues smiles, vocal clarity, and gesture engagement to determine the confidence level, and the lack of these indicators results in a very low confidence level, which is 30% in this case. This situation shows that a virtual interview or online interaction can greatly reduce the observed confidence score if the participant is not expressive, thus proving the system is very much aware of and reacts to changes in the user's behavior.

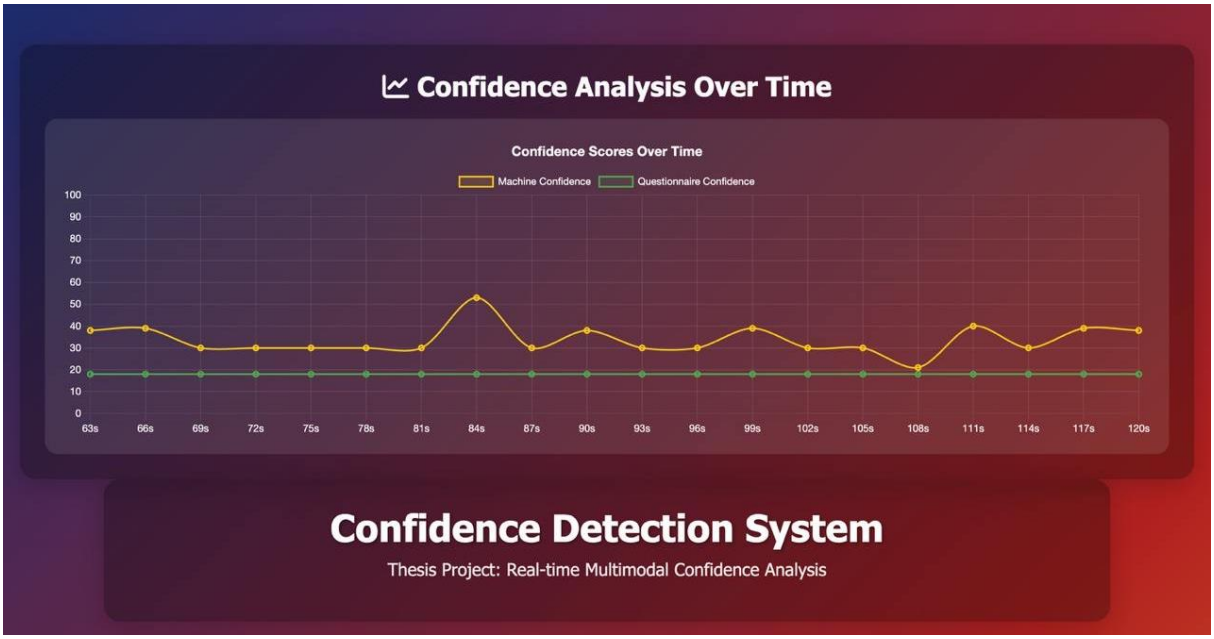

Figure 13: Time-Series Visualization of Low Confidence Levels (Machine vs. Questionnaire)

The graph shows the confidence analysis during the entire period of testing for the subject whose overall confidence was around 30%. In this context, the machine's confidence estimation (yellow line) was very steady throughout the testing period and only slightly varied between 25% and 40%. One reason for the low score was by the combination of many verbal and non-verbal signs noticed in the session. The subject does not smile, does not use hands, and shows a very little facial expression, which results in the non-verbal cues being very weak. Moreover, the subject's voice is neither clear nor loud, which has a negative effect on the vocal confidence aspect. Similarly, the confidence based on questionnaires (green line) remains very low and stable, meaning that the participant's self-reported confidence is in line with the machine's prediction. In short, the graph indicates a constant trend of low confidence, which is a clear reflection of the subject's lack of participation and very little feedback both verbally and non-verbally during the session.

Conclusion

The proposed work introduces a multimodal confidence analysis system that operates in real-time and can assess user confidence in virtual meetings through the comprehensive analysis of face, gesture, and voice. The system is built on top of powerful computer vision and machine learning methods for the extraction and processing of non-verbal signals (facial landmarks, hand gestures) and vocal features (pitch, loudness) to give an estimate of the confidence levels of the participants. The first thing that the study did was to take a look at the shortcomings of unimodal systems that do not capture the complexities of human communication and confidence in real-time, particularly in online settings where non-verbal cues are usually less visible. With these ideas combined together, the system is claimed to be more efficient and thus giving accurate and instantaneous confidence estimations. The findings of the study indicate that the system developed has made it to the top in terms of accuracy (94.97%) against self-reported confidence scores via questionnaires, thereby proving its robustness and paving the way to future applications in virtual interviews and e-learning environments. The multimodal method together with real-time processing brings about remarkable benefits for the existing systems. It enables the users' emotional state and engagement to be perceived as a whole, thus making it very necessary for usage in media connection at a distance, online learning, and virtual interviews where the understanding of non-verbal cues is very important.

Limitations

This dissertation introduces a thoroughgoing real-time confidence detection system which is capable of integrating different aspects of facial and vocal behavior to give a confidence score. The system offers instant feedback to the users which is very useful in contexts such as virtual coaching, public speaking training, and interviews thereby improving the user's communication skills. The system which is based on the real-time processing of facial and hand gestures is indeed a major progress in the detection of facial gesture communication. The detection of facial gestures, however, at this stage needs refinements before it could be used in a variety of scenarios especially in situations involving different cultural contexts. The expansion of the database and the enhancement of the model will assist in increasing the flexibility and accuracy of the system across varied user groups.

Future Implementation

In order to make the real-time confidence detection system better in terms of performance, accuracy, and applicability, the following promising directions for future development have been suggested:

Multi-Face Detection and Analysis: The detection and analysis of multiple faces would be supported simultaneously, which would allow the system to work even in group situations like meetings, classes, conferences, and group interviews. This upgrade will give the system the power to create individual confidence assessments for every participant, thus its practical usability will be extended in multi-user scenarios.

Advanced Recognition of Complex Hand Gestures: The adoption of highly developed hand-gesture recognition methods would make it possible for the system to discern a bigger range of expressive hand motions. This improvement is particularly necessary in the case of presentations and interviews where hand gestures are subtle yet critical to communication effectiveness and the perception of confidence.

Eyebrow Movement Integration: The system would be able to spot small emotional signals like surprise, skepticism, or emphasis if it had the power to analyse eyebrow movements. Given that eyebrow dynamics are a significant part of the communication of intent and feelings, this feature's inclusion could greatly impact the accuracy and depth of the confidence estimation process.

Enhanced Visualization and Detailed Confidence Breakdown: Real-time visual overlays, such as facial landmark mapping, gesture tracking, and annotated feedback, would significantly support user engagement and interpretability if the user interface were improved in this manner. Moreover, if a thorough explanation of the confidence contributors (like gaze stability, blinking pattern, hand gestures, facial expressions, vocal clarity) were provided to the users, it could allow them to receive more insightful guidance, thus, understanding and refining the targeted behaviors which have an impact on their confidence rating.

References

[1] Du, Y., Crespo, R., & Martínez, O. (2022). Human emotion recognition for enhanced performance evaluation in e-learning. *Progress in Artificial Intelligence*, 12, 199-211. https://doi.org/10.1007/s13748-022-00278-2.

[2] Zhu, X., Liu, Z., Cambria, E., Yu, X., Fan, X., Chen, H., & Wang, R. (2024). A client-server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods. *Computer methods and programs in biomedicine*, 260, 108564 . https://doi.org/10.1016/j.cmpb.2024.108564.

[3] Sharma, A., Sharma, K., & Kumar, A. (2022). Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion. *Neural Computing and Applications*, 35, 22935-22948. https://doi.org/10.1007/s00521-022-06913-2.

[4] Verma, B., & Choudhary, A. (2021). Affective state recognition from hand gestures and facial expressions using Grassmann manifolds. *Multimedia Tools and Applications*, 80, 14019 - 14040. https://doi.org/10.1007/s11042-020-10341-6.

[5] Rani, K., Muralidhar, A., Rao, G., Shareef, S., Sudheer, P., & Bhavsingh, M. (2025). Real-Time Emotion Detection in Live Video Streams using Multimodal Deep Learning Architectures. *2025 5th International Conference on Pervasive Computing and Social Networking (ICPCSN)*, 1445-1452. https://doi.org/10.1109/icpcsn65854.2025.11034807.

[6] Kraack, K. (2024). A Multimodal Emotion Recognition System: Integrating Facial Expressions, Body Movement, Speech, and Spoken Language. *ArXiv*, abs/2412.17907. https://doi.org/10.48550/arxiv.2412.17907.

[7] Gupta, S., Kumar, P., & Tekchandani, R. (2023). A multimodal facial cues based engagement detection system in e-learning context using deep learning approach. *Multimedia Tools and Applications*, 1 - 27. https://doi.org/10.1007/s11042-023-14392-3.

[8] Villegas-Ch., W., Gutierrez, R., & Mera-Navarrete, A. (2025). Multimodal Emotional Detection System for Virtual Educational Environments: Integration Into Microsoft Teams to Improve Student Engagement. *IEEE Access*, 13, 42910-42933. https://doi.org/10.1109/access.2025.3546772.

[9] Savchenko, A., Savchenko, L., & Makarov, I. (2022). Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network. *IEEE Transactions on Affective Computing*, 13, 2132-2143. https://doi.org/10.1109/taffc.2022.3188390.

[10] Gupta, S., Kumar, P., & Tekchandani, R. (2022). Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimedia Tools and Applications*, 82, 11365 - 11394. https://doi.org/10.1007/s11042-022-13558-9.

[11] Mutawa, A., & Sruthi, S. (2023). Enhancing Human–Computer Interaction in Online Education: A Machine Learning Approach to Predicting Student Emotion and Satisfaction. *International Journal of Human–Computer Interaction*, 40, 8827 - 8843. https://doi.org/10.1080/10447318.2023.2291611.

[12] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Inf. Fusion*, 59, 103-126. https://doi.org/10.1016/j.inffus.2020.01.011.

[13] Somu, R., & Kumar, P. (2024). Analysis of Learner's Emotional Engagement in Online Learning Using Machine Learning Adam Robust Optimization Algorithm. *Scientific Programming*. https://doi.org/10.1155/2024/8886197.

[14] Lee, H., Shim, M., Liu, X., Cheon, H., Kim, S., Han, C., & Hwang, H. (2025). A review of hybrid EEG-based multimodal human-computer interfaces using deep learning: applications, advances, and challenges.. *Biomedical engineering letters*, 15 4, 587-618 . https://doi.org/10.1007/s13534-025-00469-5.

[15] Du, X., Wu, J., Tang, X., Lv, X., Jia, L., & Xue, C. (2025). Predicting User Attention States from Multimodal Eye–Hand Data in VR Selection Tasks. *Electronics*. https://doi.org/10.3390/electronics14102052.

[16] Zhao, H., Bian, S., Liu, X., & Jing, S. (2023). Learning State Detection with Multimodal Information in Virtual Reality Learning. *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, 1-6. https://doi.org/10.1109/dtpi59677.2023.10365428.

[17] Salazar, C., Montoya-Múnera, E., & Aguilar, J. (2021). Analysis of different affective state multimodal recognition approaches with missing data-oriented to virtual learning environments. *Heliyon*, 7. https://doi.org/10.1016/j.heliyon.2021.e07253.

[18] Zhang, Y., Ren, E., Song, Y., & Chen, F. (2023). Evaluation Method of Online Education Learners' Emotional Input Based on Multimodal Data Fusion. , 422-440. https://doi.org/10.1007/978-3-031-51503-3_27.

[19] Jiang, F., Peng, Y., Dong, L., Wang, K., Yang, K., Pan, C., & You, X. (2023). Large AI Model Empowered Multimodal Semantic Communications. *IEEE Communications Magazine*, 63, 76-82. https://doi.org/10.1109/mcom.001.2300575.

[20] Thakur, P., Kaur, N., Aggarwal, N., & Singh, S. (2025). A Comprehensive Review of Unimodal and Multimodal Emotion Detection: Datasets, Approaches, and Limitations. *Expert Syst. J. Knowl. Eng.*, 42. https://doi.org/10.1111/exsy.70103.

[21] Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances. *ArXiv*, abs/2203.06935. https://doi.org/10.48550/arxiv.2203.06935.

[22] Samadiani, N., Huang, G., Cai, B., Luo, W., Chi, C., Xiang, Y., & He, J. (2019). A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor

Data. *Sensors (Basel, Switzerland)*, 19. https://doi.org/10.3390/s19081863.

[23] Wu, Y., Mi, Q., & Gao, T. (2025). A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. *Biomimetics*, 10. https://doi.org/10.3390/biomimetics10070418.

[24] Güler, S., & Akbulut, F. (2025). Multimodal Emotion Recognition: Emotion Classification Through the Integration of EEG and Facial Expressions. *IEEE Access*, 13, 24587-24603. https://doi.org/10.1109/access.2025.3538642.

[25] Tellamekala, M., Amiriparian, S., Schuller, B., Andr'e, E., Giesbrecht, T., & Valstar, M. (2022). COLD Fusion: Calibrated and Ordinal Latent Distribution Fusion for Uncertainty-Aware Multimodal Emotion Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46, 805-822. https://doi.org/10.1109/tpami.2023.3325770.

[26] Amaresh, R., & Aote, S. (2025). AI based Psychometric Assessment using Multimodal Signal Data. *2025 Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, 622-627. https://doi.org/10.1109/icaiss61471.2025.11041804.

[27] Ge, S., & Chen, Y. (2025). Confidence-Aware Multimodal Learning for Trustworthy Fake News Detection. *INFORMS Journal on Computing*. https://doi.org/10.1287/ijoc.2024.0655.

[28] Immadisetty, P., Rajesh, P., Gupta, A., R., A., A., S., & Subramanya, K. (2025). Multimodality in online education: a comparative study. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-20540-0.

[29] Joshi, G., Tasgaonkar, V., Deshpande, A., Desai, A., Shah, B., Kushawaha, A., Sukumar, A., Kotecha, K., Kunder, S., Waykole, Y., Maheshwari, H., Das, A., Gupta, S., Subudhi, A., Jain, P., Jain, N., Walambe, R., & Kotecha, K. (2025). Multimodal machine learning for deception detection using behavioral and physiological data. *Scientific Reports*, 15. https://doi.org/10.1038/s41598-025-92399-6.

[30] Rahayu, Y., Fatichah, C., Yuniarti, A., & Rahayu, Y. (2025). Advancements and Challenges in Video-Based Deception Detection: A Systematic Literature Review of Datasets, Modalities, and Methods. *IEEE Access*, 13, 28098-28122. https://doi.org/10.1109/access.2025.3533545.

[31] Luo, X., Jia, N., Ouyang, E., & Fang, Z. (2024). Introducing machine-learning-based data fusion methods for analyzing multimodal data: An application of measuring trustworthiness of microenterprises. *Strategic Management Journal*. https://doi.org/10.1002/smj.3597.

[32] P, D. (2025). Landmark-based Dataset Generation using Mediapipe. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2025.68160.

[33] Hemanth, S., Dinesh, V., Raj, R., Nithin, P., Niharika, V., & Kumar, T. (2025). Real-Time Sign Language Recognition Using Advanced Computer Vision. *International Research Journal on Advanced Engineering Hub (IRJAEH)*. https://doi.org/10.47392/irjaeh.2025.0368.

[34] Shao, Z., Dou, W., & Pan, Y. (2023). Dual-level Deep Evidential Fusion: Integrating

multimodal information for enhanced reliable decision-making in deep learning. *Inf. Fusion*, 103, 102113. https://doi.org/10.1016/j.inffus.2023.102113.

[35] Guo, W., Wang, J., & Wang, S. (2019). Deep Multimodal Representation Learning: A Survey. *IEEE Access*, 7, 63373-63394. https://doi.org/10.1109/access.2019.2916887.

[36] Jabeen, S., Li, X., Amin, M., Bourahla, O., Li, S., & Jabbar, A. (2022). A Review on Methods and Applications in Multimodal Deep Learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19, 1 - 41. https://doi.org/10.1145/3545572.

[37] Zhao, F., Zhang, C., & Geng, B. (2024). Deep Multimodal Data Fusion. *ACM Computing Surveys*, 56, 1 - 36. https://doi.org/10.1145/3649447.

[38] Wu, P., Lin, J., , Z., & Li, H. (2025). Enhancing Interpretability: A Hierarchical Belief Rule-Based (HBRB) Method for Assessing Multimodal Social Media Credibility. *Int. J. Intell. Syst.*, 2025. https://doi.org/10.1155/int/7184626.

[39] Kong, X., & Ge, Z. (2022). Deep PLS: A Lightweight Deep Learning Model for Interpretable and Efficient Data Analytics. *IEEE Transactions on Neural Networks and Learning Systems*, 34, 8923-8937. https://doi.org/10.1109/tnnls.2022.3154090.

[40] Mamieva, D., Abdusalomov, A., Kutlimuratov, A., Muminov, B., & Whangbo, T. (2023). Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. *Sensors (Basel, Switzerland)*, 23. https://doi.org/10.3390/s23125475.

[41] Ye, J., Yu, Y., Lu, L., Wang, H., Zheng, Y., Liu, Y., & Wang, Q. (2025). DEP-Former: Multimodal Depression Recognition Based on Facial Expressions and Audio Features via Emotional Changes. *IEEE Transactions on Circuits and Systems for Video Technology*, 35, 2087-2100. https://doi.org/10.1109/tcsvt.2024.3491098.

[42] Jiang, Z., Seyedi, S., Griner, E., Abbasi, A., Rad, A., Kwon, H., Cotes, R., & Clifford, G. (2024). Multimodal Mental Health Digital Biomarker Analysis From Remote Interviews Using Facial, Vocal, Linguistic, and Cardiovascular Patterns. *IEEE Journal of Biomedical and Health Informatics*, 28, 1680-1691. https://doi.org/10.1109/jbhi.2024.3352075.

[43] Zhu, X., Liu, Z., Cambria, E., Yu, X., Fan, X., Chen, H., & Wang, R. (2024). A client-server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods. *Computer methods and programs in biomedicine*, 260, 108564 . https://doi.org/10.1016/j.cmpb.2024.108564.

[44] Li, C., Weng, X., Li, Y., & Zhang, T. (2024). Multimodal Learning Engagement Assessment System: An Innovative Approach to Optimizing Learning Engagement. *International Journal of Human–Computer Interaction*, 41, 3474 - 3490. https://doi.org/10.1080/10447318.2024.2338616.

[45] Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning?. *Br. J. Educ. Technol.*, 51, 1450-1484. https://doi.org/10.1111/bjet.12993.

[46] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, A design science research

methodology for information systems research, Journal of management information systems 24 (3) (2007) 45–77. https://doi.org/10.2753/MIS0742-1222240302

[47] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS quarterly (2004) 75–105. https://doi.org/10.2307/25148625

[48] Boulay, E., Wallace, B., Fraser, K., Kunz, M., Goubran, R., Knoefel, F., & Thomas, N. (2024). Improving Synchronization of Eye Fixation and Saccade Measurements with Speech Recognition for Cognitive Assessment. *2024 IEEE Sensors Applications Symposium (SAS)*, 1-6. https://doi.org/10.1109/sas60918.2024.10636646.

[49] Boulay, E. (2024). Eye-Tracking and Speech Analysis Measurement and Synchronization for Cognitive Assessment. . https://doi.org/10.22215/etd/2024-16227.

[50] Simmatis, L., Naeini, S., Jafari, D., Xie, M., Tanchip, C., Taati, N., McKinlay, S., Sran, R., Truong, J., Guarin, D., Taati, B., & Yunusova, Y. (2023). Analytical Validation of a Webcam-Based Assessment of Speech Kinematics: Digital Biomarker Evaluation following the V3 Framework. *Digital Biomarkers*, 7, 7 - 17. https://doi.org/10.1159/000529685.

[51] Mori, Y., & Pell, M. (2019). The Look of (Un)confidence: Visual Markers for Inferring Speaker Confidence in Speech. *Frontiers in Communication*, 4. https://doi.org/10.3389/fcomm.2019.00063.

[52] Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P., Marias, K., & Tsiknakis, M. (2017). Stress and anxiety detection using facial cues from videos. *Biomed. Signal Process. Control.*, 31, 89-101. https://doi.org/10.1016/j.bspc.2016.06.020.

[53] Kovacs-Balint, Z., Bereczkei, T., & Hernádi, I. (2012). The telltale face: possible mechanisms behind defector and cooperator recognition revealed by emotional facial expression metrics.. *British journal of psychology*, 104 4, 563-76 . https://doi.org/10.1111/bjop.12007.

[54] Krumhuber, E., Skora, L., Hill, H., & Lander, K. (2023). The role of facial movements in emotion recognition. *Nature Reviews Psychology*, 2, 283 - 296. https://doi.org/10.1038/s44159-023-00172-1.

Appendix

Hardware Configuration

The proposed confidence detection system was tested using widely available and low-cost hardware components in order to ensure both accessibility and deployment feasibility.

Webcams: Conventional webcams that capture real-time video.

Computing Devices: Mid-range laptops and desktop computers are equipped with general-purpose CPUs and integrated GPUs.

Audio Input: Laptop computer built-in microphones or simple USB microphones for vocal feature extraction.

Software Tools and Frameworks

The system has been implemented using open-source and platform-independent technologies.

MediaPipe: Utilized for facial landmark detection, hand tracking, and gesture analysis.

JavaScript: Primary programming language for browser-based system development.

MediaPipe Camera Utils: Used to stream real-time video input and manage frames.

Hosting Platform: The web application was deployed and executed on a cloud-based hosting service.

Machine Learning and Processing Models

The following lightweight models were integrated in order to enable real-time multimodal confidence estimation:

Face Mesh Model: MediaPipe model, which detects 468 landmarks on a face for gaze, expression, or head pose analysis.

Hand Tracking Model: A model used to identify 21 hand landmarks for the analysis of gesture activities.

Audio Analysis Module: Extracting loudness, pitch variation, and clarity of vocals in real time.

Experimental Setup and Test Parameters

The system was evaluated under controlled experimental conditions to ensure measurement uniformity.

Participants: A total of 100 participants participated in the testing process.

Environment: Experiments were run in a light indoor environment with minimal background noise.

Distance of camera: The participants were positioned approximately 2 feet from the webcam.

Task duration: Each participant completed a 2-minute speaking activity for data collection.

Device Variability: Testing of the system was performed on more than one device to check consistency in performance.

Data Collected During Evaluation

The system captured several multimodal behavioral cues relevant for confidence estimation:

Facial Features: Blink rate, smile detection, lip movement patterns, and head movement.

Gesture features: Gesture frequency, activity intensity, and stability of motion.

Vocal Features: Pitch stability, loudness, clarity, and speech fluency.

Confidence Score: Real-time confidence values derived from weighted contributions of face, hand, and voice modalities.

Evaluation Metrics

The following are the metrics that were considered for assessing the performance of the system:

- Real-time frame processing accuracy
- Facial and hand landmark detection stability
- Latency in confidence computation
- Cross-user consistency
- Performance under varying lighting and motion conditions