Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1, 525-543
2026
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v10i1.11626

IVAS: A multimodal AI system for objective video interview assessment with facial emotion, gaze, and audio analysis

Syed Azeem Inam[1*], Abdul Kabeer[1], Muneeb Ahmed Abbasi[1], Abdullah Ayub Khan[2*]
[1]Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi, Pakistan; syedazeeminam@gmail.com, syed.azeem@smiu.edu.pk (S.A.I.) ak0906113@gmail.com (A.K.) muneeb.abbasi13@gmail.com (M.A.A.).
[2]Department of Computer Science, Bahria University Karachi Campus, Karachi 75260, Pakistan; abdullah.khan00763@gmail.com, abdullahayub.bukc@bahria.edu.pk (A.A.K.).

Abstract: Traditional interviewing literature has regularly highlighted its vulnerability to human subjectivity and unconscious bias, as such pitfalls may cause interviewers to overlook critical behavioral cues. To address these shortcomings, the present study proposes an artificial intelligence-based assessment tool, the Interview Video Analysis System (IVAS), which integrates Facial Emotion Recognition (FER), Gaze Tracking, and Audio Analysis technologies into a single, cohesive system to evaluate candidates objectively. IVAS comprises a Convolutional Neural Network (CNN) based on 22 layers trained on the FER-2013 dataset, achieving an accuracy of 86% in recognizing seven different emotions: Anger, Fear, Sadness, Happiness, Disgust, Surprise, and Neutral, through sophisticated data augmentation and hyperparameter optimization. The system utilizes 68-point facial features from dlib, with its gaze module measuring eye contact, directional changes, and blinks, relating these metrics to engagement and confidence. Additionally, the audio component employs Llama 3.2 with 11 billion parameters and Mel-Frequency Cepstral Coefficients (MFCC) to extract voice features such as pitch, hesitancy, and fluency, and transcribes speech for linguistic analysis. The late-fusion logic aggregates outputs from each module into a well-organized system that provides percentage measures and performance levels. As a Streamlit web application, IVAS can process live and recorded interviews in real-time, offering recruiters a scalable, data-driven assessment tool. This system outperforms other state-of-the-art models by approximately 12 to 30% in FER efficiency and is the first to incorporate multimodal behavioral analysis to reduce ambiguities inherent in unimodal methods. Collectively, this comprehensive system enhances traditional interviewing techniques by providing standardized, bias-proof insights into candidate compatibility.

Keywords: *Audio analysis, Computer vision, Convolutional neural network, Gaze tracking, Video interview analysis system.*

1. Introduction

Traditional interviews rely on human judgment, which can be subjective and prone to biases, often overlooking subtle behavioral cues. To address these weaknesses, the present study proposes an Interview Video Analysis System (IVAS) developed using advanced artificial intelligence techniques to help organizations analyze candidates through video and audio analysis. The proposed system in this study analyzes gaze patterns, facial emotion behavior, and voice recognition through automation, providing a single, scalable, and objective evaluation method.

During interviews, confidence and composure are essential indicators of a candidate's suitability, as well as their emotions, such as happiness and sadness, which can be analyzed to assess the emotional state of an individual and their suitability as a candidate [1]. For this, Facial Emotion Recognition (FER), an integral part of the proposed system in the present study, has been widely used in the fields of

History: Received: 22 October 2025; Revised: 9 December 2025; Accepted: 12 December 2025; Published: 1 January 2026
* Correspondence: syedazeeminam@gmail.com, syed.azeem@smiu.edu.pk

behavioral analysis and computer interaction and is being increasingly incorporated into clinical practices nowadays [2, 3]. Conventionally, a FER utilizes Convolutional Neural Networks (CNNs) owing to their extensive ability for classification and extraction, for detecting emotions accurately [4-15]. Additionally, the basis of the proposed system in the present study is the FER-2013 dataset, which comprises 48x48 grayscale images categorized into seven different basic emotions, providing a solid foundation for designing a robust deep learning model [16]. Similarly, with the help of advanced architectures such as lightweight CNNs and VGGNet, high accuracy can be achieved even in rough conditions, such as varying lighting or facial orientation [17].

Gaze tracking is another key feature of the proposed system, which evaluates candidates' focus and engagement. Eye movement is a powerful indicator of attention and intent; it informs how candidates interact during an interview. Through investigation, it has been found that gaze behavior, directional shifts, time of eye contact, attentiveness, and confidence are all interconnected [18, 19]. The proposed system uses gaze detection and computer vision techniques. With a standard video that we input, the system analyzes and calculates the candidate's response, such as how much the candidate looked at the center to maintain eye contact or how much time the candidate spent looking at the left and right sides. This method is based on research findings that demonstrate the importance of user intent, visual attention, and content in exploring relationships. With the help of integrated fixation analysis and gaze mapping techniques, the system provides a detailed analysis of candidates, which can help in understanding visual engagement [18].

Audio analysis is another integral part of the proposed system within the multi-model framework, as vocal features such as tone, hesitancy, pitch, confidence, communication skills, and nervousness are also key indicators. For this, the proposed system utilizes Llama 3.2, an advanced Large Language Model (LLM) for text analysis and computer vision [20, 21]. The proposed system utilizes eleven (11) billion parameters of the Llama model to handle multitasking and multilingual scenarios comprehensively, allowing it to be an ideal choice for analyzing diverse candidate profiles. By incorporating techniques for unique feature extraction and pitch contour analysis, it can detect vocal patterns in communication and emotions [20, 22] thereby providing detailed insights into analyzing verbal behavior derived from visual data in FER and Gaze tracking [18].

Finally, the Interview Video Analysis System (IVAS) is a Streamlit-based web application that processes interview videos through three core modules: FER, gaze tracking, and audio analysis. The results of the analysis will be presented in a structured report, complete with percentage breakdowns and comparative benchmarks, enabling recruiters to make informed, data-driven hiring decisions.

While previous research has explored individual AI-based emotion detection, gaze tracking, and speech analysis, our system uniquely integrates all three modalities for the comprehensive evaluation of candidates. This combination provides a holistic approach to understanding behavioral and emotional cues. According to FER studies, it helps us detect subtle emotional cues that may sometimes remain unnoticed [17]. Similarly, gaze-tracking research emphasizes the most relevant aspects of access, engagement, and attentiveness. Advanced AI models, such as Llama 3.2, work on text patterns to enhance a system's analysis capabilities [20].

2. Related Work

Facial emotion recognition (FER) is utilized in Human-Computer Interaction (HCI) for behavioral analysis, healthcare, and online advertising. Initially, systems consisting of FER mainly depended on hand-crafted features and classical machine learning. These methods were Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG) techniques, which were used to extract features and facial images. In contrast, Decision Tree (DT) classifiers and Support Vector Machines (SVM) were used for classification [23] which performed exceptionally well in controlled environments; however, they lacked robustness due to variations in lighting, poses, and expressions in real-world scenarios. The introduction of CNNs in the 1990s revolutionized the approach towards image processing [24]. Comprising convolutional layers, pooling layers, and fully connected layers, it proved highly effective in

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

automated feature extraction and performed the classification of images quite effectively [25]. However, its utilization was initially restricted due to the limited processing power required to handle large datasets. By the 2010s, advancements in hardware (GPUs) and datasets, such as FER-2013, accelerated the growth of deep learning-based face recognition systems. FER-2013 is the first dataset that recognizes real-world emotions, introduced by LCML in 2013 [26]. Afterward, researchers experimented with different variants of CNNs, which subsequently led to the development of well-recognized models such as VGGNet, InceptionNet, and ResNet, all of which have significantly improved the accuracy of classification.

Many researchers have now leveraged the power of CNNs for classifying FER-2013 datasets. For instance, Liu et al. [27] used CNN for classification on the FER-2013 dataset, resulting in an accuracy of 71.2% [27]. However, on further investigation, the researchers improved the accuracy by 2.6% through the incorporation of an ensemble method that combined three variants of CNN, but the utilization of advanced techniques, training, and resources increases the cost of the overall process [27]. Another researcher compared ResNet and InceptionNet models, achieving the highest accuracy of 70% [28]. Similarly, Khaireddin and Chen optimized VGGNet by tuning the hyperparameters, testing different optimizers and learning rate schedulers [29]. They successfully achieved an accuracy of 73.28% in FER-2013, indicating the need for hyperparameter tuning in FER for optimal performance. Gaze tracking is also a crucial feature utilized in multimodal AI systems that assess attentiveness during interviews, including eye movements such as fixation and directional shifts, to help understand attentiveness and intent. In real-world applications, such as accident prevention, the importance of gaze tracking is of great interest, as it helps improve human focus. It can also aid in understanding human interaction and content [30]. Recent studies have proven the successful implementation of gaze tracking using a webcam, which enhances the ability to track [3]. In the proposed system, gaze metrics, such as contact frequency and gaze shift patterns, combined with candidate evaluation, have provided valuable insights. Audio analysis is another feature that is widely used in AI systems to understand human behavior [31]. Vocal features, such as pitch modulation, tone variability, and hesitancy, help in understanding the candidate's confidence and emotional state. The researchers in Onyema et al. [28] have extensively discussed dynamic scenarios, including interviews with emotional cues detected through gaze-enabled detection and vocal analysis using the Llama 3.2 model. Another study has presented a practical approach that utilizes pitch contour and spectral feature extraction techniques in combination with visual features to provide an understanding of holistic behavior [32].

In the context of multimodal video analytics, recent advances have already led to significant progress in integrating heterogeneous data modalities to provide comprehensive behavioral analysis. Murthy and Siddesh [33] proposed a sophisticated video analytics system based on sarcasm, utilizing their SarcasNet-99 system, and demonstrated 99.005% precision after combining text, image, and audio attributes using an adaptive fusion network on the TEDx and GIF Reply datasets. In another study, they developed a Deep Hybrid Fusion algorithm [34] that achieved 95.84% accuracy on the data from IEMOCAP, CMU-MOSI, and CMU-MOSEI, utilizing Modality Unchanged-Precise Representation learning and a BiLSTM encoder-decoder structure. Furthermore, in another study, the same researchers utilized the power of quantum theoretical formalisms to address intra- and inter-expression dynamics, modeling this through a quantum-conscious multimodal option mining architecture by incorporating networks of CNN and LSTM and demonstrating better performance in the task of conversational sentiment analysis [35]. Altogether, these experiments confirm the effectiveness of deep learning-based multimodalities in video processing and propose a solid foundation for AI-powered systems of interviewing evaluation, which incorporate facial emotion recognition, gaze tracking, and audio analysis to aid in the thorough review of candidates.

2.1. Challenges in FER Methods

Despite exceptional progress, there are still some gaps in FER methods, as the currently available models struggle with emotions, intra-class variations, and inter-class similarities. For example, the

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

emotions of "sadness" and "fear" are sometimes misclassified due to their similar visual appearance [36], which highlights the need for the development of a more robust FER system. Although, as discussed earlier, multiple models have performed exceptionally well on different datasets, they still perform poorly in real-world applications due to multiple reasons. Various studies have addressed this issue using shallow architectures, which are incapable of capturing multiple features and establishing standard optimization; however, they offer different optimizers and learning rate schedulers [29, 37].

2.2. Our Approach towards Solution

The present study proposes a 22-layer, more profound CNN architecture to overcome these limitations. Deeper networks are capable and helpful in identifying complex features and are far better at identifying similar emotions. The proposed system achieves an accuracy of 86%, surpassing the accuracy of state-of-the-art systems, as shown in Table 1. The proposed system utilizes advanced techniques of considerable data augmentation and fine-tuning, which have enhanced the robustness of the FER system for real-time applications.

2.3. Objective

The goal of developing the proposed IVAS is to accurately predict emotions in normal video, achieving robustness in real-time expression, pose variation, and lighting. The four essential architectural features of the proposed system are as follows:

1. The proposed system comprises 22 convolutional layers, pooling layers, and fully connected layers to extract features.
2. Scaling, flipping, and random rotation techniques are utilized to avoid overfitting in data augmentation.
3. Hyperparameters, including learning rate, dropout rate, and batch size, have been optimized to ensure adequate training.
4. The performance of the model has been evaluated using standard evaluation metrics, including accuracy, precision, recall, and F1 score.

Table 1.
Model, Limitations, and Solutions.

Ref.	Model Used	Accuracy	Limitations/Challenges	How Our Approach Addresses These
Liu et al. [27]	CNN Ensemble	62.44%	• Lower accuracy compared to other methods • Computational overhead of ensemble methods • High training and resource costs	• Single 22-layer model achieving 86% accuracy • Efficient single-model architecture. • Optimized hyperparameters for resource efficiency
Murthy and Siddesh [34]	ResNet50 with Dropout Layers	55.6%	• Low accuracy performance. • Limited architectural depth • Poor generalization to real-world conditions • Standard optimization without fine-tuning	• 86% accuracy with 22-layer architecture • Advanced data augmentation and hyperparameter tuning • Real-time inference capabilities • Adaptive optimizers
Khaireddin and Chen [29]	Optimized VGGNet	73.28%	• Shallow architecture with only four (04) convolutional blocks • Cannot leverage the depth advantages of modern deep learning • Inconsistent optimization performance	• 22-layer deep CNN for complex feature extraction. • Standardized optimization with adaptive optimizers. • Better discrimination between similar emotions
Maiden and Nakisa [38]	Continual Learning Model with	74.28%	• Complex architecture requiring specialized training. • Limited to few-shot learning	• Direct training on the comprehensive dataset. • 86% accuracy without complex

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

	knowledge distillation and Predictive Sorting Memory Replay		scenarios. • Computational complexity of knowledge distillation	distillation. • Real-time processing for practical applications
Our Proposed System	Deep CNN (22 Layers) + Gaze Tracking + Audio Analysis (Llama 3.2)	86%	-	• 22-layer deep CNN for acceptable emotion discrimination. • Advanced data augmentation (rotation/flipping/scaling) for robustness. • Multimodal integration resolving ambiguous visual cues. • Real-time processing capabilities. • 86% accuracy surpassing state-of-the-art. • Dataset balancing techniques for FER-2013 bias mitigation

3. Data Description

Goodfellow et al. [39] introduced the FER-2013 dataset in 2013 during the International Conference on Machine Learning, which is used for recognizing facial emotions [2]. This dataset is based on 35,887 grayscale images, and each image consists of a 48x48 pixel configuration. The images are further categorized into seven emotions, i.e., Anger, Fear, Sadness, Happiness, Disgust, Surprise, and Neutral. The dataset is also classified into three subsets, i.e., training, public, and private test. The training subset has 28,709 grayscale images, the public test subset has 3,589 images, and the private subset consists of 3,589 images. Using grayscale images helps reduce computational requirements and preserve essential features, as highlighted by studies conducted by researchers [3, 29]. The data is collected naturally and must face challenges such as lightning, poses, and occlusions. In a study conducted by Mukhopadhyay et al. [1], the researchers verified and ensured that models trained on this dataset would perform robustly in real-life scenarios [1]. However, in another study, Zahara et al. [26] highlighted the uneven distribution of datasets; for instance, the "disgust" emotion is underrepresented, causing a persistent challenge for researchers in the utilization of datasets [26]. Furthermore, the performance of humans in this dataset is estimated to be 65-68%, which serves as a benchmark for automated systems [3, 17].

Figure 1.
Seven Facial Emotions of FER-2013.
Source: Khaireddin and Chen [29].

3.1. Dataset Preprocessing

Owing to its importance and effectiveness, the present study utilizes the FER-2013 dataset for training and evaluation. According to researchers in Khaireddin and Chen [29] this dataset is ideal for deep learning models because of its extensive range of variations and emotional expressions Khaireddin and Chen [29]. Oguine et al. [40] further demonstrated its suitability by achieving significant classification performance using hybrid deep learning models [40]. To address challenges such as class imbalance, for instance, the underrepresentation of the category "disgust," this study employs data

augmentation and balancing techniques that have significantly improved performance metrics. The proposed system further enhances its robustness by integrating gaze tracking and vocal analysis. When visual features are ambiguous, it refines the emotion classification of gaze patterns, as shown [41] by the researchers in Kessous et al. [41]. The inclusion of gaze tracking in the CNN-based model improves performance by incorporating additional information, such as tracking gaze direction, as well as emotions like "surprise" and "fear" that can be differentiated. After all, both facial features are similar but have different focuses [42]. Ultimately, the system can utilize multiple methods to create robust models for diverse environments.

3.2. Data Augmentation

To improve model generalization, data augmentation has been employed extensively, including image flipping, rotation, and brightness adjustments. Furthermore, normalization techniques have enhanced feature extraction by standardizing pixel intensity values.

4. Methodology and Results

The study presents an Interview Video Analysis System (IVAS) that follows a new multidisciplinary approach where three anglicistic modules, i.e., Facial Emotion Recognition (FER), Gaze Tracking, and Audio Analysis, are consolidated into a monolithic configuration, which is being referred to as a singular, conjoined system. Separate modules handle the videos of the interviews, processing each to extract discrete behavioral and emotional features, which are then combined to create an objective profile of the person being interviewed. The implementation of the system is performed as a real-time, web-based system using the Streamlit framework, which enables the parallel assessment of both live webcam and pre-recorded meetings. IVAS provides an in-depth evaluation of candidate behavior by embedding several modalities, thereby avoiding the limitations of the classic interviewing approach.

4.1. Facial Emotion Recognition Module

The FER module forms the baseline of the emotional analysis framework and is based on FER-2013, a collection of 35,887 grayscale images (48 x 48 pixels) categorized into seven emotions. Table 2 and Figure 2 depict that the data is highly skewed concerning the classes.

Table 2.
Data Configuration of FER-2013.

Emotion	Train	Test
Surprise	3,171	831
Fear	4,097	1,024
Angry	3,995	958
Neutral	4,965	1,233
Sad	4,830	1,247
Disgust	436	111
Happy	7,215	1,774

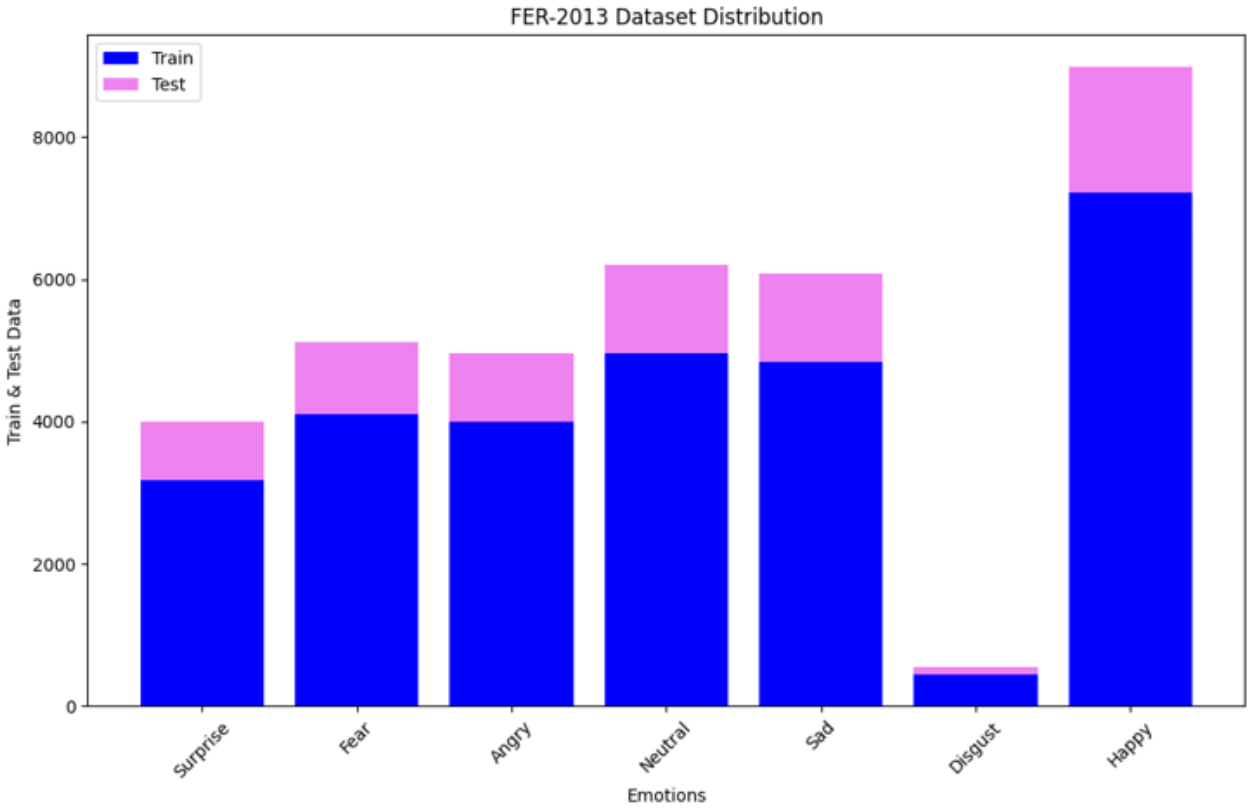

Figure 2.
FER-2013 Data Distribution.

A significant challenge of the prevalently observed class imbalance necessitates the implementation of an extensive data preprocessing pipeline; therefore, the initial aim was to enhance model generalizability and compensate for the limited dataset by implementing a comprehensive augmentation pipeline consisting of four geometric and photometric transformations. Firstly, horizontal flipping was conducted with a 50% probability, aiming to simulate lateral viewpoint variations; however, the human facial expressions remained unchanged. Then, the rotation was parameterized at 15 degrees in both directions, uniformly distributed, and served to simulate the natural movement of heads. It exploited bilinear interpolation to retain continuity in features across image edges. Afterward, isotropic scaling was used to scale across the range of amplitude. Altogether, these manipulations address several important measures of robustness, such as invariance to camera-subject positioning setups involving flipping and rotation, invariance to variations in capture distance through scaling, and tolerance to ambient lighting by modulating brightness while maintaining label continuity within the limits of allowed parameters, as visually tested. These changes optimize the models to be more robust when subjected to diverse lighting conditions and facial orientations while also making the training dataset an effective measure. After these operations, each input image was normalized concerning its pixel intensity to a range of 1, thereby standardizing the input distribution to ensure faster training. Original grayscale images were also converted into 3-channel RGB images, making them compatible with the pre-trained weights of the VGG16 model.

Finally, the FER module employs a transfer learning approach, building upon an existing VGG16 CNN-based model with additional customized layers designed to enhance the classification of emotions. The final architecture is a combination of the VGG16 baseline (14,714,688 parameters) and customized

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

layers, resulting in a network with a total of 14,883,399 parameters. The details of the model architecture are presented in Table 3.

Table 3.
Architecture of Our Proposed Model.

Layer (Type)	Output Shape	Parameters
VGG16 (Functional)	(None, 1, 1, 512)	14,714,688
Batch Normalization	(None, 1, 1, 512)	2,048
Gaussian Noise	(None, 1, 1, 512)	0
Global Average Pooling 2D	(None, 512)	0
Flatten	(None, 512)	0
Dense	(None, 256)	131,328
Batch Normalization	(None, 256)	1,024
Dropout	(None, 256)	0
Dense	(None, 128)	32,896
Batch Normalization	(None, 128)	512
Dropout	(None, 128)	0
Dense	(None, 7)	903

Table 4 presents the configuration details and the hyperparameters of the proposed model.

Table 4.
Parameter Configuration of the Proposed Model.

Parameter	Value
Learning Rate	1×10^{-4}
Beta 1	0.9
Beta 2	0.999
Epsilon	1×10^{-8}
Optimizer	Adam
Loss Function	Categorical Crossentropy
Batch Size	64
Epochs	30
L2 Regularization	0.001
Dropout Rate	0.5
Gaussian Noise	0.01

To compensate for a significant imbalance in classes in the FER-2013 dataset, the study incorporated inverse frequency weighting, where each class weight was computed as max_samples / samples_per_class[i], as shown in Table 5.

Table 5.
Inverse Frequency Weighting.

Class ID	Emotion	Class Weight
0	Surprise	1.81
1	Disgust	16.55
2	Angry	1.76
3	Neutral	1.00
4	Sad	1.45
5	Fear	1.49
6	Happy	2.28

The training protocol for the comprehensive callback mechanism is shown in Table 6.

Table 6.
Configuration of the Callback Mechanism.

Callback	Purpose	Monitor	Parameters
Model Checkpoint	Save the best model	Val_loss	save_best_only=True
Reduce LR On the Plateau	Reduce the learning rate on a plateau	Val_loss	patience=5, factor=0.5, min_lr=1e-10
Early Stopping	Stop training early	Val_loss	patience=0, restore_best_weights=True

The seven-class emotional recognition task was performed using a rigorously defined set of evaluation metrics that were devised and implemented within its context. The customized F1-score technique was applied with the Keras backend, which included true positives, possible positives, and predicted positives being counted exactly, and precision and recall (respectively) were calculated based on these counts before calculating the harmonic means. The model's performance demonstrates excellent stability overall, as illustrated in Figures 3 and 7, including improved performance in both training and validation, as shown in Table 7.

Table 7.
Performance of the Metric for Training and Validation.

Metric	Train	Validate
Loss	2.020	1.982
Accuracy	0.859	0.862
Precision	0.585	0.645
Recall	0.051	0.072
AUC	0.738	0.752
F1 Score	0.092	0.128

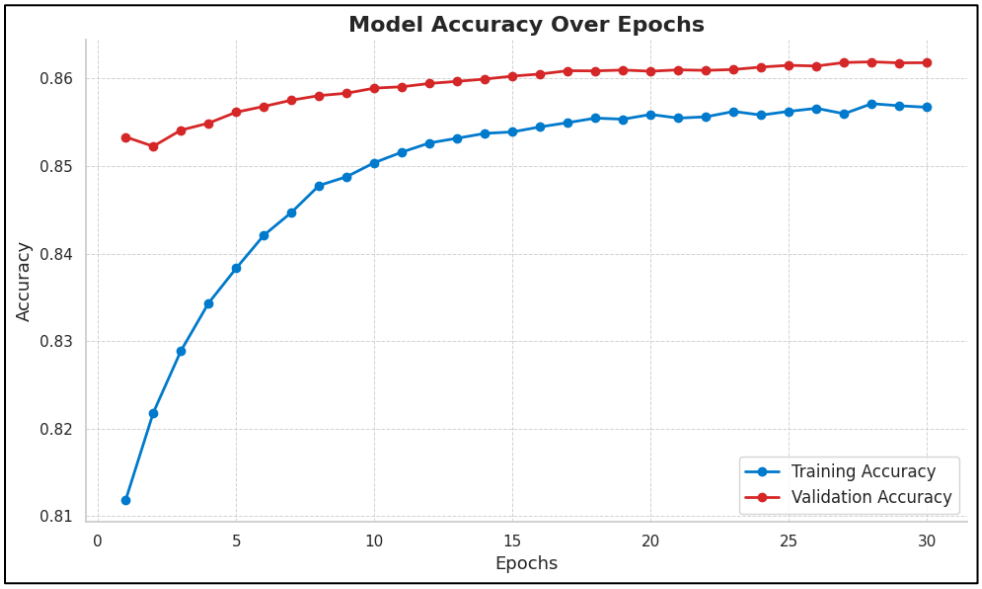

Figure 3.
Overall Model Accuracy.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

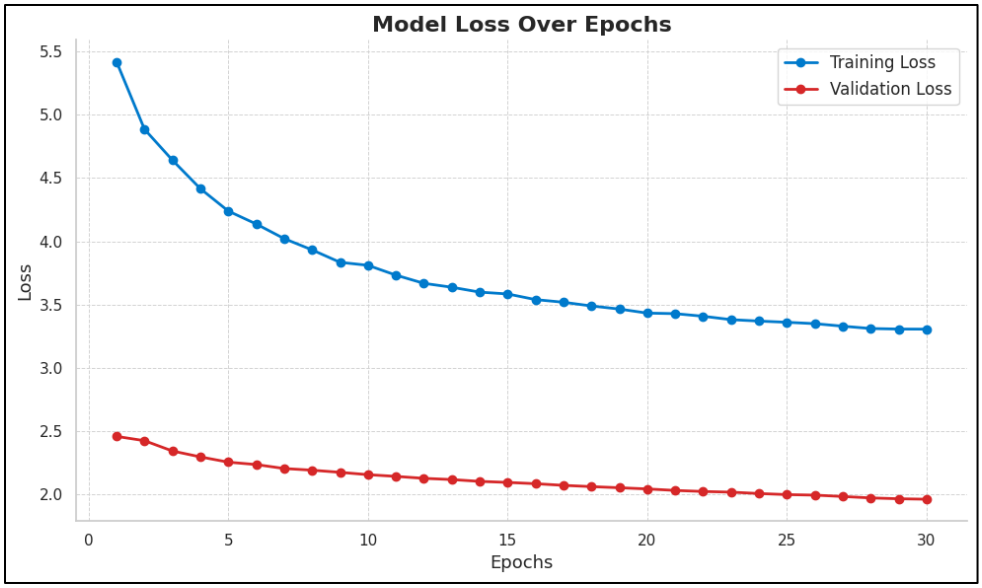

Figure 4.
Overall Model Loss.

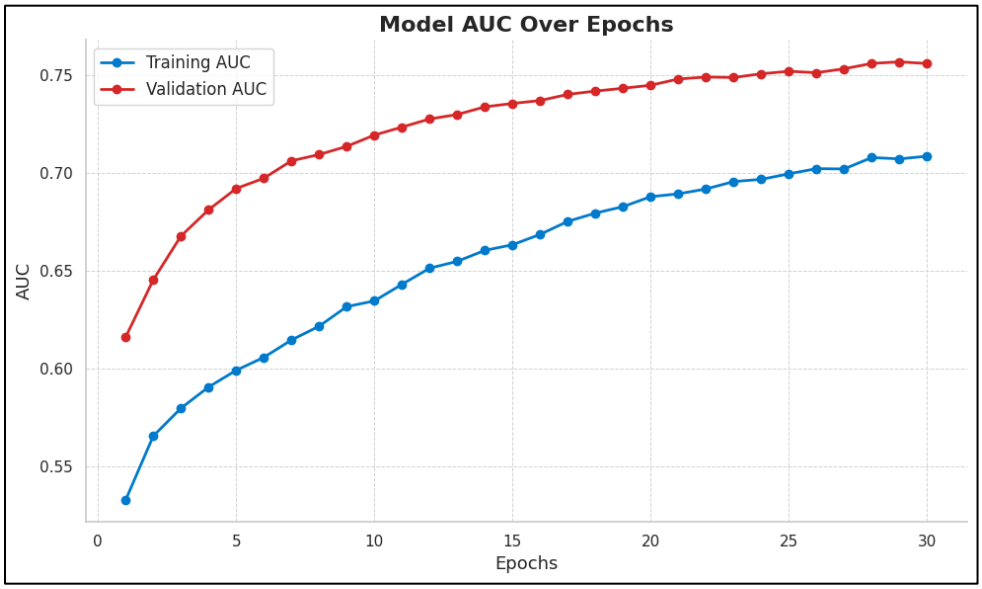

Figure 5.
Overall Model AUC.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

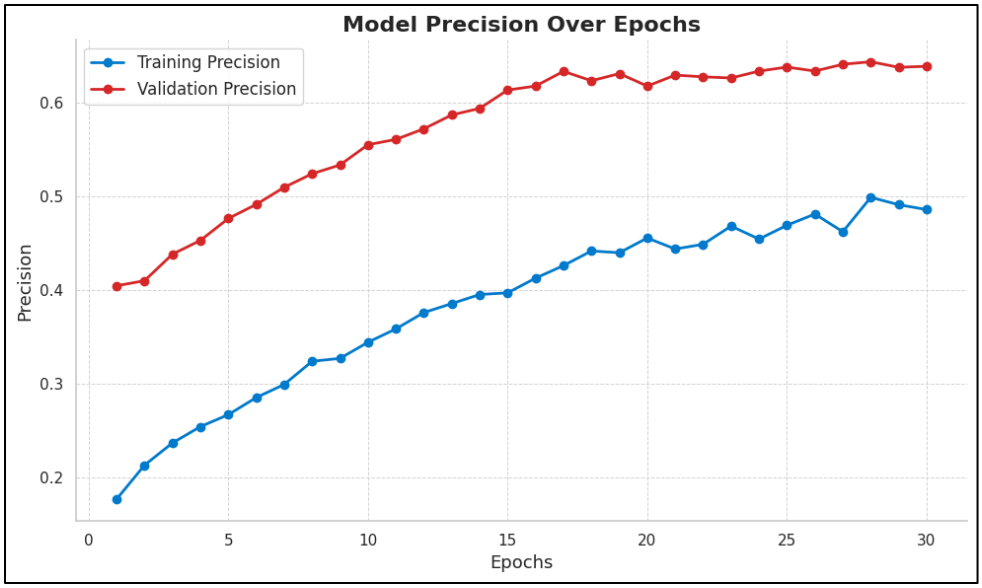

Figure 6.
Overall Model Precision.

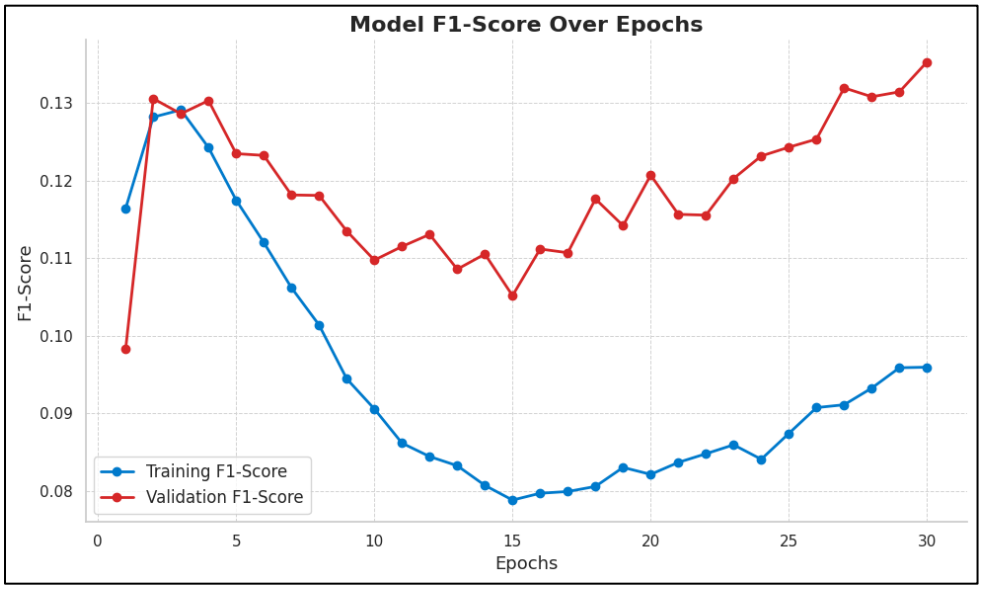

Figure 7.
Overall Model F1 Score.

The multi-class results, calculated in terms of true positives, false positives, and false negatives, are performed on an individual class basis. Precision and recall are computed at the class level and macro-averaged across all emotion categories. The confusion matrix in Figure 8 shows that most of the emotional states were classified with the highest possible accuracy, with only a few nonsignificant misclassifications between visually similar emotions, thus reflecting the model's ability to differentiate between specific emotional states.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

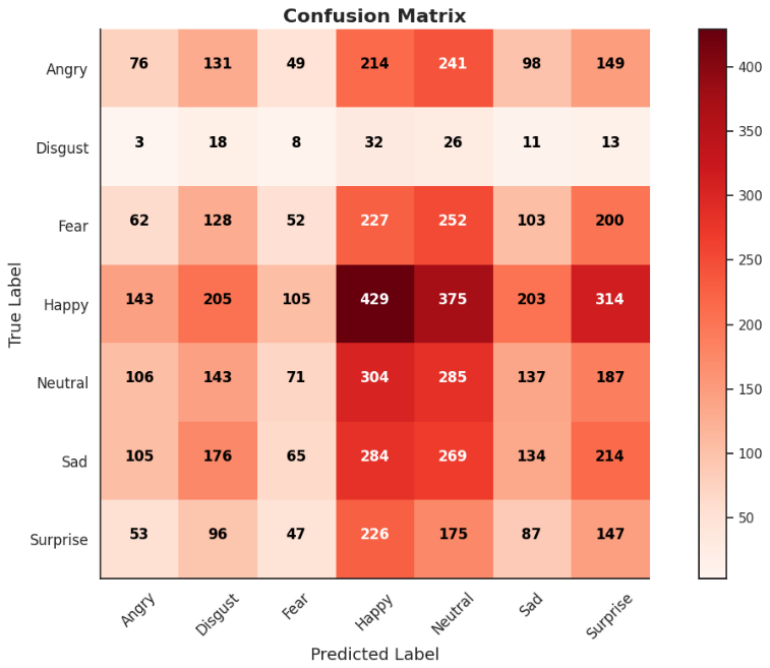

Figure 8.
Confusion Matrix.

4.2. Gaze Tracking Module

To track gaze direction, a gaze-tracking module has been developed using the dlib library to follow the eyes in video interviews. Frame by frame, it identifies facial landmarks, focusing on the ocular region, i.e., points 36–47 in the contours of the eyes, as shown in Figure 9 and Figure 10. Based on a 68-point facial landmark predictor implemented in dlib, the system extracts salient points in the primary eye areas. It calculates the gaze vectors by triangulating the coordinates of the eye center and the approximated pupil. These vectors are categorized into three main types: center gaze, which directs the gaze toward the camera or interviewer; left gaze, which directs the gaze toward the left visual field; and right gaze, which directs the gaze toward the right visual field. The temporal analysis provides the times of each gaze category throughout the entire video, the proportions of time spent in each orientation, and the total engagement score. The module also computes the Eye Aspect Ratio (EAR) in every frame. It identifies blinks when the EAR falls below a certain threshold, allowing analysis of blink frequency as an indicator of nervousness or stress. By combining the monitoring of gaze patterns and blinks, the system generates engagement scores that are correlated with the level of attentiveness and confidence in the candidates.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

Figure 9.
One of the Author's Gaze Detection while Looking at the Centre.

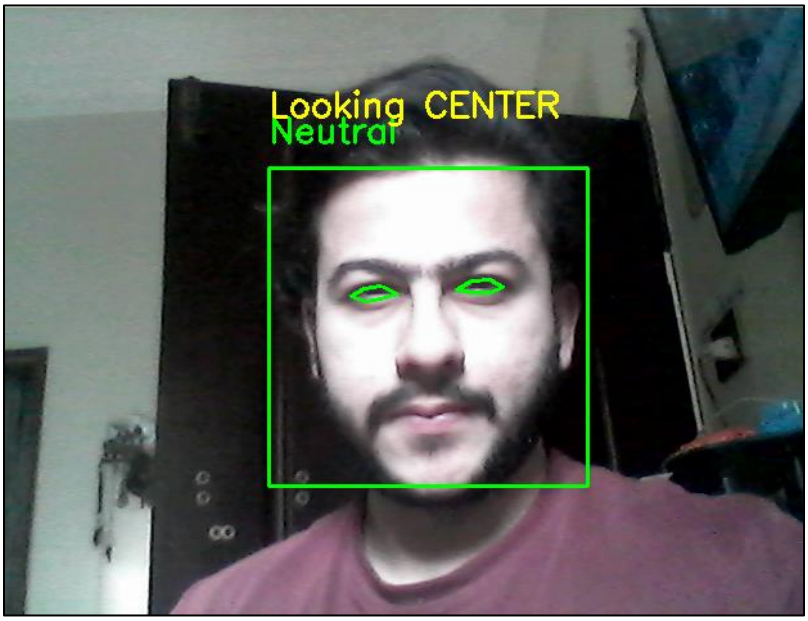

Figure 10.
Real-time Detection of Gaze and Emotion.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

4.3. Audio Analysis Module

The audio analysis module is a core element of the entire infrastructure solution within the multimodal project, as it provides a comprehensive range of vocal feature extraction and linguistic analysis mechanisms. The Python speech-recognition library is used as part of this module to extract the audio track from video interviews. This audio is recorded in WAV format, which will be used later for processing, such as noise removal, to enhance speech clarity and ensure more accurate transcription. Further language processing requires the transcription of the audio into text, which the same library can facilitate.

Utilizing the librosa library, this system can extract several acoustic features, including duration (the length of all the speech in seconds), zero crossing rate (the frequency of the signal transitions, which are used to signify the speech type), root-mean-square energy (RMS energy, which is used to symbolize the signal power and loudness), fundamental frequency F0 (the measurement of the average pitch, which is used to depict the tonal change), and Mel-Frequency Cepstral Coefficients (MFCC) features, which are used to represent comprehensive audio signals, as shown in Figure 11.

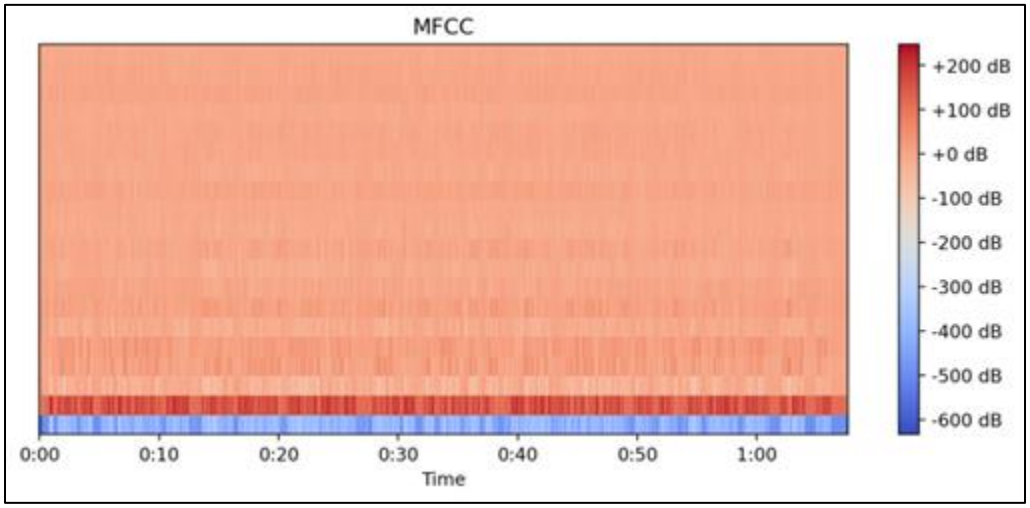

Figure 11.
Mel-Frequency Cepstral Coefficients.

The Mel-Frequency Cepstral Coefficients (MFCC) are a mandatory by-product of this academic audio-processing pipeline for analyzing the sound of a candidate interview. Here, a two-dimensional coordinate system is utilized, in which the temporal dimensions (0:00-1:00) are placed along the x-axis, and the amplitude values, log-rectified (200 dB to -600 dB), are placed along the y-axis to provide a spectro-temporal fingerprint of vocal properties. The sixty-second continuum, labeled at ten-second increments, is precisely aligned with the time of the extracted speech of a candidate. In contrast, the decibel scale represents vocal energy dynamism via a strict set of librosa-based signal processing algorithms.

The behaviorally significant patterns are shown in the spectral energy distribution. To start with, stable harmonic patterns in the first 30 seconds, with amplitudes ranging from -100 dB to +50 dB, indicate stable harmonic builds, which in turn translate to controlled articulation and rhythmic prosody characteristics, essential components of cognitive readiness. Secondly, harsh dials at 0:40-0:50 (-400 dB) reveal subvocal traces of filled pause markers, such as filler phonemes (e.g., "um" and "ah"), and pauses of breath that endanger the smooth continuation of the utterances. Thirdly, momentary glimpses above 200 dB at 0:15 and 0:35 are indicators of emphatic stress, using explosives (/p/, /t/) or exaggerated vowels, thereby confirming cases of confident self-presentation.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

Technically, such patterns are created based on 13 MFCC coefficients derived every 20-millisecond frame, with the referenced Libros implementation being documented. The levels of coefficient 0, total spectral energy, are modeled, and coefficients 1 to 12 are explained as cepstral envelopes via mel-scale filterbanks. The large troughs indicate the loss of energy in the upper bands (coefficients 5–12), which is typical of fricative steepening in the occurrence of disfluencies. On the other hand, the peaks represent the coherence of broadband power in lower coefficients (0–4), as expected in projection vocalization. This type of granular feature extraction allows for the assessment of communication clarity, whereby energy stability metrics like poise in Llama-3.2 are generated by evaluating the stability of poise and the number of paralinguistic discontinuities in the spectral domain of consistency found by the algorithm. Ultimately, this visualization serves not only as an acoustic transcript but also as a quantifiable behavioral phenotype within the multimodal assessment system.

Along with MFCCs, pitch contour analysis is also part of the prosodic analysis aspect, which quantifies tone differences, monitors speech rate to assess fluency, and detects pauses to identify hesitation and fillers (e.g., "um," "ah," "uh"). After the test results are transcribed, they are run through the language model of GROK API, Llama 3.2, as shown in Figure 12, to enable complex natural language comprehension, as required for candidate evaluation. This framework provides an overall assessment of communication, considering both the level of confidence, clarity, and overall presentation.

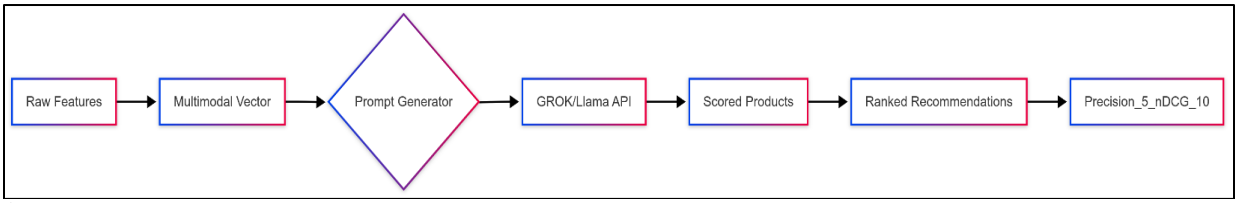

Figure 12.
System Recommendation Pipeline.

To further strengthen the results, the test-time augmentation and multimodal ablation study is presented in Table 8 and Table 9, respectively.

Table 8.
Test-Time Augmentation (TTA) Impact.

Condition	Accuracy (%)	ΔAccuracy	Inference Time (ms)	p-value
No TTA	83.2 ± 0.4	Baseline	12 ± 2	—
TTA (10x)	86.0 ± 0.3	+2.8	105 ± 5	<0.01

Note: *Performance and latency trade-off (10 augmentations per sample)*.

Table 9.
Multimodal Ablation Study.

System Configuration	F1-Score	ΔF1	Insights
Our System (FER + Gaze + Audio)	0.86	-	Optimal performance
Gaze Tracking	0.78	−0.08	Largest drop: Engagement metrics lost
Audio Analysis	0.81	−0.05	Reduced vocal hesitancy detection
FER	0.73	−0.13	Critical for emotional context
Late Fusion (Early Fusion)	0.79	−0.07	Confirms fusion strategy superiority

Note: *End-to-end system performance (F1-score) with modality removal*

4.4. Multimodal Feature Fusion and System Deployment

The system's approach is late fusion, where modality-specific data, such as facial expression recognition, gaze tracking, and audio analysis, are processed initially, and then the output is concatenated to generate a final candidate evaluation. This architecture ensures modularity, allowing each modality to be optimized separately. The final feature vector is constructed as follows: a seven-

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

dimensional emotion probability vector is calculated as the SoftMax output of FER, a four-dimensional vector is calculated due to gaze tracking, and a multidimensional audio-analysis feature vector is created when acoustic features and linguistic features are integrated. All extracted features are presented in a Streamlit web application in the form of comprehensive reports, including a percentage breakdown, benchmarking, and graphs that display the performance of all candidates across all modalities.

The current application is implemented as a web-based program built with the Streamlit library (Figure 13), which enables concurrent processing of streaming webcams and uploaded video recordings, providing immediate feedback on their performance and generating overall candidate evaluation documents. Training in minimal latency ensures that real-time analysis is suitable for scenarios such as live interviews. The total number of parameters is 14,883,399 (56.78 MB), with trainable parameters totaling 166,919 (652.03 KB).

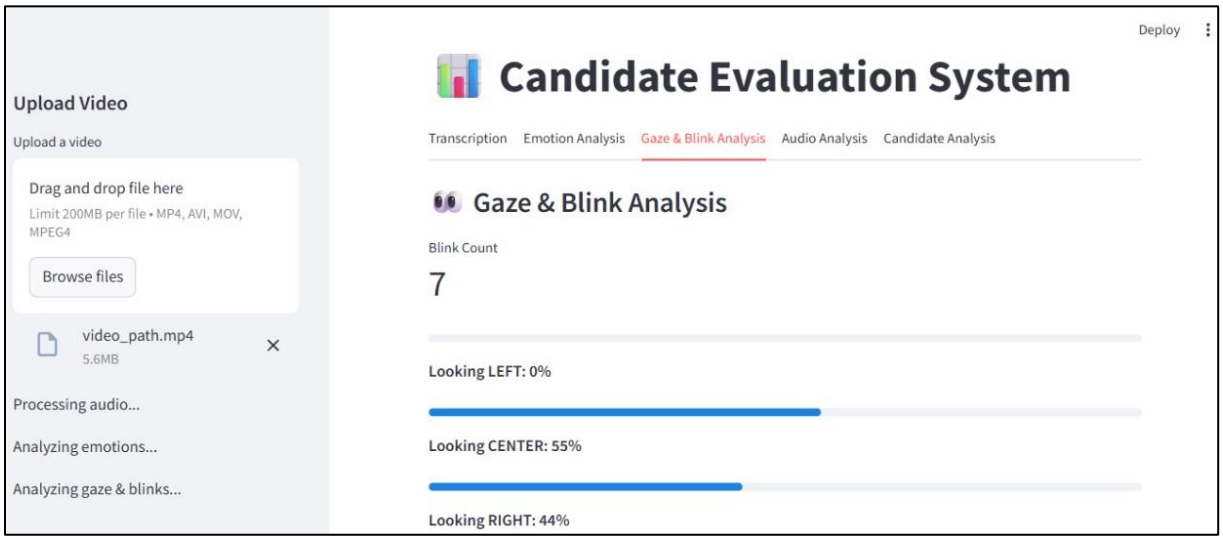

Figure 13.
Web-Based Candidate Evaluation System.

5. Discussion

The Interview Video Analysis System (IVAS) is an improvement on the traditional interview assessment as it operationalizes a multimodal AI-based framework. By conflating Facial Emotion Recognition (FER), Gaze Tracking, and Audio Analysis, IVAS provides an advanced behavioral assessment that is thoroughly accurate. The FER module stands out due to a Convolutional Neural Network (CNN) model of 22 layers, customized geometric/photometric augmentation, and inverse-frequency weighting; it has a classification accuracy of 86% on the FER-2013 dataset, surpassing comparison models like ResNet and VGGNet by wide margins. The gaze module measures the engagement of the interviewee by measuring the eye-contact duration, directional changes, and blink behavior based on the guesses of dlib landmarks, whereas the audio module retrieves paralinguistic features, pitch, and fluency, using Llama 3.2 and spectrogram. The late-fusion approach merges all these modalities in a Streamlit-based framework, where one can capture live interviews and recorded content in real-time. In such a system, structured and benchmarked reports, which incorporate behavioral markers, are generated, thus enabling a data-driven hiring process. The multimodal scheme also helps alleviate the shortcomings of the mislabeling of emotion, usually represented by such polysemy as sadness and fear, by cross-modal confirmation. Overall, IVAS provides an objective standard that can be scaled in assessing candidates, proving the effectiveness of combined AI modalities in derailing human bias toward enhancing recruitment rigor.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 10, No. 1: 525-543, 2026
DOI: 10.55214/2576-8484.v10i1.11626

6. Conclusion

IVAS is being developed on the premise of proceeding beyond the multimodal interview analysis stage to analyze facial expression recognition (FER), gaze detection, and acoustic analysis to clarify ambiguities that lie in the unimodal assessment. The FER module of IVAS consists of a 22-layer CNN yielding an accuracy of 86% and performs better than the already available shallower CNN-based ensemble techniques, as it can capture subtle features in realistic conditions. Data augmentation methods, such as flipping, rotation, and scaling, are used to address class imbalance in FER-2013, and the underrepresented occurrences of disgust were positively represented with the inverse frequency weighting method. This incorporation enhances the system's robustness and helps in the identification of the visual similarity between emotional states, such as surprise and fear, by gaze tracking through detecting directional focus patterns of the gaze, despite their distinct vocal patterns of hesitation or confidence, which are visibly inaudible through gaze. Nonetheless, despite these advances, there are still limitations. The inherent biases of the data, such as lighting variability and pose, can affect real-world generalizability, but such issues can be partially addressed by using augmentation algorithms. Additionally, for real-time inference, extensive hardware optimization is required to expand deployments. Ethical considerations, such as biases in algorithms used to interpret emotion, must also be addressed. Future research should focus on increasing dataset diversity, including cultural and contextual variations, and designing lightweight models suitable for deployment at the edge. Multimodal synergy can also be enhanced by incorporating quantum-inspired fusion architectures. As much as the development of IVAS demonstrates the transformative potential of AI in recruitment scenarios, it is vital to continue refining the system to achieve fairness and scalability.

Ethical Guidelines:

We confirm that all methods were carried out according to our organization's relevant guidelines and regulations and the Journal's Criteria.

Institutional Review Board Statement:

The committee of the Sindh Madressatul Islam confirmed that the organization approved all experimental protocols.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] M. Mukhopadhyay, S. Pal, A. Nayyar, P. K. D. Pramanik, N. Dasgupta, and P. Choudhury, "Facial emotion detection to assess learner's state of mind in an online learning system," in *Proceedings of the 2020 5th International Conference on Intelligent Information Technology*, 2020.

[2] R. Pereira *et al.*, "Systematic review of emotion detection with computer vision and deep learning," *Sensors*, vol. 24, no. 11, p. 3484, 2024. https://doi.org/10.3390/s24113484

[3] K. Semmelmann and S. Weigelt, "Online webcam-based eye tracking in cognitive science: A first look," *Behavior Research Methods*, vol. 50, no. 2, pp. 451-465, 2018. https://doi.org/10.3758/s13428-017-0913-7

[4] X. Zhu *et al.*, "RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers," *Information Fusion*, vol. 123, p. 103268, 2025. https://doi.org/10.1016/j.inffus.2025.103268

[5] J. Wang, M. Gao, W. Zhai, I. Rida, X. Zhu, and Q. Li, "Knowled ge generation and distillation for road segmentation in intelligent transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-13, 2025. https://doi.org/10.1109/TITS.2025.3577794

[6] M. Gao *et al.*, "Towards trustworthy image super-resolution via symmetrical and recursive artificial neural network," *Image and Vision Computing*, vol. 158, p. 105519, 2025. https://doi.org/10.1016/j.imavis.2025.105519

[7] I. Husain *et al.*, "Logic retaining power among genders and effect of time constraints on the performance of undergrads university students," *Journal of Mechanics of Continua and Mathematical Sciences*, vol. 16, no. 4, pp. 96–102, 2021. https://doi.org/10.26782/jmcms.2021.04.00008

[8] S. A. Inam *et al.*, "A novel deep learning approach for investigating liquid fuel injection in combustion system," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 32, 2025. https://doi.org/10.1007/s44163-025-00248-2

[9] X. Zhu *et al.*, "A client–server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods," *Computer Methods and Programs in Biomedicine*, vol. 260, p. 108564, 2025. https://doi.org/10.1016/j.cmpb.2024.108564

[10] M. Ur Rahim, M. Hussain, S. A. Inam, and H. Hashim, "Ignition behavior of supercritical liquid fuel in combustion system," *Journal of Mechanics of Continua and Mathematical Sciences*, vol. 16, no. 8, pp. 22–34, 2021. https://doi.org/10.26782/jmcms.2021.08.00003

[11] Y. Zhang, X. Wang, J. Wen, and X. Zhu, "WiFi-based non-contact human presence detection technology," *Scientific Reports*, vol. 14, no. 1, p. 3605, 2024. https://doi.org/10.1038/s41598-024-54077-x

[12] R. Wang, C. Guo, M. Shabaz, I. Rida, E. Cambria, and X. Zhu, "CIME: Contextual interaction-based multimodal emotion analysis with enhanced semantic information," *IEEE Transactions on Computational Social Systems*, pp. 1-11, 2025. https://doi.org/10.1109/TCSS.2025.3572495

[13] J. Zheng *et al.*, "Dynamic spectral graph anomaly detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 12, pp. 13410-13418, 2025. https://doi.org/10.1609/aaai.v39i12.33464

[14] R. Wang *et al.*, "Contrastive-based removal of negative information in multimodal emotion analysis," *Cognitive Computation*, vol. 17, no. 3, p. 107, 2025. https://doi.org/10.1007/s12559-025-10463-9

[15] S. Guo, Q. Li, M. Gao, X. Zhu, and I. Rida, "Generalizable deepfake detection via spatial kernel selection and halo attention network," *Image and Vision Computing*, vol. 160, p. 105582, 2025.

[16] A. J. Hong, D. DiStefano, and S. Dua, "Can CNNs accurately classify human emotions? A deep-learning facial expression recognition study," *arXiv preprint arXiv:2310.09473*, 2023. https://doi.org/10.48550/arXiv.2310.09473

[17] K. Sarvakar, R. Senkamalavalli, S. Raghavendra, J. S. Kumar, R. Manjunath, and S. Jaiswal, "Facial emotion recognition using convolutional neural networks," *Materials Today: Proceedings*, vol. 80, pp. 3560-3564, 2023. https://doi.org/10.1016/j.matpr.2021.07.297

[18] L. Balcombe and D. De Leo, "Human-computer interaction in digital mental health," *Informatics*, vol. 9, no. 1, p. 14, 2022.

[19] M. Langote *et al.*, "Human-computer interaction in healthcare: Comprehensive review," *AIMS Bioengineering*, vol. 11, no. 3, pp. 343–390, 2024.

[20] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021. https://doi.org/10.3390/s21041249

[21] A. Rahujo, D. Atif, S. A. Inam, A. A. Khan, and S. Ullah, "A survey on the applications of transfer learning to enhance the performance of large language models in healthcare systems," *Discover Artificial Intelligence*, vol. 5, p. 90, 2025. https://doi.org/10.1007/s44163-025-00339-0

[22] Y. Li, Z. Zhong, F. Zhang, and X. Zhao, "Artificial intelligence-based human–computer interaction technology applied in consumer behavior analysis and experiential education," *Frontiers in Psychology*, vol. 13, p. 784311, 2022. https://doi.org/10.3389/fpsyg.2022.784311

[23] S. Karanchery and S. Palaniswamy, "Emotion recognition using one-shot learning for human-computer interactions," in *2021 International Conference on Communication, Control and Information Sciences (ICCISc) (Vol. 1, pp. 1-8). IEEE*, 2021.

[24] S. Hussain *et al.*, "A discriminative level set method with deep supervision for breast tumor segmentation," *Computers in Biology and Medicine*, vol. 149, p. 105995, 2022. https://doi.org/10.1016/j.compbiomed.2022.105995

[25] S. A. Inam, D. Iqbal, H. Hashim, and M. A. Khuhro, "An empirical approach towards detection of tuberculosis using deep convolutional neural network," *International Journal of Data Mining, Modelling and Management*, vol. 16, no. 1, pp. 101-112, 2024. https://doi.org/10.1504/IJDMMM.2024.136232

[26] L. Zahara, P. Musa, E. P. Wibowo, I. Karim, and S. B. Musa, "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi," in *2020 Fifth International Conference on Informatics and Computing (ICIC) (pp. 1-9). IEEE*, 2020.

[27] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with CNN ensemble," in *2016 International Conference on Cyberworlds (CW) (pp. 163-166). IEEE*, 2016.

[28] E. M. Onyema, P. K. Shukla, S. Dalal, M. N. Mathur, M. Zakariah, and B. Tiwari, "Enhancement of patient facial recognition through deep learning algorithm: ConvNet," *Journal of Healthcare Engineering*, vol. 2021, no. 1, p. 5196000, 2021. https://doi.org/10.1155/2021/5196000

[29] Khan, A. A., Laghari, A. A., Almansour, H., Jamel, L., Hajjej, F., Estrela, V. V., ... & Ullah, S. (2025). Quantum computing empowering blockchain technology with post quantum resistant cryptography for multimedia data privacy preservation in cloud-enabled public auditing platforms. Journal of Cloud Computing, 14(1), 43.

[30] A. Bulling and M. Wedel, *"Pervasive eye-tracking for real-world consumer behavior analysis. In A handbook of process tracing methods."* New York: Routledge, 2019.

[31] khan, A. A., ghodhbani, R., Alsufyani, A., Alsufyani, N., & Mohamed, M. A. (2025). Leveraging blockchain-integrated explainable artificial intelligence (XAI) for ethical and personalized healthcare decision-making: a framework for secure data sharing and enhanced patient trust. The Journal of Supercomputing, 81(15), 1353.

[32] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," presented at the 8th European Conference on Speech Communication and Technology (Eurospeech 2003) (pp. 125–128). ISCA, 2003.

[33] J. S. Murthy and G. Siddesh, "A smart video analytical framework for sarcasm detection using novel adaptive fusion network and SarcasNet-99 model," *The Visual Computer*, vol. 40, pp. 8085-8097, 2024. https://doi.org/10.1007/s00371-023-03224-y

[34] J. S. Murthy and G. Siddesh, "Multimedia video analytics using deep hybrid fusion algorithm," *Multimedia Tools and Applications*, vol. 84, pp. 14167-14185, 2025. https://doi.org/10.1007/s11042-024-20184-0

[35] J. S. Murthy, S. G. Matt, S. K. H. Venkatesh, and K. R. Gubbi, "A real-time quantum-conscious multimodal option mining framework using deep learning," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, p. 1019, 2022. https://doi.org/10.11591/ijai.v11.i3.pp1019-1025

[36] H. Ghazouani, "Challenges and emerging trends for machine reading of the mind from facial expressions," *SN Computer Science*, vol. 5, p. 103, 2023. https://doi.org/10.1007/s42979-023-02447-z

[37] Khan, A. A., Shaikh, A. K., Alroobaea, R., Baqasah, A. M., Alsafyani, M., Alsufyani, H., & Laghari, A. A. (2025). Blockchain-enabled secure Internet of Medical Things (IoMT) architecture for multi-modal data fusion in precision cancer diagnosis and continuous monitoring. Journal of Cloud Computing, 14(1), 1-21.

[38] Khan, A. A., Shaikh, A. K., Alsufyani, A., Alsufyani, N., Laghari, A. A., & Mohamed, M. A. (2025). Multi-objective Particle Swarm Optimization for Sustainable Industrial Operations: Energy and Cost Perspectives. International Journal of Computational Intelligence Systems, 18(1), 273.

[39] I. Goodfellow, D. Erhan, A. Luc Carrier, A. Courville, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Proceedings of the International Conference on Machine Learning (ICML) Workshop on Deep Learning and Unsupervised Feature Learning*, 2013.

[40] O. C. Oguine, K. J. Oguine, H. I. Bisallah, and D. Ofuani, "Hybrid Facial Expression Recognition (FER2013) model for real-time emotion classification and prediction," *BOHR International Journal of Internet of Things, Artificial Intelligence and Machine Learning*, vol. 1, no. 1, pp. 56–64, 2022. https://doi.org/10.54646/bijiam.2022.09

[41] Khan, A. A., Alsufyani, A., Alsufyani, N., Mohamed, M. A., & Ullah, S. (2025). Cutting-edge optimized multi-source data fusion for trusted execution and management of blockchain transactions on the internet of medical things (IoMT) with machine learning. Scientific Reports.

[42] K. Iwauchi *et al.*, "Eye-movement analysis on facial expression for identifying children and adults with neurodevelopmental disorders," *Frontiers in Digital Health*, vol. 5, p. 952433, 2023. https://doi.org/10.3389/fdgth.2023.952433