PrVia

Automated Evaluation of Pre-Recorded Video Interviews

By: Ammar Mohamed Mohamed Samy Mohamed Ashraf
Youssef Tamer Yomna Mohamed Nadine Haitham

Supervised by: Dr. Mona Abdelazim , TA. Aya Nasser
Faculty of Computer and Information Sciences - Ain Shams University

Abstract

PRVIA is an AI-powered system that automates the evaluation of pre-recorded job interviews by analyzing candidates' speech, language, facial expressions, and personality traits. Delivered via a web platform, it generates comprehensive scores in under two minutes, helping reduce evaluation bias, match human performance, and streamline large-scale hiring with fairness and consistency.

Introduction

With the rise of remote hiring, pre-recorded video interviews have become standard but they bring challenges in evaluating communication, confidence, and response quality at scale. Manual reviews are often slow and subjective. PRVIA addresses these limitations by introducing an AI-powered evaluation system that analyzes each interview holistically capturing both verbal and non-verbal traits to deliver fast, consistent, and fair assessments.

System Architecture

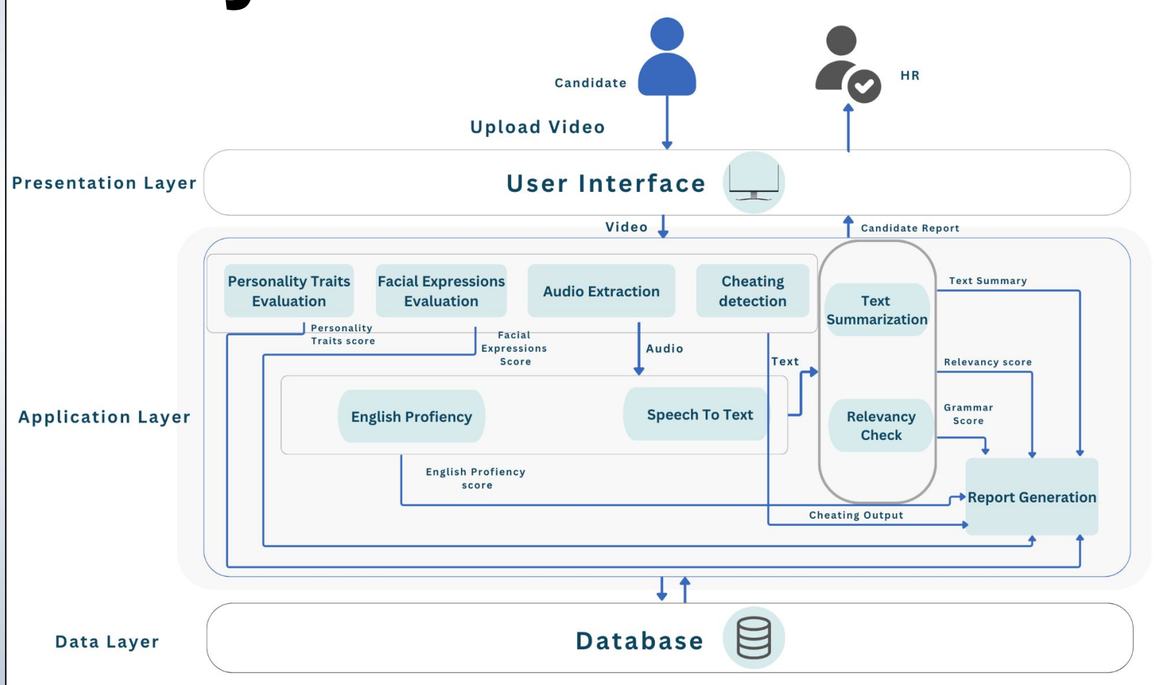


Fig 1 System Architecture

PRVIA processes interview videos through five core modules—**facial analysis**, **audio scoring**, **speech-to-text**, **text evaluation**, and **cheating detection**—to generate an automated candidate report combining verbal and non-verbal insights.

Methods

Video Analysis: Candidate videos are processed using **MTCNN** for face detection. **X3D** extracts spatiotemporal facial features to predict Big Five personality traits. **BlazeFace** and **MediaPipe Face Mesh** monitor eye movements frame-by-frame; sustained off-screen gaze (>3s) is flagged as potential cheating. For emotional behavior analysis, **DeepFace** and **MediaPipe** are used to extract and track facial emotions throughout the video.

Audio Processing: Audio is extracted using MoviePy and transcribed with Whisper. Wav2Vec2 and ModernBERT features are fused via cross-attention and passed through a Transformer encoder to assess pronunciation (fluency, accuracy, prosody, completeness).

Text Evaluation:
Transcribed responses are summarized using **BART-Large-CNN** and **Gemini**. Semantic relevance between candidate answers and questions is assessed using prompt-based **Gemini API** classification.

Text-Based Personality Prediction:
Transcripts are passed through a **3-layer MLP** trained on the Essays dataset to infer personality traits from language, complementing video-based trait scoring.

System Integration: All module outputs are aggregated and served via a FastAPI backend and PostgreSQL database, with results presented through an interactive React web interface for HR review.

Results

Personality traits from text were predicted from the Essays Dataset (2,468 essays) using BERT-Large ([CLS] token) and a 3-layer MLP, achieving strong results with 10-fold cross-validation.

Model	OPN	CON	AGR	NEU	EXT
Bert-Large+MLP	0.68	0.65	0.63	0.61	0.62

Text Summarization:A custom dataset of 800 interview transcripts and human-written summaries was created to train and evaluate models. BART-Large-CNN and FLAN-T5 were assessed using **ROUGE**, **BLEU**, and **BERTScore**. While ROUGE and BLEU scores were limited, BERTScore demonstrated strong semantic alignment with reference summaries.

Model	Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum	BLEU Score	AverageBERT Score F1
BART-Large	0.4930	0.2686	0.4345	0.4344	0.1738	**0.9110**
FLAN-T5 Large	0.3530	0.2164	0.3425	0.3423	0.1117	0.8996
Gemini 2.0 Flash	0.4530	0.2373	0.4073	0.4072	0.1047	0.9079

We used the First Impressions V2 dataset (10,000 ~15s clips of individuals speaking) to predict Big Five **personality traits from video:** Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.

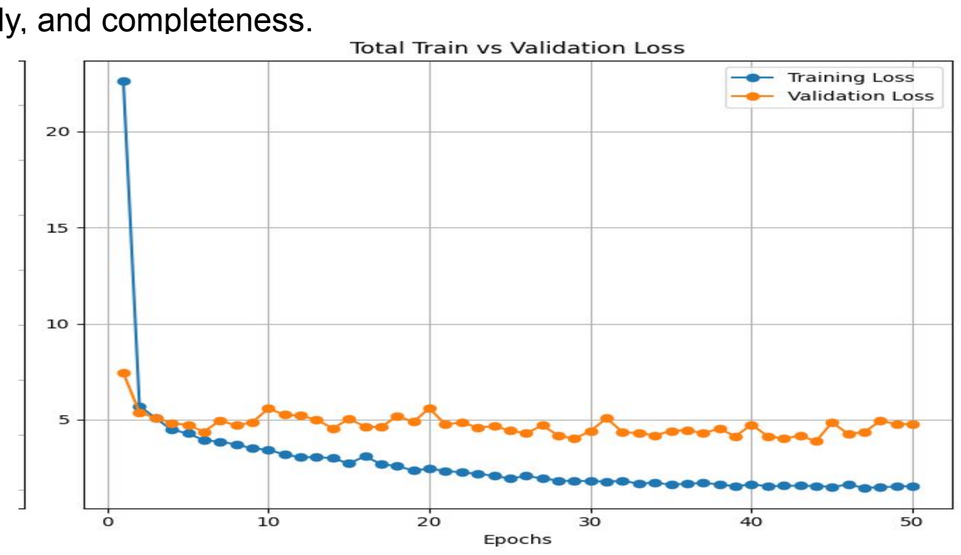
Model	Mean Accuracy	Mean Squared Error
X3D-S	0.9051	0015

For **Emotion Detection from video** DeepFace was validated on 350 samples from **CAER dataset** across seven emotion classes, showing strong visual emotion classification.

Library	Frame Success Rate	Emotion Accuracy
DeepFace	90%	80%

Cheating Detection: MediaPipe Face Mesh tracks gaze direction frame-by-frame; sustained deviation outside the [0.35–0.65] range for over 3s is flagged as cheating.

The **audio model** is Trained on SpeechOcean762 using Wav2Vec2 to extract acoustic features and ModernBERT to encode the expected transcript. A cross-attention layer fuses both modalities, followed by a Transformer encoder and MLP heads to predict fluency, accuracy, prosody, and completeness.



Conclusions

PRVIA combines audio, text, and video models to evaluate interviews accurately and consistently. Strong results,highlight its potential to support faster, fairer, and more scalable hiring.

Bibliography

1. Kassab, K. and Kashevnik, A., 2024, April. Novel Framework for Job Interview Processing Automation Based on Intelligent Video Processing. In *2024 35th Conference of Open Innovations Association (FRUCT)* (pp. 336-342). IEEE.

2. Naim, I., Tanveer, M.I., Gildea, D. and Hoque, M.E., 2016. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, *9*(2), pp.191-204.