
AI-Augmented Recruitment Interviews: A Feasibility Study of Real-Time Facial Video Analysis under Recruitment-Like Conditions

Submitted 23/04/25, 1st revision 09/05/26, 2nd revision 20/05/26, accepted 29/06/26

Przemysław Jaworski¹, Miłosz Makowski², Maciej Pondel³

Abstract:

Purpose: This paper examines whether real-time facial video analysis can function as an interpretable decision-support layer in technology-mediated recruitment interviews. It addresses a gap between research on AI-supported recruitment, which mainly emphasises efficiency and process standardisation, and research on affect-related video analysis, which typically prioritises model performance outside realistic organisational hiring contexts.

Design/Methodology/Approach: The study adopts a staged research design combining: (1) laboratory grounding through comparison of optically derived facial activity traces with EMG-related measures, (2) development of a remote-capable and interview-compatible capture procedure, and (3) prototype evaluation in recruitment-like scenarios. The analytical pipeline combines convolutional neural network-based landmark detection, FLAME-based facial reconstruction, and FACS-consistent descriptors to transform facial activity recorded from standard video into biosignal-like temporal traces. Recruitment-oriented evaluation was conducted on a sample of 75 participants, with system outputs compared against the ratings of a single expert observer.

Findings: The results indicate prototype-level feasibility rather than validated recruitment effectiveness. The strongest agreement with expert assessment was observed for emotion-related outputs (85%) and stress-related inference (80%), while lower agreement was found for interactional responsiveness (60%) and nonverbal behaviour based on microexpression analysis (55%). Aggregate agreement reached 75%. The prototype also proved operationally feasible within a structured interview workflow, requiring up to 20 seconds of behavioural observation for selected constructs and generating outputs within near-real-time latency. However, the findings do not establish predictive validity for live hiring decisions.

Practical Implications: The proposed approach may support structured recruiter observation by providing auditable, standardised, and temporally interpretable behavioural indicators during selected interview segments. Its use should remain strictly supportive, human-supervised, and bounded by governance safeguards relating to privacy, consent,

¹Corresponding author, Shen.AI OÜ, Tallinn, Estonia, ORCID: 0009-0009-1849-193X, e-mail: pjaworski@shen.ai;

²Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznań, Poland and Shen.AI OÜ, Tallinn, Estonia

³Corresponding author, Institute of Information Systems in Economics, Wrocław University of Economics and Business, Wrocław, Poland, ORCID: 0000-0002-1978-6571, e-mail: maciej.pondel@ue.wroc.pl;

transparency, and fairness.

Originality/Value: The paper contributes a recruitment-oriented framework that links physiological grounding, remote-capable data collection, and prototype-level expert-concordance testing. Its originality lies in treating facial video analysis not as a tool for autonomous candidate judgement, but as a cautious and reviewable analytical layer designed to support, rather than replace, human decision-making in recruitment.

Keywords: Personnel selection, human resource analytics, affective computing, behavioural assessment, decision support, technology-mediated screening.

JEL codes: M12, M15, M51.

Paper type: Research paper

Funding: This publication was prepared as part of the project entitled “Research and Development Work on a Prototype of a Generic Personal Contact Matters Platform Supported by Artificial Intelligence”, implemented under the Regional Operational Programme for the Dolnośląskie Voivodeship for 2014–2020 (call no. RPDS.01.02.01-IP.01-02-372/19), and co-financed by the European Union through the European Regional Development Fund.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this manuscript, the authors used ChatGPT Business, a generative AI tool developed by OpenAI, solely for language editing purposes, including improvement of grammar, clarity, style, and readability of the text. The tool was not used to generate research data, conduct data analysis, create results, formulate conclusions, or make substantive scientific decisions. All AI-assisted edits were reviewed, verified, and approved by the authors. The authors remain fully responsible for the accuracy, originality, integrity, and final content of the manuscript.

1. Introduction

Recruitment interviews remain one of the most influential selection instruments because they combine competence signalling, interpersonal evaluation, and contextual judgement in a single interaction.

Yet interviews also concentrate multiple sources of uncertainty: candidates engage in impression management; interviewers rely on bounded attention; and organisational constraints (time-to-hire, interviewer workload, and inconsistent interviewing skills) can weaken standardisation across candidates and roles. In this setting, the practical objective is not to “read” candidates’ internal states, but to improve the quality and consistency of behavioural evidence that informs hiring decisions—particularly for early-stage screening and competency-focused assessment.

Over the last decade, recruitment has become increasingly technology-mediated. Video interviewing and other digital selection technologies are now embedded in many organisational pipelines, enabling geographically distributed hiring, faster scheduling, and scalable screening.

Evidence from employer surveys indicates that video interviews and recorded video clips used for screening belong to the portfolio of technologies actively used in recruitment processes, alongside online tests and applicant tracking systems. At the same time, the use of AI in recruitment is described as “limited but increasing,” with reported growth in applications such as AI-supported candidate shortlisting, chatbot-based candidate communication, and—crucially for this paper—AI analysis of interview responses for selection purposes (Hogarth and McCartney, 2024).

The migration of interviews into remote formats has methodological consequences for assessment. In a face-to-face interview, evaluators can integrate a wide range of behavioural cues (posture, spatial behaviour, gaze and joint attention in shared space, and subtle turn-taking dynamics) within a richer situational context. In remote interviews, however, evaluators often have access to a constrained behavioural channel: a compressed audiovisual stream, typically limited to a head-and-shoulders view and a single microphone.

This constraint can reduce observability of behavioural context and increase reliance on facial and vocal cues that survive the medium (for example, facial dynamics, prosody, and conversational timing). Research on employment interviewing supports the premise that visual and vocal cues shape interviewer perceptions and hiring recommendations, often through trait inferences that then mediate performance evaluations (Martín-Raugh *et al.*, 2023).

The practical implication is straightforward: when the medium restricts behavioural information, the remaining cues may carry disproportionate weight—making both measurement quality and interpretability particularly important.

Prior work has already shown that video interviews can be treated as a source of structured behavioural evidence, with facial Action Units extracted from interview recordings used to analyse candidate-related behavioural patterns and to support HR decision processes (de Jesus *et al.*, 2023).

Within this environment, affective computing has been proposed as a candidate decision-support layer for recruitment: systems that quantify aspects of candidates’ affective dynamics (for example, indicators consistent with stress, engagement, or positive/negative activation) could, in principle, complement conventional evaluation instruments.

However, robust video-based emotion recognition is still difficult to operationalise in realistic interviewing conditions. Optical methods may degrade under head

movements, variable lighting, differences in camera distance, and other uncontrolled factors; additionally, sensitivity to low-amplitude facial dynamics (including micro-movements) is often limited.

This paper addresses the recruitment-specific research problem of how to extract reliable and interpretable affect-related signals from facial video in a way that is compatible with remote, technology-mediated interviews. The proposed approach is grounded in the idea that recruitment-oriented decision support benefits from signals that are (i) measurable from standard video, (ii) temporally informative (time-series rather than single labels), and (iii) interpretable in behavioural terms.

To reduce the conceptual gap between laboratory-grade physiological measurement and deployable optical systems, the project developed a framework that records facial activity from video and represents it as biosignal-like time series analogous to electromyography (EMG).

In laboratory work, facial activity was analysed using a morphable face model (FLAME) and FACS-based indicators, and compared with EMG recordings from facial muscles commonly implicated in affective expression (zygomaticus major, orbicularis oculi, corrugator supercilii).

The comparison indicated correlations sufficient for detecting medium and salient facial movements, whereas very low-amplitude micro-movements were close to the detectability threshold in the current implementation.

These limitations are especially salient in remote interviews, where candidates' devices and environments are heterogeneous, and where even minor disruptions (frame drops, compression artefacts, or unstable camera placement) can affect facial measurement.

To move beyond laboratory evidence toward recruitment-relevant feasibility, the framework was organised into staged evaluation and tool development. First, controlled recordings supported physiological alignment (optical versus EMG).

Second, a remote-capable application was designed to execute a standardised protocol (including calibration and repeated facial video capture), enabling collection under conditions compatible with distributed participation and remote administration.

Third, a portable prototype device was constructed to capture and process facial video at high resolution, with the intention of supporting near-real-time inference and operator-facing inspection.

In the recruitment domain, the system was conceptually positioned as support for specific interview-relevant stages and observable behavioural dimensions. The

recruitment scenarios were designed around: (a) preliminary interviews, where affect-related signals could complement evaluators' judgements of engagement and visible confidence; (b) interactional responsiveness during interview prompts, particularly in situations requiring turn-taking, nonverbal alignment, and socially appropriate behavioural regulation; and (c) stress management, where observed affective dynamics under challenging prompts may relate to coping and self-regulation.

This framing is intentionally conservative: the objective is not to infer stable personality or "truthfulness," but to quantify behavioural dynamics that are already informally attended to in interviews, and to provide these measurements in a consistent, reviewable form.

Empirical testing in recruitment-like scenarios was conducted on a sample of 75 participants, with algorithm outputs compared against an expert observer's ratings. Aggregate agreement with expert assessment was reported at 75% across the set of tasks, while the evaluation documentation explicitly notes justified concerns regarding objectivity and ecological validity.

The same source also states that fully representative evidence would require studies involving real candidates engaged in genuine recruitment processes, but that legal and ethical objections associated with near-realistic testing conditions led to discontinuation of further exploration in recruitment scenarios at that stage.

These observations delimit the intended scope of claims in the present article: we treat the framework as a recruitment-oriented decision-support prototype whose value must be assessed jointly in terms of technical performance, robustness under remote constraints, and governance feasibility (privacy, consent, fairness, and auditability).

Against this background, the paper makes three principal contributions. First, it proposes a recruitment-oriented framework for real-time video emotion recognition that transforms facial activity captured from standard video streams into interpretable biosignal-like time series grounded in FLAME and FACS-based representation. Second, it reports a staged research design combining laboratory alignment of optical and EMG-related measures, remote-capable data collection, and prototype-based testing in recruitment-relevant scenarios.

Third, it evaluates the practical value and limitations of this approach for interview assessment, with particular attention to interpretability, robustness, and governance constraints in AI-supported hiring. The remainder of the paper is organised as follows. Section 2 presents the literature review, covering prior work on AI-supported recruitment, video emotion recognition, and the ethical challenges of affective computing in selection contexts. Section 3 describes the methodology, including the signal representation framework, research design, prototype

assumptions, and evaluation procedure. Section 4 reports the research results and discussion, focusing on system performance, agreement with expert assessment, and the implications of deployment in recruitment settings. Section 5 concludes with conclusions, proposals, and recommendations for the responsible development and future validation of AI-augmented interview tools.

2. Literature Review

The literature relevant to this study lies at the intersection of three research streams: AI-supported recruitment, video-based affect recognition, and the ethical governance of affective computing. Together, these streams suggest that while digital interviewing and automated behavioural analysis are becoming more widespread, the use of affect-related AI in recruitment remains methodologically demanding and ethically contested (Hu, 2023; Yi *et al.*, 2025; Barker *et al.*, 2025).

In the recruitment literature, artificial intelligence is increasingly framed as a tool for augmenting rather than replacing human judgement. Its most common applications include candidate sourcing, CV screening, automated communication, online assessment, and support for technology-mediated interviewing. Evidence from organisational practice suggests that AI-supported recruitment may improve efficiency, reduce administrative workload, and increase process standardisation, especially when digital tools are embedded in integrated recruitment pipelines (Hu, 2023).

At the same time, these operational gains do not eliminate persistent concerns regarding fairness, privacy, transparency, and accountability in personnel selection. Accordingly, the literature indicates that the value of AI in recruitment depends not only on predictive performance, but also on whether its use can be embedded within governance mechanisms that preserve procedural legitimacy and human oversight (Hu, 2023).

A related body of research examines AI interviewing systems that combine natural language processing, computer vision, and machine learning to structure interview evidence. Such systems typically analyse verbal content, vocal characteristics, and visible behavioural cues in order to support candidate evaluation, especially in early-stage screening. The underlying premise is that automated analysis may improve consistency and scalability.

However, many studies in this area advance strong performance claims while paying less attention to ecological validity, interpretability, and the operational constraints of realistic hiring contexts. This matters because interview behaviour is inherently contextual, role-dependent, and shaped by the interactional setting itself.

These limitations become even more salient in remote and technology-mediated interviews. Compared with face-to-face interviews, remote interviews constrain the

behavioural channel available to assessors, often reducing observation to a head-and-shoulders video stream and a single audio channel. Under such conditions, the cues that remain observable may carry disproportionate weight in interviewer impressions, especially facial dynamics, gaze-related behaviour, and vocal delivery.

This increases the practical importance of measuring such cues reliably, but also raises the risk of over-interpreting signals that are weak, context-dependent, or only partially observable. For this reason, recent scholarship increasingly suggests that AI-derived behavioural indicators should be treated as structured support for human judgement rather than as autonomous evidence of candidate suitability.

The second major research stream addresses the technical foundations of video emotion recognition. Recent review work shows that the field has evolved from handcrafted visual and audiovisual features toward end-to-end neural architectures, especially multimodal and transformer-based approaches. At the same time, major technical challenges remain unresolved. Emotions are temporally dynamic, context-sensitive, and often ambiguously labelled, while available datasets frequently simplify emotional states into coarse categories that do not reflect real interactional complexity.

Yi *et al.* (2025) emphasise that current video emotion recognition systems still struggle with the quality of emotional representations, dataset limitations, and efficient multimodal fusion. These issues are directly relevant to interview settings, where affective responses are embedded in task performance and social interaction rather than expressed as isolated and clearly classifiable displays.

A consistent finding in the literature is that multimodal fusion generally outperforms unimodal analysis when the objective is to maximise recognition performance across heterogeneous real-world conditions. Emotion is expressed through multiple partially interdependent channels, including facial movement, speech, prosody, body posture, gesture, and contextual information.

Transformer-based approaches integrating visual, audio, and textual features have shown that cross-modal learning improves recognition quality by capturing dependencies that single-modality systems cannot model effectively (Le *et al.*, 2023). Similarly, cross-modal audio-video fusion models demonstrate that performance improves when systems learn both modality-specific salience and inter-modal relationships rather than relying on simple feature concatenation (Mocanu *et al.*, 2023). In recruitment interviews, therefore, facial behaviour alone cannot provide a complete account of candidate meaning.

However, this observation does not make bounded unimodal approaches irrelevant in organisational settings. In technology-mediated interviews, access to behavioural channels is uneven and operationally constrained. Audio quality may vary, spoken content introduces additional language- and context-dependence, and body-related

information is frequently incomplete in head-and-shoulders video. Under such conditions, facial video remains one of the most consistently available and standardisable observational channels, even if it does not exhaust the behavioural meaning of the interview.

From this perspective, the question is not whether facial analysis alone can replace multimodal assessment, but whether it can provide a restricted and interpretable layer of affect-related evidence that is operationally usable within a structured, human-led interview process.

Within the visual domain, recent work further shows that facial expression should not be treated in isolation from broader visible behaviour. Wei *et al.* (2024) demonstrate that combining facial expression with body gesture improves video emotion recognition, particularly when models capture both local spatio-temporal variation and longer temporal dependencies.

Even so, this line of research also underlines an important methodological constraint for remote interviewing: webcam-based analysis typically captures only a partial behavioural field. As a result, any inference drawn from interview video is necessarily based on incomplete behavioural information. This limitation is not peripheral to the present study; rather, it defines the scope within which a facial-video-based analytical layer can be meaningfully evaluated.

Another important strand of research concerns the relationship between physiological and video-based representations of emotion. Studies using EEG and wearable sensing often treat physiological signals as a more direct or less consciously controlled channel of affective response than overt behaviour (Wu *et al.*, 2023; Saganowski *et al.*, 2023).

Although such methods are impractical in routine recruitment interviews, they remain relevant as a conceptual reference point. In particular, they highlight the value of temporally resolved signal analysis and the importance of interpretability when linking observable data to affect-related states.

This is significant for the present study, which does not claim that facial video is equivalent to physiological measurement, but instead proposes a framework in which facial activity is represented as biosignal-like temporal traces. In this sense, the study occupies an intermediate position between computer-vision-based affect recognition and physiologically grounded measurement approaches (Wu *et al.*, 2023; Saganowski *et al.*, 2023).

A more recruitment-specific point of reference is provided by prior work showing that facial Action Units extracted from interview video may be analysed as temporally structured signals and linked to HR-relevant behavioural assessment. De Jesus *et al.* (2023), for example, demonstrated that AU intensities derived from

candidate video interviews can be modelled as multivariate time series to support behavioural-profile analysis.

Nevertheless, this evidence remains methodologically partial. It does not yet establish a recruitment-oriented framework that combines temporal interpretability, operational feasibility in remote interviewing, and bounded decision-support logic rather than opaque end-point classification.

The third research stream focuses on ethics and governance. This literature argues that emotion-recognition technologies may generate substantial risks when used in high-stakes environments such as healthcare, education, surveillance, and employment. Barker *et al.* (2025) identify privacy intrusion, emotional profiling, demographic bias, observer bias, and weak consent mechanisms as recurrent concerns in emotion-recognition research. They argue that such systems require safeguards including transparency, informed consent, data minimisation, and algorithmic accountability (Barker *et al.*, 2025).

These concerns are especially important in recruitment, where classification errors or unjustified inferences may affect access to employment and where candidates often have limited capacity to challenge opaque algorithmic outputs.

Taken together, the literature reveals a clear research gap. Recruitment studies tend to focus on efficiency, scalability, and process innovation, whereas emotion-recognition studies focus primarily on modelling performance, multimodal fusion, and benchmark datasets. Considerably less attention has been paid to recruitment-specific systems that are at the same time interpretable, operationally feasible in remote interviews, and cautious in the claims they make.

The present study addresses this gap not by proposing a full multimodal interview-analysis system, but by examining whether facial activity extracted from standard video can be transformed into temporally structured, behaviourally interpretable signals that function as a limited decision-support layer under remote-interview constraints.

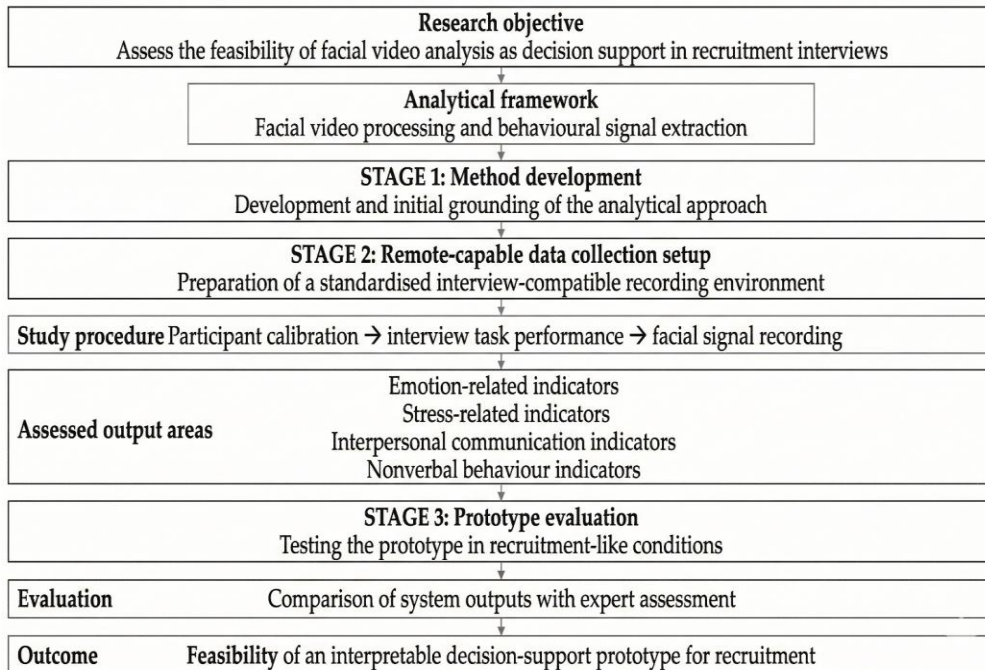
Its contribution lies in the recruitment-specific orientation of the framework, its emphasis on auditable temporal signals rather than opaque end-point classification, and its staged validation logic spanning physiological grounding, remote-capable data collection, and prototype-level testing in recruitment-like conditions.

3. Research Methodology

This study employed a staged, recruitment-oriented research design to assess the feasibility of facial video analysis as decision support in technology-mediated recruitment interviews. The methodological objective was intentionally limited. The proposed system was not examined as a substitute for interviewer judgement or as

an autonomous hiring mechanism, but as an interpretable analytical layer designed to structure behavioural evidence derived from facial video. In this sense, the study addressed whether facial video processing can generate reviewable indicators that may support more consistent assessment in selected interview situations. The overall research procedure adopted in the study is presented in **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε..**

Figure 1. General research procedure of the study.



Source: Own elaboration.

The analytical framework was based on facial video processing and behavioural signal extraction. At the conceptual level, the system was designed to transform observable facial activity into structured temporal representations that could be interpreted in relation to interview-relevant behaviour.

This design choice reflected the practical conditions of remote and hybrid interviews, in which the behavioural channel available to assessors is narrower than in face-to-face interaction and facial dynamics consequently acquire greater evaluative importance. Rather than relying exclusively on opaque end-point classifications, the framework prioritised interpretable outputs that could be inspected and compared with external assessment.

At the lower level of visual processing, the analytical pipeline relies on convolutional neural network-based face analysis to detect and track a set of 68 facial landmarks, covering anatomically informative regions such as the eyebrows,

eyes, nose, mouth, and jawline.

This landmark structure provides the geometric basis for face alignment, head-pose estimation, localisation of facial muscle regions, and frame-wise tracking of visible facial dynamics. On this basis, the pipeline applies parametric facial reconstruction using FLAME and subsequently translates the resulting trajectories into FACS-consistent descriptors. In line with the Facial Action Coding System developed by Paul Ekman and Wallace V. Friesen, facial activity is represented through observable movement patterns encoded as Action Units rather than treated as direct evidence of latent psychological states (Ahmed Khan *et al.*, 2025).

The use of automated facial coding in this context is also consistent with prior evidence indicating that OpenFace may provide a more sensitive and precise measure of facial behaviour than manual FACS coding in predictive analytical settings (Pfeffer, 2023).

In the project implementation, the downstream emotion-recognition component was supported by a convolutional model operating on landmark-derived representations, including keypoint heatmaps and AU-related descriptors, thereby preserving an interpretable link between raw video, facial geometry, and behaviourally meaningful facial signals (Ahmed Khan *et al.*, 2025).

The study was organised into three sequential stages. The first stage concerned method development and initial grounding of the analytical approach. Its purpose was to establish the technical basis for extracting behaviour-related information from facial recordings and to define the operational boundaries of the method under more controlled conditions.

The second stage focused on the preparation of a remote-capable, interview-compatible recording environment. This stage addressed the procedural requirements of technology-mediated assessment, including calibration, standardised capture conditions, and repeatable facial signal acquisition across participants.

The third stage consisted of prototype evaluation in recruitment-like conditions and represented the main empirical component of the study. Here, the prototype was examined not only in terms of analytical output, but also with regard to its compatibility with an interview-oriented workflow.

A standardised study procedure was applied throughout the evaluation phase. Each session began with participant calibration, followed by interview task performance and facial signal recording. Calibration was used to initialise the facial capture process and improve the stability of subsequent analysis.

The task phase was structured to elicit behaviour relevant to interview settings, while video data were collected for further processing. This ensured procedural consistency and made it possible to compare system-generated outputs with external

expert assessment.

The outputs examined in the study were grouped into four analytical areas: emotion-related indicators, stress-related indicators, interactional responsiveness indicators, and nonverbal event indicators. This operationalisation was deliberately conservative.

The system was not assumed to provide direct access to hidden psychological states or stable personality traits; rather, it was used to derive structured indicators from observable facial behaviour within specific interview prompts. Accordingly, the interpersonal dimension addressed in the study refers to prompt-specific visible interactional cues rather than trait-level competence or personality inference.

Evaluation was conducted by comparing the outputs generated by the system with expert assessment. This comparison constituted the principal validation logic of the recruitment-oriented phase of the research. The aim was to determine whether the prototype produced outputs that were sufficiently aligned with external human interpretation to justify treating it as a feasible support instrument in recruitment-like assessment. In addition to analytical correspondence, the study also considered the practical usability of the prototype within a structured interview procedure.

The methodological outcome of the study was thus defined in feasibility terms. More specifically, the research examined whether an interpretable decision-support prototype based on facial video analysis could be developed, implemented in an interview-compatible procedure, and meaningfully evaluated against expert judgement in recruitment-like conditions. Accordingly, the value of the proposed approach was assessed jointly in terms of analytical relevance, interpretability, and procedural compatibility with technology-mediated recruitment.

An optional personality inventory (IPIP-50 / Big Five) was administered for exploratory analysis only. These data were not treated as direct target variables for facial-video-based inference and were not used to justify personality classification from facial behaviour. Their role was limited to exploratory comparison at the level of broader behavioural patterning and hypothesis generation for future research.

3.1 Validation Strategy and Scope of Inference

The study applies a staged validation logic. First, laboratory work addressed physiological grounding by comparing optically derived facial activity traces with EMG-related measures under controlled conditions. Second, the study examined operational feasibility in a remote-capable, interview-compatible capture procedure, including calibration, segment-based observation, and near-real-time output generation. Third, the recruitment-oriented phase assessed expert-concordance in recruitment-like scenarios by comparing system outputs with the ratings of a single expert observer.

Importantly, this third stage does not evaluate predictive validity with respect to hiring outcomes, later job performance, or decision quality in live recruitment. The empirical claims of the present study should therefore be interpreted as prototype-level feasibility and preliminary expert-concordance evidence rather than deployment-ready recruitment validity.

3.2 Ethics, Privacy, and Governance Constraints

Given the sensitivity of behavioural and biometric data in employment-related contexts, the methodology incorporated governance safeguards from the outset. These included informed consent, data minimisation, transparency regarding the research purpose, and explicit separation between prototype evaluation and actual employment decision-making. The system was designed to favour auditable outputs, such as time-series traces and interpretable intermediate representations, over opaque single-score predictions.

The study therefore treats the prototype as a decision-support instrument whose legitimacy depends not only on technical performance, but also on its interpretability, bounded use, and compatibility with procedural safeguards. For this reason, the present methodology does not support the use of the system as a standalone or high-stakes selection mechanism.

4. Research Results and Discussion

4.1 Operator-facing Outputs in a Recruitment-style Setting

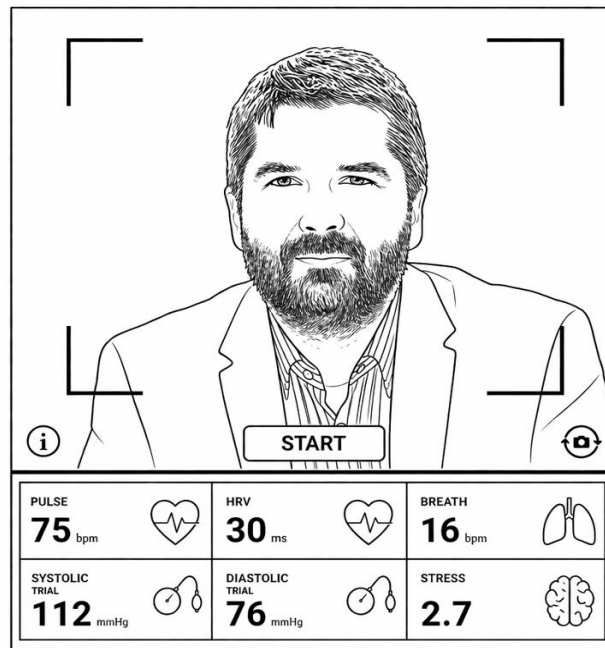
This subsection documents what the system produces in the context of a technology-mediated recruitment interview—i.e., the operator/interviewer view used to (a) start/stop acquisition, (b) monitor signal quality in real time, and (c) inspect whether the model is tracking interpretable facial activity rather than generating opaque scores.

Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε. presents a sketch of the real-time acquisition screen used during a recruitment-style recording segment. The interface shows a live camera preview with framing guides and a single-action control to begin the standardised capture (“START”).

Alongside the preview, the UI displays continuous indicators (e.g., pulse/HRV/breath and stress-related index values in the prototype build) intended as operator-facing decision-support telemetry.

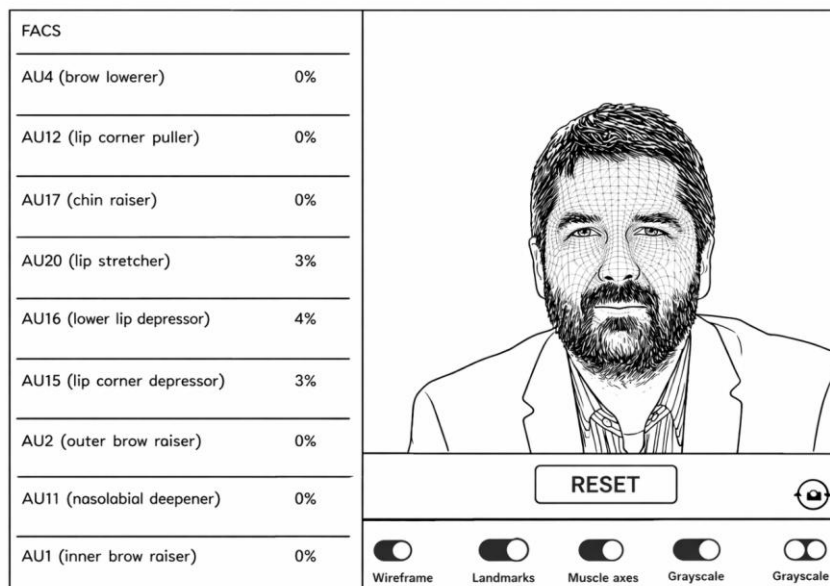
In a recruitment workflow, this screen primarily supports standardisation and quality control (ensuring adequate face visibility, stable framing, and sufficient recording duration per interview prompt) and provides a consistent, reviewable record of the system’s outputs for that segment.

Figure 2. Operator view of the real-time capture interface during a recruitment-style recording



Source: Own elaboration.

Figure 3. FACS/mesh interpretability wireframe



Source: Own elaboration.

Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε. shows the interpretability/diagnostic view of the same system, exposing intermediate representations used by the pipeline: (i) FACS Action Unit (AU) readouts (left panel; example AUs such as AU4, AU12, AU17, AU20, AU16, AU15 with instantaneous values), and (ii) an overlay of the reconstructed facial mesh/landmarks (right panel), with toggles for wireframe, landmarks, muscle axes, and grayscale rendering.

This view is included to demonstrate that the system provides auditable, behaviour-linked signals (how specific facial movements evolve over time) and allows an operator to verify tracking stability—critical in recruitment settings where video quality and head motion can materially affect measurement reliability.

Table 1. Agreement with single-expert assessment for recruitment-like prototype outputs

Output (recruitment framing)	Expert agreement (%)
Emotion analysis	85%
Interactional responsiveness (prompt-specific expert agreement)	60%
Stress management (stress index)	80%
Nonverbal behaviour (microexpression analysis)	55%

Source: Author's calculations.

Table 1 reports the agreement across the evaluated recruitment-oriented outputs; within this performance profile, emotion analysis and the stress index emerge as the most promising signals for near-term use as decision-support indicators in structured interview segments.

Beyond the per-output agreement levels reported in Table 1, the evaluation also addressed operational conditions determining whether the prototype could function within a structured interview workflow. For selected constructs, including stress-related inference, the system required up to 20 seconds of behavioural observation before an estimate could be generated. Once the relevant evidence window had been collected, the maximum computation time was 3 seconds.

In addition, an a priori operational criterion was defined for repeated use: the prototype would be regarded as sufficiently robust at this stage if it could analyse 20 consecutive individuals without downtime or maintenance intervention. This condition was met in testing, indicating that the system can operate within a segment-based interview procedure and return outputs with near-real-time latency.

The present results support two narrower conclusions. First, the prototype appears operationally feasible within a structured interview workflow, as selected constructs can be estimated after a limited observation window and returned within near-real-

time latency. Second, under recruitment-like conditions, the system showed preliminary agreement with a single expert observer, with higher agreement for emotion-related and stress-related outputs than for broader interpersonal and microexpression-related indicators. More specifically, the strongest agreement was observed for emotion analysis (85%) and stress-related inference (80%), whereas lower agreement was obtained for interactional responsiveness indicators (60%) and nonverbal behaviour based on microexpression analysis (55%).

This suggests that the present system is better suited to supporting the observation of short-term affective dynamics during specific interview prompts than to the interpretation of broader social-performance patterns. Importantly, these outputs should not be read as valid measures of stable interpersonal competence or personality traits.

These findings should be interpreted as evidence of workflow compatibility and exploratory expert-concordance, not as proof of decision validity in real hiring contexts. The recruitment-oriented evaluation was conducted in recruitment-like rather than real recruitment conditions and relied on comparison with a single expert observer, while the underlying project documentation explicitly notes concerns regarding objectivity and ecological validity.

For this reason, the present results do not justify treating the system as a standalone assessment mechanism or as evidence sufficient for high-stakes deployment in personnel selection. At this stage, the more defensible interpretation is that the prototype may serve as an auditable and standardised support layer that helps structure recruiter observation and identify moments warranting closer human review, while leaving evaluative authority with the human decision-maker.

5. Conclusions, Proposals, Recommendations

For avoidance of doubt, the study does not claim that facial-video analysis can validly infer Big Five traits or other stable personality characteristics from interview behaviour alone.

This study examined whether facial video analysis can be developed and embedded as an interpretable decision-support layer in technology-mediated recruitment interviews. The findings support a cautious conclusion: the proposed framework demonstrates prototype-level feasibility rather than validated recruitment effectiveness.

In particular, the study shows that facial activity captured from standard video can be transformed into structured, auditable, and temporally interpretable signals grounded in FLAME- and FACS-consistent representations, and that such signals can be incorporated into an interview-compatible analytical workflow.

Under recruitment-like conditions, the strongest exploratory results were obtained for emotion-related and stress-related outputs, whereas broader interpersonal and microexpression-based indicators proved less robust and should be treated as substantially more tentative.

The contribution of the study should therefore not be understood as demonstrating that affective AI can validly determine candidate suitability or improve hiring decisions in real organisational settings.

Rather, the paper provides initial evidence that selected affect-related behavioural outputs may be generated in a consistent and reviewable form and may serve as supplementary observational material in structured interview segments.

This interpretation is particularly important because the recruitment-oriented evaluation was conducted on 75 participants in recruitment-like scenarios and relied on comparison with a single expert observer, with aggregate agreement reported at 75%.

These results indicate preliminary expert-concordance under simulated conditions, but they do not establish criterion validity, predictive validity, or readiness for deployment in high-stakes personnel selection.

The study also makes a narrower operational contribution. The prototype was shown to function within a structured interview workflow, requiring a limited observation window for selected indicators, generating outputs in near-real time, and meeting a basic robustness requirement during consecutive use. These results support the claim that facial-video analytics can be technically integrated into recruitment-like procedures.

However, technical integration should not be conflated with practical legitimacy in live hiring environments. The project documentation explicitly identifies concerns related to objectivity, ecological validity, privacy, and the legal and ethical acceptability of conducting such assessment under near-real recruitment conditions. For this reason, the present findings should be interpreted as evidence of workflow feasibility and exploratory expert-concordance, not as support for autonomous or high-stakes use.

Accordingly, the main recommendation arising from this study is that any future development of recruitment-oriented affective AI should follow a conservative augmentation logic. Human decision authority must remain primary, while system outputs should be treated only as supplementary indicators requiring contextual interpretation by trained evaluators.

Further research should focus on validation in ecologically valid recruitment environments, the use of stronger reference standards than a single-observer

benchmark, and more precise determination of which categories of outputs can be considered sufficiently stable, interpretable, and governance-compatible for responsible organisational use.

In parallel, future work should continue to address robustness under heterogeneous remote-interview conditions and systematically examine privacy, transparency, fairness, consent, and data-protection safeguards. In this sense, the main value of the study lies not in claiming recruitment automation, but in defining a cautious and auditable prototype logic for future research on AI-augmented interview assessment.

References:

- Ahmed Khan, H.U.D., Bajwa, U.I., Ratyal, N.I., Zhang, F., Anwar, M.W. 2025. Deception detection in videos using the facial action coding system. *Multimedia Tools and Applications*, 84(9), 6429-6443.
- Barker, D., Tippireddy, M.K.R., Farhan, A., Ahmed, B. 2025. Ethical considerations in emotion recognition research. *Psychology International*, 7(2), 43.
- de Jesus, A.C.C., Mariano, R.V.R., Vieira, A.G., de Lima, J.A.D.A., de Castro, G.Z., Brandão, W.C. 2023. An Unsupervised Approach for Study of the Relationship Between Behavioral Profile and Facial Expressions in Video Interviews. In: *ICEIS* (1), pp. 493-500.
- Hogarth, A., McCartney, C. 2024. *Resourcing and Talent Planning Report 2024*. London: Chartered Institute of Personnel and Development. Available at: <https://www.cipd.org/globalassets/media/knowledge/knowledge-hub/reports/2024-pdfs/8662-resource-and-talent-planning-2024-report-web.pdf>.
- Hu, Q. 2023. Unilever's practice on AI-based recruitment. *Highlights in Business, Economics and Management*, 16, 256-263.
- Kamble, P., Kulkarni, U. 2022. An innovative approach of personality recognition for e-recruitment. In: *2022 5th International Conference on Advances in Science and Technology (ICAST)* (pp. 324-328). IEEE.
- Le, H.D., Lee, G.S., Kim, S.H., Kim, S., Yang, H.J. 2023. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *Ieee Access*, 11, 14742-14751.
- Martín-Raugh, M.P., Leong, C.W., Roohr, K.C., Chen, X. 2023. Perceived Conscientiousness and Openness to Experience Mediate the Relationship Between Vocal and Visual Features and Hiring Decision in the Interview. *Computers in Human Behavior Reports*, 10, 100272.
- Mocanu, B., Tapu, R., Zaharia, T. 2023. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. *Image and vision computing*, 133, 104676.
- Pfeffer, C.C. 2023. *Facial Expressions as Predictors of Long-Term Outcomes Following a Traumatic Event: Comparing Automated and Manual Coding Systems*. Columbia University.
- Saganowski, S., Perz, B., Polak, A.G., Kazienko, P. 2022. Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Transactions on Affective Computing*, 14(3), 1876-1897.
- Wei, Q., Huang, X., Zhang, Y. 2022. FV2ES: A fully end2end multimodal system for fast yet effective video emotion recognition inference. *IEEE Transactions on*

- Broadcasting, 69(1), 10-20.
- Wei, J., Hu, G., Yang, X., Luu, A.T., Dong, Y. 2024. Learning facial expression and body gesture visual information for video emotion recognition. *Expert Systems with Applications*, 237, 121419.
- Wu, M., Teng, W., Fan, C., Pei, S., Li, P., Lv, Z. 2023. An investigation of a factory-enhanced video on EEG-based emotion recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31, 1602-1613.
- Yi, Y., Zhou, Y., Wang, T., Zhou, J. 2025. Advances in Video Emotion Recognition: Challenges and Trends. *Sensors*, 25(12), 3615.