Design of a Vision Transformer-Based Architecture for Automatic Facial Emotion Monitoring in Workplace Environments

Renzo Sebastian Gonzalez Caceres, Jeramel Melissa Avila Saldaña, Patricia Gissela Pereyra Salvador
Faculty of Engineering, Peruvian University of Applied Sciences, Lima, Peru

***Abstract*—Facial emotion recognition is increasingly considered in affective computing as a mechanism for unobtrusive emotional awareness in organizational environments. This study proposes the design of a Vision Transformer (ViT)-based system architecture for automatic facial emotion monitoring, focusing on deployability, modular integration, and data-governance considerations rather than model benchmarking. The architecture defines a complete pipeline comprising a visual data acquisition layer, a processing backend for transformer-based inference, and a web-based visualization interface intended for aggregated emotional analytics. Publicly available datasets such as FER2013 and AffectNet are identified as reference sources for model adaptation within the proposed framework. The work details system components, data flow, scalability strategies, and privacy-by-design mechanisms, including transient image handling and non-persistent processing. Rather than presenting experimental performance, this study provides a technical blueprint and feasibility analysis intended to guide future implementation and validation of transformer-driven emotion monitoring systems in workplace contexts. The proposed framework aims to bridge the gap between advances in deep learning models and their practical integration into real-world organizational infrastructures.**

***Keywords*—Facial emotion recognition; vision transformer; affective computing; system architecture design; workplace emotion monitoring; privacy-by-design; deep learning inference**

I. Introduction

Emotional well-being in the workplace is a recognized determinant of organizational sustainability, employee retention, and productivity. In Metropolitan Lima, this issue carries particular relevance given a sustained rise in occupational stress and fatigue, with recent reports estimating that approximately 62% of employees exhibit significant stress-related symptoms, a condition further intensified by the COVID-19 pandemic [1], [2]. These affective states are associated with elevated absenteeism, increased staff turnover, and diminished operational performance, underscoring the need for scalable and systematic monitoring approaches.

Current organizational responses rely predominantly on periodic self-reporting instruments, such as psychometric surveys or structured interviews, which are limited by their discrete nature, susceptibility to social desirability bias, and inability to capture emotional fluctuations in real time. There is therefore, a notable gap in automated, non-intrusive systems for continuous affective state monitoring in workplace environments [3]. Although recent deep learning approaches have reported facial emotion recognition accuracies exceeding 90% under controlled conditions [4], their translation into deployable organizational systems remains largely unaddressed, further compounded in the Peruvian context by the scarcity of locally representative datasets and cultural variability in facial expression patterns [5].

Vision Transformers (ViT) offer a promising pathway to address this gap by modeling global spatial dependencies across facial image patches through self-attention mechanisms, enabling more generalizable representations of facial expressions under unconstrained conditions compared to convolutional approaches [6]. However, model performance alone does not translate into operational systems. Workplace deployment introduces architectural requirements beyond inference accuracy, including modular integration, real-time data flow, visual analytics for non-technical stakeholders, and data governance compliance, constituting a distinct design problem that motivates the present work.

This study addresses how a ViT-based pipeline can be systematically designed to meet the scalability, modularity, and privacy requirements of continuous workplace emotion monitoring. The study proposes a complete system architecture comprising a visual data acquisition layer, a cloud-based inference backend, and a web-based visualization interface for aggregated emotional analytics. The contribution is explicitly architectural: this work provides a technical blueprint and feasibility analysis to guide future implementation and empirical validation, rather than reporting experimental performance benchmarks.

II. Related Works

Research on transformer-based affective computing has advanced considerably at the international level, though most existing work focuses on multimodal inference under controlled experimental conditions rather than on deployable system architectures. These contributions are reviewed here to contextualize the architectural gaps that the present work addresses, rather than to establish direct performance comparisons with the proposed unimodal system.

In the domain of multimodal emotion recognition, [5] introduced the KoHMT model (Knowledge-Oriented Hybrid Model Transformer), a hybrid architecture combining the HuBERT model for auditory feature extraction and

KoELECTRA for textual representation, fused through a cross-attention mechanism. Trained on AI-Hub datasets and deployed on Linux with NVIDIA GPUs, KoHMT demonstrates the expressive potential of combining speech and text modalities for affective state classification. However, its reliance on multiple synchronized input streams and the absence of an integrated sensing and visualization layer limit its applicability as a blueprint for continuous, camera-based workplace monitoring systems.

Similarly, [3] proposed the TCMA model (Transformer-based Cross-Modal Attention), which combines facial imagery with non-invasive physiological signals using a 1D-CNN for rPPG extraction and ResNet50 for visual features, fused through cross-modal attention. Trained on MAHNOB-HCI and DEAP datasets, TCMA achieves high accuracy under laboratory conditions but requires physiological sensing hardware and lacks a modular deployment strategy suitable for enterprise-level integration. Both KoHMT and TCMA illustrate the state of the art in affective inference quality while simultaneously highlighting the absence of end-to-end architectural frameworks oriented toward organizational deployment, data governance, and non-technical user interfaces.

Within the unimodal vision domain, transformer-based models such as POSTER [7] and TokenFace [8] have demonstrated that ViT-derived architectures can achieve competitive FER performance using facial image data alone, without reliance on auxiliary modalities. These works provide direct architectural precedent for the inference component proposed in the present study and support the selection of ViT as the core model within a unimodal, camera-based monitoring pipeline.

At the national level, the landscape of applied computer vision research in Peru remains nascent. The most representative work identified in this context is [6], which presented a vision-based system for automatic vehicle license plate recognition using YOLOv4 and convolutional neural networks, trained on a locally collected dataset of one thousand images captured in Lima. While this work demonstrates the technical feasibility of deploying vision-based sensing systems within the Peruvian context, it is confined to vehicular recognition and does not engage with facial analysis, affective computing, or the architectural requirements of distributed emotional monitoring. Notably, no peer-reviewed work was identified addressing transformer-based facial emotion recognition systems designed or validated within Peru, representing a significant gap that the present proposal directly targets.

The reviewed literature collectively reveals two interconnected deficiencies: first, the predominance of multimodal and laboratory-oriented FER systems that prioritize inference accuracy over deployability; and second, the complete absence of ViT-based emotion monitoring architectures adapted to the Peruvian organizational context. The present work addresses both dimensions by proposing a system architecture that prioritizes modularity, scalability, and privacy-by-design over experimental benchmarking.

III. Architecture Design

This section presents the main contribution of this study: an integrated architecture for automatic facial emotion monitoring in workplace environments. The proposed system is designed for deployment on the Google Cloud Platform (GCP), leveraging Vision Transformers (ViT) as the core inference component for facial image analysis. ViT architectures process images by dividing them into fixed-size patches and modeling global spatial dependencies through self-attention mechanisms, enabling robust facial expression representation under unconstrained real-world conditions [8]. The overall design supports automated processing, scalable data management, and continuous inference, structured across three interconnected layers: the technological layer, the application layer, and the business layer.

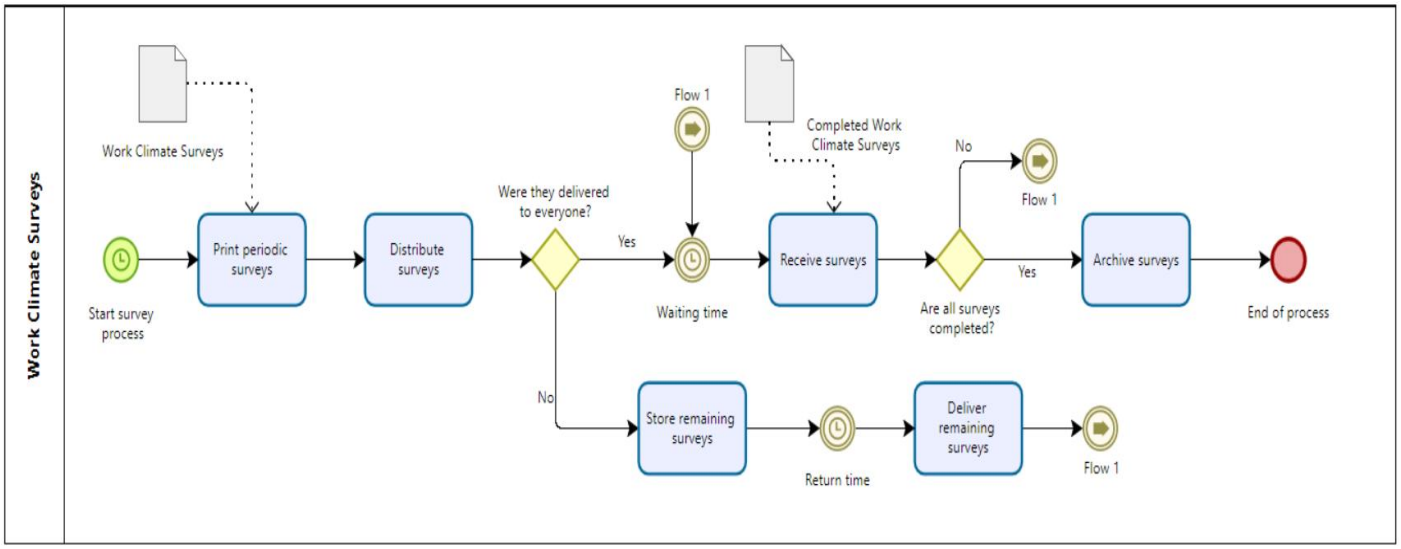

Fig. 1. Workflow of the workplace climate survey process.

B. Current Analysis

To contextualize the architectural requirements of the proposed system, a diagnostic interview was conducted with Human Resources personnel at a private organization operating in Metropolitan Lima. The interview aimed to characterize current practices for workplace emotional monitoring and identify operational limitations that an automated system should address. The organization requested anonymity; its profile is considered representative of mid-sized enterprises in the Peruvian service sector.

The interview revealed that emotional monitoring in this organization is conducted exclusively through periodic paper-based or digital survey instruments, distributed manually to employees at fixed intervals. As illustrated in Fig. 1 above, the current workflow is triggered by a scheduled timer event and initiates with the physical printing and manual distribution of survey forms. The process includes a verification checkpoint to confirm whether forms were delivered to all employees, with a dedicated sub-flow for handling pending deliveries and an intermediate waiting period before collection. Once surveys are received, a second decision point determines whether all forms have been completed before the results are archived. The entire process concludes with physical or digital storage of the completed documents, with no automated processing, inference, or real-time visualization at any stage.

This workflow exhibits several structural limitations relevant to the design of the proposed system. First, the timer-triggered nature of the process means emotional data is captured only at discrete intervals, preventing the detection of affective fluctuations that develop between survey cycles. Second, the manual distribution and collection of forms introduce operational overhead for HR personnel and potential disruptions to employee workflow. Third, self-reported survey responses are inherently susceptible to social desirability bias and subjective interpretation, limiting the reliability of the collected data. Fourth, the absence of any analytical or visualization layer means that aggregated emotional indicators cannot be derived in real time, constraining the organization's capacity for proactive decision-making.

These limitations collectively define the operational gap that the proposed architecture targets: the need for a continuous, non-intrusive, and automated system capable of capturing facial emotional indicators, performing transformer-based inference, and delivering aggregated analytics to organizational stakeholders without interrupting employee activities.

C. Design Decision Rationale

The selection of core architectural components for the proposed system was grounded in a systematic literature-based comparative analysis conducted across five dimensions: deep learning model architecture, specialized FER models, primary training dataset, complementary dataset, and cloud deployment platform. Each comparative analysis evaluated candidate options against weighted criteria derived from the technical requirements of a deployable workplace emotion monitoring system. The methodology assigned impact-weighted scores to each candidate based on performance metrics, characteristics, and limitations reported in their respective primary sources, enabling an evidence-based ranking of alternatives.

Model Architecture Selection. The first comparative analysis evaluated four deep learning architectures: CNN [9], RCNN [10], QCNN [11], and ViT [8], assessed against five criteria: emotion classification accuracy, inference time, robustness to environmental variation, adaptability to facial diversity, and architectural maturity. As shown in Table I, ViT achieved the highest weighted score (4.467), outperforming CNN (4.267), RCNN (4.267), and QCNN (3.867). ViT's self-attention mechanism enables the modeling of global spatial dependencies across facial image patches, yielding superior accuracy (89.90%) and robustness under variable illumination and pose conditions compared to convolution-based alternatives, which are constrained by local receptive fields. Although QCNN demonstrated competitive inference speed (95 ms), its robustness limitations and lower maturity for real-world deployment reduced its overall suitability.

TABLE. I. WEIGHTED COMPARATIVE ANALYSIS OF DEEP LEARNING ARCHITECTURES FOR FACIAL EMOTION RECOGNITION

Criterios	Impacto	ViT	RCNN	QCNN	CNN
		Puntaje	Puntaje	Puntaje	Puntaje
Accuracy in emotion classification	26.67%	5	4	4	4
Inference time	6.67%	3	2	2	1
Robustness in the face of changing conditions	20.00%	4	5	3	5
Adaptability to a diversity of expressions and faces	20.00%	4	5	3	4
Technologies, techniques, and models used	26.67%	5	4	5	5
Total		**21**	**20**	**17**	**19**

Specialized Model Validation. A second analysis extended the comparison to advanced architectures including 1D-CNN [12], SMaTE [13], TCMA [3], ResNet50 [14], and ViT [8], evaluated on classification accuracy, inference time, number of emotional classes, and generalization capacity. As shown in Table II, ViT and ResNet50 achieved equivalent weighted scores (4.555), with SMaTE (4.000), 1D-CNN (3.777), and TCMA (3.555) ranking lower. While SMaTE achieved the highest reported accuracy (99.19%) and ResNet50 demonstrated superior inference speed (28 ms), ViT was selected for its balanced profile: accuracy above 90%, multiclass coverage of seven emotion categories, robust generalization in non-controlled environments, and unimodal operation requiring only visual input. The exclusion of multimodal architectures such as TCMA was deliberate, as their dependency on physiological sensors is incompatible with the non-intrusive sensing constraints of the proposed workplace deployment.

TABLE. II. WEIGHTED COMPARATIVE ANALYSIS OF SPECIALIZED DEEP LEARNING MODELS FOR FER

Criterios	Impacto	1D-CNN	SMaTE	TCMA	ResNet50	ViT
		Puntaje	Puntaje	Puntaje	Puntaje	Puntaje
Accuracy in emotional classification	22,22%	4	5	4	5	5
Inference time	22,22%	4	3	2	4	4
Number of emotional classes evaluated	22,22%	3	4	4	4	4
Level of model generalization	33,33%	4	4	4	5	5
Total		15	16	14	18	**18**

Primary Dataset Selection. The third analysis compared FER2013 [15], CK+ [16], Aff-Wild2 [17], and RAF-DB [18] across five criteria: data volume, emotional coverage, demographic diversity, capture quality, and class balance. As shown in Table III, FER2013 achieved the highest weighted score (3.688), followed by RAF-DB (3.375), CK+ (2.938), and Aff-Wild2 (2.625). FER2013's 35,887 grayscale images (48×48 pixels) collected under unconstrained conditions provide a representative training base for real-world deployment, covering seven basic emotion categories with acceptable demographic diversity. Its class imbalance, particularly for the disgust category, can be mitigated through data augmentation strategies within the proposed pipeline.

TABLE. III. WEIGHTED COMPARATIVE ANALYSIS OF PRIMARY TRAINING DATASETS

Criterios	Impacto	Aff-Wild2	CK+	FER2013	RAF-DB
		Puntaje	Puntaje	Puntaje	Puntaje
Data volume	25%	3	2	5	3
Emotional coverage	18.75%	3	3	4	3
Demographic diversity	18.75%	3	2	5	5
Capture quality and environment	18.75%	3	5	3	3
Balance of emotional classes	18.75%	1	3	1	3
Total		**13**	**15**	**18**	**17**

Complementary Dataset Selection. The fourth analysis identified a supplementary dataset to enhance model generalization, evaluating AFEW [19], AffectNet [20], DEFE [21], and MELD [22] on emotional coverage, volume, demographic diversity, and capture conditions. As shown in Table IV, AffectNet achieved the highest weighted score (4.999), substantially outperforming MELD (4.000), AFEW (3.555), and DEFE (3.444). With over 450,000 manually annotated images covering eight discrete emotion categories and continuous valence-arousal dimensions, AffectNet provides the demographic diversity and in-the-wild variability necessary to improve ViT generalization across populations and environments relevant to the Peruvian workplace context.

TABLE. IV. WEIGHTED COMPARATIVE ANALYSIS OF COMPLEMENTARY DATASETS

Criterios	Impacto	AFEW	AffectNet	DEFE	MELD
		Puntaje	Puntaje	Puntaje	Puntaje
Emotional coverage	22.22%	4	5	4	4
Dataset volume	33.33%	2	5	3	4
Demographic diversity	22.22%	5	5	3	3
Capture conditions	22.22%	4	5	4	5
Total		**15**	20	14	16

Cloud Platform Selection. The fifth analysis evaluated deployment platforms including OsmOticGate2 [23], VaBUS [24], Google Cloud Platform [25], and MArk [26], assessed on processing latency, inference accuracy, operational scalability, and estimated cost. As shown in Table V, GCP achieved the highest weighted score (5.000), followed by OsmOticGate2 and MArk (both 4.000) and VaBUS (3.100). GCP's ecosystem of specialized services, including Vertex AI for model serving, Cloud Storage for transient image handling, Firestore for structured data management, and Compute Engine for scalable inference, provides a comprehensive and cost-effective infrastructure for the proposed architecture. Its support for GPU and TPU acceleration enables inference latency optimization within the 100–300 ms range suitable for near-real-time workplace monitoring.

TABLE. V. WEIGHTED COMPARATIVE ANALYSIS OF CLOUD DEPLOYMENT PLATFORMS

Criterios	Impacto	OsmOticGate2	VaBUS	Google Cloud	MArk
		Puntaje	Puntaje	Puntaje	Puntaje
Processing latency	30%	5	2	3	3
Accuracy in emotional inference	20%	3	5	5	5
Operational scalability	20%	5	3	5	3
Estimated operating cost	30%	3	3	5	3
Total		16	13	**18**	**14**

These five decisions collectively define the technical foundation of the proposed architecture: a ViT-based inference model fine-tuned on FER2013 and AffectNet, deployed on Google Cloud Platform within a modular pipeline designed for organizational scalability and privacy compliance.

D. Logical Architecture

The proposed logical architecture is structured into three interconnected layers: the business layer, the application layer, and the technological layer, as illustrated in Fig. 2. Together,

these layers define the system's functional components, information flows, and actor interactions, ensuring modularity and scalability across local and cloud-based environments.

The business layer defines the two organizational actors. The Human Resources department accesses the system via a web browser to visualize aggregated emotional indicators, apply departmental filters, and export reports. Employees interact through a local desktop application (Lumora-Employee.exe) that handles facial image capture and transmission, requiring no manual intervention beyond initial setup. No individual-level emotional profiling is exposed to any user role.

The application layer mediates between user interaction and deep learning inference. The front-end, hosted on Vercel, provides a dashboard with charts and filters, a visualization module, and a document export module. The GCP-hosted back-end integrates a REST API, a ViT-based inference module, and a database storing results, timestamps, and organizational metadata. A local edge component on the employee workstation performs preliminary facial capture processing before transmitting results to the cloud, reducing bandwidth requirements and supporting near-real-time inference.

The technological layer provides the supporting infrastructure, including firewall-governed network connectivity and Cloud Storage for transient image handling, ensuring facial images are not retained beyond the inference cycle in accordance with the privacy-by-design principles of the proposed framework.

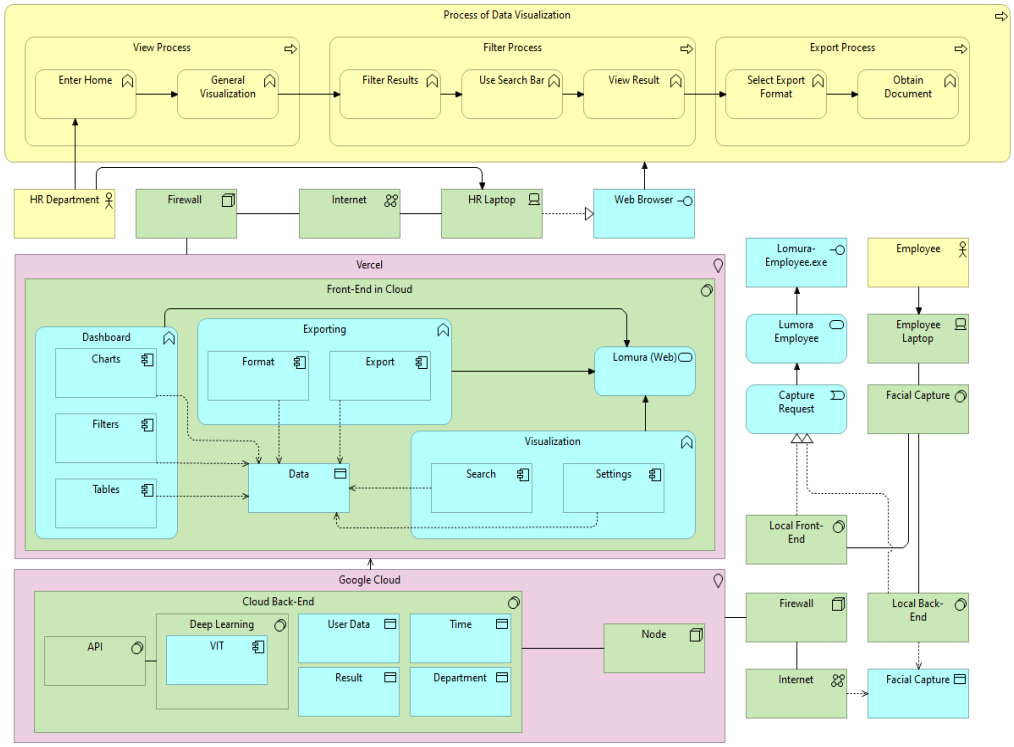

Fig. 2. Logical architecture of the proposed facial emotion monitoring system.

E. Physical Architecture

The physical architecture defines the hardware and infrastructure components supporting the end-to-end data flow from facial image acquisition to aggregated visualization, as illustrated in Fig. 3. The design is organized across four layers: client and collection, communication, cloud back-end, and presentation.

At the client and collection layer, each employee workstation consists of a laptop equipped with a camera running a local application composed of a frontend module handling login and user interaction, and a backend module managing capture logic through standard controllers, services, and entities. Facial images are captured at predefined intervals and transmitted via HTTP through a REST API to the cloud infrastructure. Images are used exclusively for inference and discarded immediately upon completion, with no persistent storage of visual data.

Within the cloud back-end layer, hosted on GCP, a server-side application receives incoming data through the API and processes it using a deep learning module containing the trained ViT model. As shown in Fig. 3, the ViT implementation processes facial image patches through a Transformer encoder to generate emotion classification outputs, which are stored as structured prediction results alongside timestamps and organizational metadata.

At the presentation layer, HR personnel access the system from a dedicated client device through a web-based frontend built with React and deployed on Vercel. This interface provides interactive dashboards for visualizing aggregated emotional indicators and generating exportable

reports at the organizational level.

F. Projected System Flow

The projected system flow defines the expected operational behavior once the proposed architecture is deployed, replacing the current manual survey-based process with an automated, continuous monitoring pipeline. The redesigned workflow is illustrated in Fig. 4 and organized across two interconnected processes: the emotion monitoring process and the HR department consultation process.

The emotion monitoring process is triggered automatically by a timer event, eliminating the need for employee intervention. The system captures facial images from the employee workstation, uploads them to cloud storage, and processes them through the ViT inference module. The resulting emotion classification outputs are then managed through a data gateway: if no results table exists, a new one is generated; otherwise, new records are appended to the existing table. The updated results table subsequently triggers a signal to the HR consultation process, ensuring dashboard data remains current without manual consolidation.

The HR department process initiates when an authorized HR manager accesses the web platform. The system reads the updated results table and evaluates whether data is available: if no results are found, an empty dashboard is displayed; if results exist, the dashboard is refreshed with the current aggregated indicators. The HR manager can then sequentially consult global indicators, personnel metrics, emotional trend charts, and individual employee records, concluding with the delivery of consolidated information to the HR area for organizational decision-making.

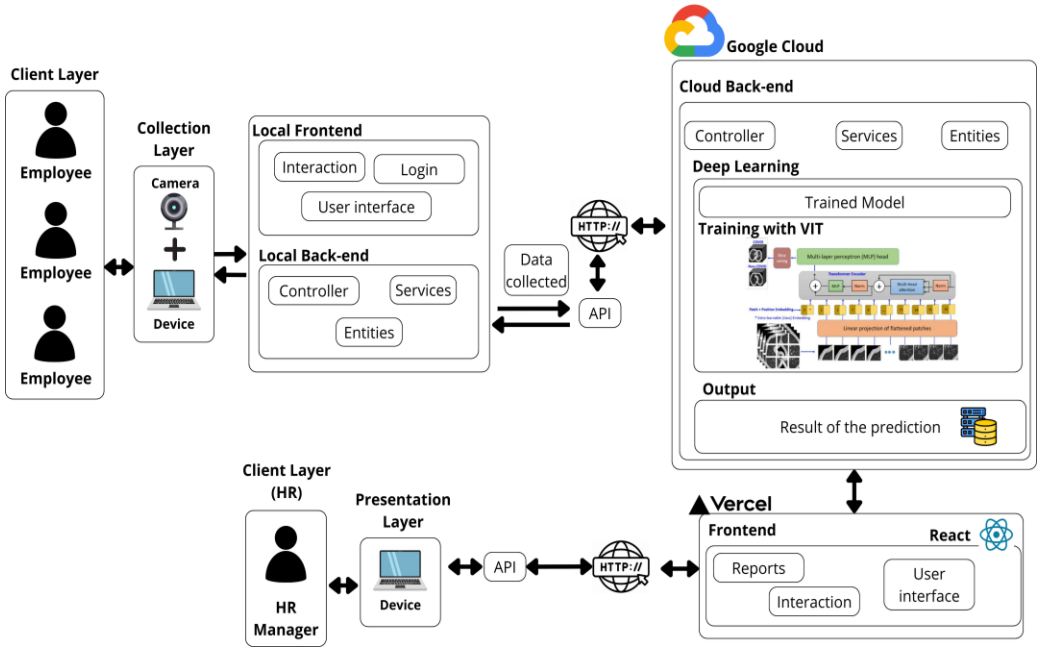

Fig. 3. Physical architecture of the proposed facial emotion monitoring system.

Compared to the current AS-IS workflow described in Section III(A), this proposed flow eliminates document printing, manual survey distribution, waiting periods, and physical archiving. The HR department gains continuous access to aggregated emotional indicators without administrative overhead, while employees contribute data passively without interrupting their work activities.

G. Ethical and Privacy Considerations

The proposed system incorporates privacy-by-design as a foundational architectural principle, ensuring that data protection mechanisms are embedded from the initial design stage rather than applied retroactively. Facial images captured by the local client are processed transiently in memory, transmitted via encrypted connections to the cloud inference module, and automatically discarded upon inference completion. No facial image is written to persistent storage at any stage, fulfilling the data minimization and storage limitation principles established under Peruvian Law N° 29733 on Personal Data Protection [27] and aligned with international Privacy by Design guidelines.

The system operates exclusively at an aggregated organizational level, associating inference outputs with departmental and temporal metadata rather than persistent individual identifiers, thereby preventing individual emotional profiling. Access to the visualization dashboard is restricted to authorized HR personnel through secure authentication mechanisms. These design choices are consistent with Law N° 29783 on Occupational Safety and Health [28], which mandates the protection of employees' physical and psychological integrity in the workplace, and with Law N° 31814 [29], Peru's first legal framework promoting ethical and responsible AI use in organizational contexts.

At the architectural level, system quality attributes including security, reliability, and maintainability are governed by ISO/IEC 25010:2023 [30], while responsible AI design principles follow the guidelines of ISO/IEC JTC 1/SC

42 [31]. Together, these standards and legal references substantiate the ethical feasibility of deploying the proposed architecture in Peruvian workplace environments.

IV. Conclusion

This study proposed the design of a Vision Transformer-based architecture for automatic facial emotion monitoring in workplace environments, addressing the architectural gap between advances in deep learning models and their practical integration into organizational contexts. The proposed system defines a complete pipeline comprising a visual data acquisition layer, a ViT-based cloud inference backend, and a web-based visualization interface, structured across three interconnected layers that support scalability, modularity, and privacy-by-design compliance.

The literature-based comparative analyses conducted across five dimensions, covering model architecture, specialized FER models, training datasets, complementary datasets, and cloud deployment platforms, substantiate the technical decisions underlying the proposed design. The selection of ViT as the core inference component, FER2013 and AffectNet as reference datasets, and GCP as the deployment platform reflects an evidence-based approach oriented toward deployability and organizational scalability rather than benchmark optimization.

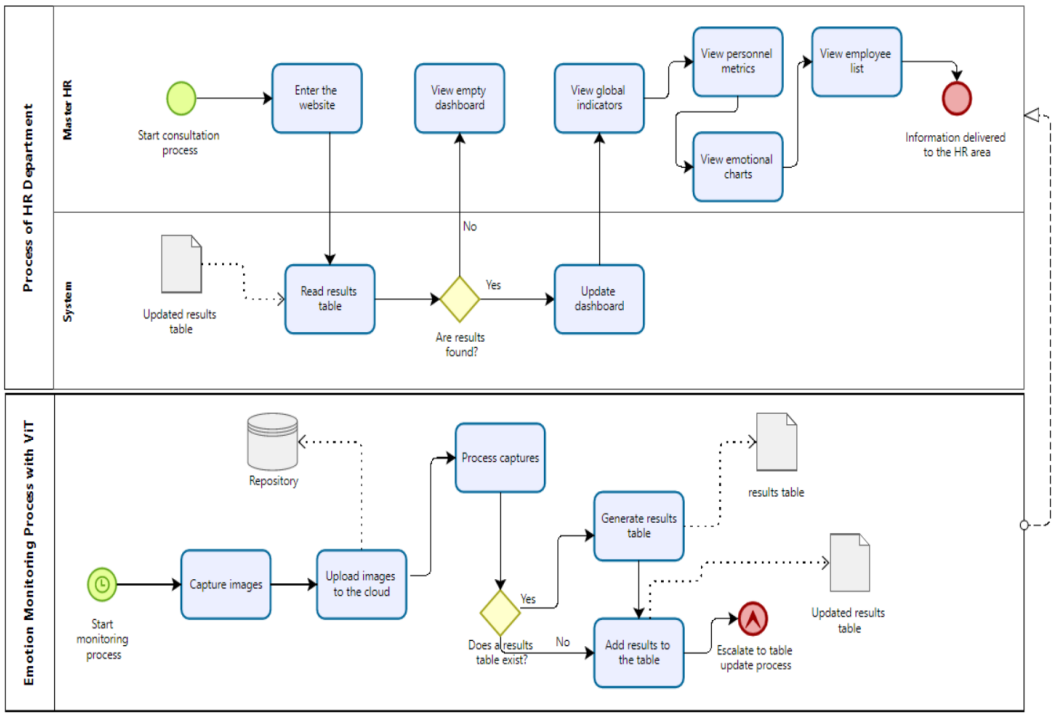

Fig. 4. Projected system flow for the proposed emotion monitoring architecture.

The projected system flow demonstrates that replacing periodic manual surveys with continuous automated monitoring can reduce administrative overhead, eliminate workflow interruptions, and provide HR departments with timely aggregated emotional indicators to support preventive organizational decision-making. These contributions are framed within applicable Peruvian legal standards and international quality frameworks, establishing the ethical and regulatory feasibility of the proposed architecture.

This work is explicitly scoped as an architectural blueprint. The primary limitation is the absence of empirical validation, as no prototype implementation or pilot deployment has been conducted. Future work should focus on the development and experimental evaluation of the proposed system, including performance benchmarking of the ViT inference module on FER2013 and AffectNet, latency measurement under real workplace network conditions, and a pilot deployment in a Peruvian organizational context to assess system usability and HR acceptance.

References

[1] G. Requejo Pacheco, M. S. Villa Santillán, L. Ruiz Barrera, y E. E. Rojas de la Puente, "Síndrome de burnout en trabajadores empresariales en Perú," Revista de Ciencias Sociales (RCS), vol. XXIX, no. 3, pp. 470–483, Sep. 2023, doi: https://doi.org/10.31876/rcs.v29i3.40731.

[2] H. Ko, D. Kim, S.-S. Cho, D.-W. Lee, J. Choi, M. Kim, M. Y. Park, y M.-Y. Kang, "The association of emotional labor and workplace violence with health-related productivity loss," Journal of Occupational Health, vol. 66, no. 1, Art. no. uiae057, Sep. 2024, https://doi.org/10.1093/joccuh/uiae057.

[3] J. Li and J. Peng, "End-to-End Multimodal Emotion Recognition Based on Facial Expressions and Remote Photoplethysmography Signals," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 10, pp. 6194–6204, Oct. 2024, https://doi.org/10.1109/JBHI.2024.3430310.

[4] P. Siritanawan, H. Kojima, and K. Kotani, "Exploring the Cultural Gaps in Facial Expression Recognition Systems by Visual Features," in Proc. 2023 IEEE Region 10 Conf. (TENCON), Chiang Mai, Thailand, Oct. 2023, pp. 1–6, https://doi.org/10.1109/TENCON58879.2023.10322368.

[5] M.-H. Yi, K.-C. Kwak, and J.-H. Shin, "KoHMT: A Multimodal Emotion Recognition Model Integrating KoELECTRA, HuBERT with Multimodal Transformer," Electronics, vol. 13, no. 23, p. 4674, Dec. 2024, https://doi.org/10.3390/electronics13234674.

[6] D. M. Valdeos Acevedo, A. S. Vadillo Velazco, M. G. S. Pérez Paredes, and R. M. Arias Velásquez, "Methodology for an automatic license plate recognition system using convolutional neural networks for a Peruvian case study," IEEE Latin America Transactions, vol. 20, no. 6, pp. 1032–1039, Jun. 2022, https://doi.org/10.1109/TLA.2022.9757747.

[7] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–26, 2023, https://doi.org/10.1109/TIM.2023.3243661.

[8] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A survey on efficient vision transformers: Algorithms, techniques, and performance benchmarking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 7682–7703, Dec. 2024, https://doi.org/10.1109/TPAMI.2024.3392941.

[9] M. A. Saleem, N. Senan, F. Wahid, M. Aamir, A. Samad y M. Khan, "Comparative analysis of recent architecture of convolutional neural network," Mathematical Problems in Engineering, vol. 2022, 2022, https://doi.org/10.1155/2022/7313612.

[10] S. Reddy, N. Pillay y N. Singh, "Comparative evaluation of convolutional neural network object detection algorithms for vehicle detection," Journal of Imaging, vol. 10, no. 7, 2024, https://doi.org/10.3390/jimaging10070162.

[11] S. Hossain, S. Umer, R. K. Rout y H. A. Marzouqi, "A deep quantum convolutional neural network based facial expression recognition for mental health analysis," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pp. 1556–1565, 2024, https://doi.org/10.1109/TNSRE.2024.3385336.

[12] G. Rescio, A. Manni, M. Ciccarelli, A. Papetti, A. Caroppo y A. Leone, "A deep learning-based platform for workers' stress detection using minimally intrusive multisensory devices," Sensors, vol. 24, no. 3, 2024, https://doi.org/10.3390/s24030947.

[13] N. Kim, S. Cho y B. Bae, "SMaTE: A segment-level feature mixing and temporal encoding framework for facial expression recognition," Sensors, vol. 22, no. 15, 2022, https://doi.org/10.3390/s22155753.

[14] S. H. Al-Gburi, K. A. Al-Sammak, I. Marghescu, C. C. Oprea, A.-M. C. Drăgulinescu, N. A. M. Alduais, K. M. A. Alheeti y N. A. H. Al-Sammak, "EffRes-DrowsyNet: A novel hybrid deep learning model combining EfficientNetB0 and ResNet50 for driver drowsiness detection," Sensors, vol. 25, no. 12, art. no. 3711, 2025, https://doi.org/10.3390/s25123711.

[15] A. Kashef, Y. Wang, M. N. Assafi, J. Ma, J. Wang, J. A. Jones, and L. Thiamwong, "Developing a novel AI-enabled extended reality system for real-time automatic facial expression recognition and system performance evaluation," Advanced Engineering Informatics, vol. 55, Feb. 2025, Art. no. 103207, https://doi.org/10.1016/j.aei.2025.103207.

[16] X. Dong, X. Ning, J. Xu, L. Yu, W. Li y L. Zhang, "A recognizable expression line portrait synthesis method in portrait rendering robot," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 1440–1450, 2024, https://doi.org/10.1109/TCSS.2023.3241003.

[17] D. Dresvyanskiy, E. Ryumina, H. Kaya, M. Markitantov, A. Karpov y W. Minker, "End-to-end modeling and transfer learning for audiovisual emotion recognition in-the-wild," Multimodal Technologies and Interaction, vol. 6, no. 2, 2022, https://doi.org/10.3390/mti6020011.

[18] U. Nawaz, Z. Saeed y K. Atif, "A novel transformer-based approach for adult's facial emotion recognition," IEEE Access, 2025, https://doi.org/10.1109/ACCESS.2025.3555510.

[19] B. Pan, K. Hirota, Y. Dai, Z. Jia, S. Shao y J. She, "Learning sequential variation information for dynamic facial expression recognition," IEEE Transactions on Neural Networks and Learning Systems, 2025, https://doi.org/10.1109/TNNLS.2025.3548669.

[20] Z.-Y. Huang, C.-C. Chiang, J.-H. Chen, Y.-C. Chen, H.-L. Chung, Y.-P. Cai, and H.-C. Hsu, "A study on computer vision for facial emotion recognition," Scientific Reports, vol. 13, no. 8425, 2023, https://doi.org/10.1038/s41598-023-35446-4.

[21] W. Li, Y. Cui, Y. Ma, X. Chen, G. Li, G. Zeng, G. Guo y D. Cao, "A spontaneous driver emotion facial expression (DEFE) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios," IEEE Transactions on Affective Computing, vol. 14, no. 1, pp. 747–760, 2023, https://doi.org/10.1109/TAFFC.2021.3063387.

[22] A. Aguilera, D. Mellado y F. Rojas, "An assessment of in-the-wild datasets for multimodal emotion recognition," Sensors, vol. 23, no. 11, 2023, https://doi.org/10.3390/s23115184.

[23] B. Qian, Y. Xuan, D. Wu, Z. Wen, R. Yang, S. He, J. Chen y R. Ranjan, "Edge-cloud collaborative streaming video analytics with multi-agent deep reinforcement learning," IEEE Network, 2024, https://doi.org/10.1109/MNET.2024.3398724.

[24] H. Wang, Q. Li, H. Sun, Z. Chen, Y. Hao, J. Peng, Z. Yuan, J. Fu y Y. Jiang, "VaBUS: Edge-cloud real-time video analytics via background understanding and subtraction," IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 90–106, 2023, https://doi.org/10.1109/JSAC.2022.3221995.

[25] V. D. Nguyen, H. V. Khoa, T. N. Kieu, and E.-N. Huh, "Internet of Things-based intelligent attendance system: Framework, practice implementation, and application," Electronics, vol. 11, no. 3151, Sep. 2022, https://doi.org/10.3390/electronics11193151.

[26] C. Zhang, M. Yu, W. Wang y F. Yan, "Enabling cost-effective, SLO-aware machine learning inference serving on public cloud," IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 1765–1779, 2022, https://doi.org/10.1109/TCC.2020.3006751.

[27] Congress of the Republic of Peru, "Law No. 29733, Personal Data Protection Law," Peru, 2011. [Online]. Available: https://www.gob.pe/institucion/congreso-de-la-republica/normas-legales/243470-29733.

[28] Congress of the Republic of Peru, "Law No. 29783, Occupational Safety and Health Law," Peru, 2012. [Online]. Available: https://www.gob.pe/institucion/congreso-de-la-republica/normas-legales/462576-29783.

[29] Congress of the Republic of Peru, "Law No. 31814, Law promoting the use of Artificial Intelligence for the development of the country," Peru, 2023. [Online]. Available: https://www.gob.pe/institucion/congreso-de-la-republica/normas-legales/4565760-31814.

[30] ISO/IEC, "Systems and Software Engineering—Systems and Software Quality Requirements and Evaluation (SQuaRE)—Product Quality Model," ISO/IEC 25010:2023, International Organization for Standardization, Geneva, Switzerland, Nov. 2023. [Online]. Available: https://www.iso.org/standard/78176.html.

[31] ISO/IEC JTC 1/SC 42, "Artificial Intelligence — Subcommittee SC 42," International Organization for Standardization, Geneva, Switzerland. [Online]. Available: https://www.iso.org/committee/6794475.html.