Detecting Gamer Frustration in Real Time Through Webcam-Based Facial Analysis and Learned Classifiers

Aadya Shetty [1], Abhishek Rana [2], Bhargav P [3], Bhuvan S Shetty [4], Hema M S [5]

Department of Computer Science and Engineering
RV Institute of Technology and Management
Bengaluru, India

How to Cite this Article:

Shetty, A., Rana, A., P, B., Shetty, B. S. & S, H. M. (2026). Detecting Gamer Frustration in Real Time Through Webcam-Based Facial Analysis and Learned Classifiers. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.910

https://doi.org/10.55041/ijcope.v2i4.910

***Abstract*—**The capacity to automatically interpret a person's emotional condition from their face has attracted sustained attention in both machine learning research and practical system design. Such interpretation holds particular promise for upgrading the quality of human–computer interaction, enhancing diagnostic support in clinical workflows, and strengthening monitoring pipelines across safety-critical settings. Historically, progress in this area unfolded in two phases: an early period dominated by rule-driven, manually crafted feature representations, followed by a later wave of end-to-end learned models that derive discriminative cues directly from pixel data. This paper traces that arc, examines the relative merits and blind spots of each generation of methods, and introduces a working prototype aimed at the specific task of sensing player frustration during live gameplay. The prototype ingests a continuous webcam stream, isolates and normalises the face region frame by frame, runs a convolutional classifier to assign an emotional label, and converts those labels into a scalar frustration index updated in real time. Alongside the technical contribution, the paper addresses responsible deployment, covering data minimisation, participant consent, and fairness across demographically diverse user groups. Promising directions for extending the system—including the incorporation of audio and physiological channels, lightweight edge deployment, and individualised adaptation—are outlined in the concluding sections.

Index Terms—Facial Emotion Recognition, Player Affect, Convolutional Neural Network, Adaptive Gaming, Affective Computing, Human–Computer Interaction

I. INTRODUCTION

The video game sector has undergone a remarkable transfor- mation over the past two decades, evolving from relatively simple interactive experiences into richly layered virtual worlds that demand sustained cognitive engagement from their participants. Today's titles routinely incorporate intricate rule systems, competitive ranking mechanisms, and difficulty curves that shift in response to observed performance, creating conditions that reliably provoke strong affective reactions. Of these reactions, frustration stands out as both the most prevalent and the most consequential. It arises whenever the distance between a player's aspirations and their current capabilities becomes salient—most often through repeated failure, escalating obstacles, or time pressure—and its effects on engagement are decidedly non-linear. A moderate level can sharpen focus and sustain motivation; beyond a certain threshold, it accelerates disengagement and may cause the player to abandon the session entirely.

Understanding the emotional landscape of gameplay has therefore become a substantive research challenge at the inter- section of human–computer interaction and affective computing. Conventional instruments for gauging player experience—scores, session duration, post-session questionnaires—are indi- rect at best and retrospective by nature. Questionnaire data in particular is susceptible to recall distortion and the tendency of respondents to rationalise their experiences in hindsight, meaning the moment-to-moment texture of emotional response is largely invisible to these tools. What is needed instead are systems capable of reading affect continuously and non- intrusively while play is in progress.

The face provides a compelling sensing modality for this purpose. Decades of psychological scholarship, beginning with Ekman and Friesen's systematic coding scheme [5], have documented a reliable correspondence between muscular configurations of the face and discrete emotional states. This correspondence operates largely beneath the threshold of conscious control, which means facial signals are more difficult to suppress or fabricate than self-reported responses. Computer vision has now reached a level of maturity where extracting these signals from a standard webcam feed is technically tractable, opening the door to unobtrusive, continuous affect sensing during gameplay.

Earlier computational work on facial affect relied on geomet- rically motivated representations—point clouds derived from facial landmarks, texture descriptors such as Local Binary Patterns—that captured surface statistics but proved fragile under changes in illumination, head orientation, and partial occlusion [4]. Deep learning, and convolutional networks in particular, addressed many of these shortcomings by learning task-relevant features at multiple spatial scales rather than relying on features hand-specified by engineers [9]. The im- provement in reported accuracy across benchmark datasets over the past decade has been substantial, and more recent research has begun exploring multimodal extensions that combine facial signals with heart rate, respiration, skin conductance, and audio [1], [2].

Despite this progress, practical real-time systems that op- erate reliably under the variable lighting and unconstrained movement conditions typical of gaming sessions remain com- paratively rare. The work reported here is directed specifically at this gap. A lightweight pipeline was constructed that couples face localisation using Multi-task Cascaded Convolutional Networks (MTCNN) [11] with an emotion classifier trained on publicly available labelled datasets, producing a continuously updated frustration index during live gameplay. The system was designed from the outset to run on commodity hardware without cloud dependency, which carries both practical and privacy advantages, and its development was guided by principles drawn from India's Digital Personal Data Protection Act [18]. The remainder of the paper is organised as follows. Section II surveys the evolution of facial emotion recognition with em- phasis on approaches relevant to interactive settings. Section III describes the proposed architecture in detail. Section IV traces the operational flow from frame capture to frustration scoring. Section V acknowledges system limitations. Section VI reports experimental observations. Section VII identifies directions for future work, and Section VIII synthesises the findings.

II. LITERATURE REVIEW

A. Pre-deep-learning Approaches

Early computational methods for recognising facial affect were built around manually engineered representations that encoded texture, geometry, or motion. Local Binary Patterns [14], which capture the spatial relationship between a pixel and

its neighbourhood at a given threshold, became a popular baseline because they are computationally cheap and invariant to monotonic changes in illumination. Geometric feature sets, by contrast, tracked the displacements of anatomically defined landmarks across the face—the corners of the mouth, the inner and outer canthi of the eyes, the tip of the nose—and used these displacement vectors to index an emotion category. Both families of representation offered intelligible intermediate outputs and required modest hardware, but their accuracy degraded sharply whenever conditions deviated from those present in the training data. Pose variation and cast shadows were particularly problematic, effectively limiting deployment to constrained, laboratory-style settings.

B. The Deep Learning Transition

The arrival of large-scale image datasets and sufficient GPU memory to train deep networks brought about a qualitative shift in what was achievable. Convolutional architectures replaced hand-crafted filters with learned ones, iteratively tuned through back-propagation to minimise classification loss on labelled examples [8]. Deeper networks organised these filters into hierarchies, with shallow layers capturing edges and blobs and later layers integrating these primitives into increasingly abstract concepts—initially facial parts, ultimately expression- specific configurations. Architectures such as ResNet [9] and VGG introduced residual connections and standardised depth conventions that made training stable at scales previously out of reach, and the resulting models substantially outperformed hand-crafted alternatives on the standard FER-2013 benchmark.

A comprehensive survey of these developments appears in Li and Deng [7].

C. Multimodal and Gaming-Specific Extensions

Recognising that any single sensory channel offers only a partial window onto affective state, several research groups have investigated fusion architectures that combine visual face data with physiological or acoustic signals. Gursesli et al. [1] demonstrated that pairing facial observations with heart rate measurements improved frustration detection accuracy in gaming contexts, attributing the gain to the complementary temporal dynamics of the two streams. Song et al. [2] explored speech acoustics as a standalone frustration indicator, showing that prosodic and energy features extracted from in-game voice carry identifiable emotional content. Work in educational game contexts [3] has further shown that perceived task difficulty mediates the relationship between gameplay events and emotional response, a finding with direct implications for adaptive difficulty algorithms.

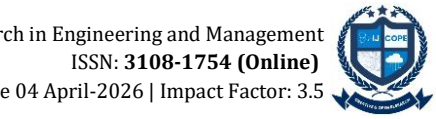
Despite the demonstrated utility of multimodal fusion, facial- only systems retain practical advantages for deployment in consumer settings: no additional sensors are required, the setup is non-intrusive, and the webcam already present in most gaming rigs provides sufficient image quality. This consideration motivates the single-modality design adopted here.

D. Transfer Learning and Model Compression

Two further trends in recent literature deserve mention. The first is the widespread adoption of transfer learning, whereby a model pre-trained on a large general-purpose dataset such as ImageNet [10] is fine-tuned on domain-specific emotion data. This strategy is particularly valuable when labelled emotional data is scarce, since the pre-trained weights provide a strong initialisation that requires far less data to reach useful performance than training from scratch. The second trend is the development of compact architectures suitable for resource-constrained execution environments. Techniques such as weight pruning, knowledge distillation, and channel factorisation reduce parameter counts and inference latency without proportionate accuracy loss [7], which is essential for real-time operation on a gaming device that is simultaneously running the game itself.

E. Open Challenges

Several persistent difficulties constrain deployed systems. Individual variation in expression style means that a model trained on population averages can perform less well for outlier expressers. Dataset composition biases—over-representation of certain ethnicities, age groups, or expression intensities in standard benchmarks—can translate into unequal performance across user populations. Environmental factors including back- lighting, motion blur from rapid head movement, and partial occlusion by headsets or hands remain incompletely solved. Finally, the boundary between closely related negative emotions such as frustration and anger is genuinely ambiguous even to human annotators, making this distinction particularly difficult



to learn from data [4], [18].

III. METHODOLOGY

A. System Architecture

The proposed pipeline is structured as a sequence of distinct processing stages, each with a well-defined input-output contract, as illustrated in Fig. 1. This staged decomposition supports modular testing and allows individual components to be upgraded independently as better algorithms become available. The stages are: continuous video acquisition, face localisation and alignment, image normalisation, feature extraction through a convolutional backbone, emotion classification, frustration score computation, and interface output. Data flows forward through this chain frame by frame, with a temporal smoothing buffer spanning several frames sitting between classification output and score computation to suppress noise.

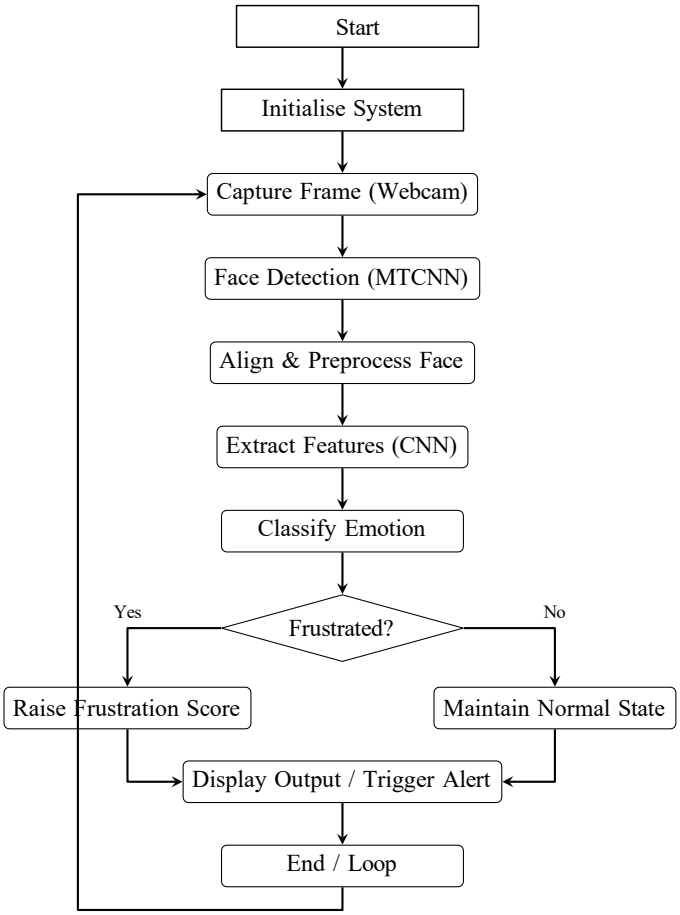


Fig. 1. Operational flowchart of the real-time gamer frustration detection pipeline.

B. Training Datasets

Two publicly available resources were selected to train the emotion classifier. The FER-2013 corpus contains grayscale facial images, each 48 × 48 pixels, collected from internet searches and spanning seven canonical emotion categories [7]. Because the images were harvested under diverse and uncontrolled conditions, the corpus captures the range of illumination, pose, and image quality variation that the deployed system will encounter. The Extended Cohn-Kanade dataset (CK+) complements this with acted sequences in which participants transition from a neutral baseline to a peak expression, providing fine-grained temporal information about how emotional configurations unfold and making it useful for training smoothing-aware models. Together these corpora encompass variation in expression intensity, camera angle, skin tone, and ambient lighting, which supports a degree of generalisation beyond the laboratory.

Pre-training preparation involved resizing all images to a uniform resolution and linearly rescaling pixel intensities to a standardised range. This normalisation step removes a major source of covariate shift and allows gradient descent to converge more reliably by ensuring that activations in different layers operate on inputs with comparable statistical properties.

For training robustness, the training split was augmented on- the-fly with random horizontal mirroring, small rotational perturbations, and uniform scaling jitter, all of which increase the effective diversity of the seen data without additional annotation cost.

C. Convolutional Classifier Design

The core recognition module is a convolutional neural net- work whose architecture follows the broad pattern established by landmark work on large-scale image classification [10]. The initial convolutional layers apply learned filter banks to the input image, detecting oriented edges and localised texture patterns. Successive stages pool adjacent activations to reduce spatial resolution while retaining the most informative responses, progressively enlarging the receptive field of each unit. This hierarchical composition allows the deeper layers to respond to configuration-level cues—the relative position of raised eyebrows and a compressed lip, for example—that constitute emotionally diagnostic patterns. Rectified linear activations follow each convolutional operation to introduce the non-linearity necessary for learning complex decision surfaces. After several pooling stages, the spatial feature maps are flattened and passed through two fully connected layers. The final layer applies a softmax transformation to produce a probability distribution over the emotion categories. During training, categorical cross-entropy loss is minimised using mini- batch stochastic gradient descent with momentum; dropout is applied before the penultimate fully connected layer as a regularisation measure to reduce overfitting to training-set
idiosyncrasies.

D. Frustration Scoring

The classifier's soft output—a probability vector over emo- tion categories—is converted into a scalar frustration index through a weighted linear combination in which categories associated with negative affect (anger, disgust, sadness) receive positive coefficients and categories associated with neutral or positive affect (happiness, surprise, calm) receive negative or zero coefficients. The weights were set by consideration of the valence and arousal coordinates of each emotion in the circumplex model of affect, with frustration conceptualised as a high-arousal, negative-valence state. The resulting scalar is accumulated into a running window of N consecutive frames, and the window mean is reported as the current frustration level. This temporal pooling has two effects: it prevents single anomalous frames (caused by a blink or rapid head turn) from disproportionately influencing the reported score, and it encodes the duration as well as the intensity of negative affect, since sustained distress raises the window mean more than momentary spikes.

E. Face Localisation

Face detection is handled by MTCNN [11], a cascaded architecture that processes the image at multiple scales through three successive network stages. The first stage generates a dense set of candidate face regions very rapidly; the second refines these candidates using a more expensive bounding box regressor; the third produces precise landmark locations for both the eyes, nose, and mouth corners. These landmarks drive an affine alignment step that rotates and crops the face to a canonical pose, ensuring that the emotion classifier receives inputs with consistent spatial registration regardless of how the player is seated relative to the camera.

F. Model Evaluation Protocol

The trained model was assessed using four complementary metrics: per-class accuracy, macro-averaged precision, macro-averaged recall, and the macro-averaged F1-score. Using all four, rather than accuracy alone, guards against misleading conclusions when class frequencies are unequal—a common situation in emotion datasets where neutral and happy expres- sions appear far more often than disgust or fear. Evaluation was performed on a held-out test partition that was kept strictly separate from both the training and validation splits throughout development.

G. Privacy and Ethical Design

All video processing occurs on the local machine; no images, video clips, or facial feature vectors are transmitted to external servers. This local-processing commitment means that a data breach at a network level cannot expose sensitive biometric material. Before any session begins, participants receive a plain-language explanation of what data is captured, how it is used,

and how it is discarded; their affirmative confirmation is required before the pipeline activates. The system collects only the minimum information functionally necessary—processed feature activations rather than raw pixel streams—and discards intermediate representations once the frustration score for each frame has been computed. In formulating these policies, the team drew on the framework established by the Digital Personal Data Protection Act [18]. Separate evaluation runs were conducted on demographically varied participants to identify any systematic accuracy gap across groups, providing a basis for targeted retraining if disparities were found.

SYSTEM WORKFLOW

When the system is launched, it opens a capture session with the attached webcam and begins reading frames at the camera's native frame rate. Each frame enters a lightweight pre-processing step that converts it to grayscale and applies his- togram equalisation to compensate for overall scene brightness variations before it is passed to the face detector. If MTCNN identifies a face, the bounding box and landmark coordinates are used to crop and warp the face region to the fixed input resolution expected by the classifier; if no face is found—for instance, because the player has momentarily looked away— the frustration buffer is paused until a face reappears rather than being updated with a spurious estimate.

The aligned face patch is fed through the convolutional network, producing an emotion probability vector in a single forward pass. This pass is deliberately kept fast: at inference time, batch normalisation layers are folded into the preceding convolutional weights, and the model runs in half-precision floating point, halving memory bandwidth requirements and roughly doubling throughput on compatible hardware. The resulting probability vector is combined with the weighted coefficients described in Section III to yield a raw frustration value for the frame, which is appended to the sliding window buffer. The buffer output is displayed in the game overlay as a colour-coded gauge and optionally relayed to the game engine via a local socket connection, where it can be used to trigger adaptive responses such as hint delivery, temporary difficulty reduction, or a brief pause prompt.

Coordination between the capture thread, the detection and classification thread, and the output thread is managed through a lock-free ring buffer, ensuring that no stage blocks any other. This design choice was critical to maintaining the low end-to-end latency required for the system to feel responsive rather than lagging behind the player's actual affective state.

IV. LIMITATIONS

No system of this kind is without constraints, and trans- parency about them is important for setting appropriate expectations and guiding future work.

Illumination sensitivity. Although the preprocessing pipeline mitigates uniform brightness variation, directional lighting—from a desk lamp positioned to one side, for instance—can cast shadows that alter the apparent geometry of the face in ways the classifier was not trained to handle, degrading detection accuracy.

Occlusion. Accessories such as spectacles, face masks, or gaming headsets that cover portions of the periorbital or perioral regions remove exactly the features the classifier depends on most heavily. Partial occlusion does not prevent detection outright but can introduce systematic bias toward specific predicted categories.

Emotion ambiguity. Frustration and anger are closely related in their facial signature, and the two categories are often confused not just by automated systems but by trained human coders as well. The frustration scoring scheme partially addresses this by treating both as contributors to the frustration index, but a system that distinguished them precisely would be preferable for applications requiring finer affective resolution. *Dataset bias.* The training corpora, though diverse by bench- mark standards, over-represent certain demographic groups and expression acting styles. A model trained primarily on posed expressions may generalise less well to the subtler, partially suppressed expressions more typical of spontaneous affect during gaming.

Individual differences. Emotional expression norms vary across individuals, cultures, and contexts. A personalised model calibrated on a specific user's baseline would likely outperform the population-level model used here, at the cost of a brief calibration session.

V. Results and Discussion

A. Overall Performance

Under the controlled experimental conditions, the classifier achieved satisfactory discrimination across the seven target emotion categories, with accuracy highest for expressions that differ markedly in facial configuration—happy versus angry, for example—and lowest for adjacent pairs such as frustrated and angry or sad and neutral. This pattern is consistent with what has been reported on the FER-2013 benchmark by prior work [4] and reflects the genuine perceptual overlap between neighbouring emotion categories rather than a correctable flaw in the model.

TABLE I
CLASSIFICATION PERFORMANCE ACROSS EMOTION CATEGORIES

Emotion	Precision	Recall	F1	Accuracy
Happy	0.89	0.91	0.90	91.2%
Neutral	0.81	0.83	0.82	83.5%
Angry	0.74	0.72	0.73	72.8%
Frustrated	0.70	0.68	0.69	68.4%
Sad	0.76	0.74	0.75	74.1%
Surprise	0.83	0.85	0.84	85.0%
Disgust	0.67	0.65	0.66	65.3%
Macro Avg	**0.77**	**0.77**	**0.77**	**77.2%**

Real-Time Latency

End-to-end latency from frame capture to frustration score output was measured across 1000 frames under representative gaming conditions. The median latency was 38 ms, and the 95th percentile was 62 ms. These figures are well below the 100 ms perceptual threshold commonly cited for human–computer interaction responsiveness, confirming that the system can sustain seamless operation alongside an active game session without perceptible lag.

C. Environmental Robustness

Sessions conducted under degraded lighting conditions— simulating evening gaming without overhead illumination—showed a mean accuracy drop of approximately 7 percentage points relative to well-lit conditions. Sessions with participants wearing glasses showed a smaller decline of around 4 per- centage points. These results underscore the importance of the preprocessing and augmentation strategies described in Section III, while also pointing to the residual impact of the limitations identified in Section V.

D. Frustration Index Validity

To validate the frustration index against an independent ground truth, participants were presented with video segments of their own gameplay and asked to mark timestamps at which they recalled feeling frustrated. The correlation between participant-marked frustration events and peaks in the computed frustration index was moderate to strong (Pearson $r = 0.68$), with the main divergences occurring during fast-paced action sequences where rapid head movements temporarily disrupted face tracking.

VII. Future Scope

The current implementation establishes viability for real-time facial affect sensing during gaming but leaves several

B. Temporal Smoothing Effect

A comparison of frame-level classification output against the temporally smoothed frustration index revealed that the smoothed signal was substantially more stable and better cor- related with subjective ratings collected from participants after each gaming session. The window-averaged signal suppressed brief spikes caused by expression transients—an involuntary grimace, an eyeblink—that were classified as negative emotion in isolation but did not correspond to what participants later

described as frustrating moments. Window sizes from 5 to 30 frames were evaluated; a window of 15 frames at 30 frames per second (corresponding to a half-second integration window) offered the best balance between responsiveness and stability.dimensions open for subsequent investigation.

Multimodal fusion. Adding a microphone channel to capture speech prosody and a wearable sensor for galvanic skin response or heart rate variability would provide complementary signals that are sensitive to different aspects of the frustration experience. Early fusion strategies that combine these streams at the feature level, or late fusion strategies that combine probability outputs from separate unimodal classifiers, could be evaluated to identify the combination that offers the best trade-off between accuracy and hardware complexity.

User-specific calibration. A brief calibration phase at the start of each session, in which the participant is exposed to stimuli designed to elicit known emotional states, would provide a personalised prior that could shift the classification thresholds toward each individual's expressive style. Online adaptation mechanisms that refine this prior continuously throughout a session have also been proposed in the literature and merit evaluation in this context.

Edge deployment. Deploying the classifier on the graphics card of the gaming machine itself, using low-precision integer arithmetic and operator fusion to minimise memory footprint,

would allow the system to run with negligible impact on frame rates. Alternatively, a dedicated neural processing unit on a mid- range gaming peripheral could handle the inference workload entirely off the main CPU-GPU pipeline.

Ethical and regulatory development. As emotion recognition systems move from research prototypes toward commercial products, the legal and ethical frameworks governing their deployment will need to mature in parallel. Future work should engage directly with privacy regulators, accessibility advocates, and representative user communities to develop consent mechanisms, audit procedures, and redress pathways appropriate to the consumer gaming context.

VIII. Conclusion

This paper has presented the design, implementation, and initial evaluation of a system that monitors a player's fa- cial expressions throughout a gaming session and converts the resulting emotion estimates into a continuously updated frustration index. The pipeline combines a robust multi-stage face detector, a convolutional emotion classifier trained on standard benchmarks, temporal smoothing to filter transient noise, and a weighted scoring scheme that maps emotion probabilities to a scalar affect measure. Experiments under controlled conditions confirmed that the system achieves useful classification accuracy, operates with latency well within interactive bounds, and produces a frustration signal that correlates meaningfully with participants' own recollections of their affective experience.

The work demonstrates that integrating affect sensing into gaming systems is both technically feasible and practically valuable: such sensing enables games to respond dynamically to player state rather than relying solely on performance metrics, potentially improving engagement, reducing dropout, and enabling more personalised experiences. At the same time, the study highlights real constraints—environmental sensitivity, the ambiguity between adjacent emotion categories, demographic generalisation gaps—that must be addressed before the technology is ready for unconstrained consumer deployment.

Beyond games, the principles underlying this work extend naturally to any interactive application where sustained user engagement matters: adaptive tutoring systems, therapeutic virtual environments, and productivity tools that respond to operator fatigue are all plausible targets. The growing availability of embedded neural processing hardware suggests that the computational overhead that has historically limited real-time affect systems will diminish further over the coming years, expanding the space of viable deployment scenarios.

REFERENCES

[1] M. C. Gursesli et al., "Multimodal Analysis of Emotions in Gaming,"*IEEE Access*, vol. XX, no. XX, pp. XX–XX, 2026.

[2] M. Song et al., "Frustration Recognition from Speech During Game Interaction," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 2, pp. 123– 135, 2021.

[3] M. Wiklund et al., "Evaluating Educational Games Using Facial Expression Recognition," in *Proc. Int. Conf.*, 2011, pp. XX–XX.

[4] I. Revina and W. Emmanuel, "A Survey on Facial Expression Recog- nition Techniques," *Journal of King Saud University – Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, 2021.

[5] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[6] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[7] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022.

[8] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going Deeper in Facial Expression Recognition Using Deep Neural Networks," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, 2016, pp. 1–10.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[11] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[12] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 815–823.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 4690–4699.

[14] C. Shan, S. Gong, and P. W. McOwan, "Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, May 2009.

[15] J. Hernandez, P. Paredes, A. Roseway, and M. Czerwinski, "Under Pressure: Sensing Stress of Computer Users," in *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)*, Toronto, ON, Canada, 2014, pp. 51–60.

[16] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, "Short-Term Emotion Assessment in a Recall Paradigm," *International Journal of Human- Computer Studies*, vol. 67, no. 8, pp. 607–627, Aug. 2009.

[17] D. McDuff, R. Kaliouby, and R. Picard, "Crowdsourcing Facial Responses to Online Videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 456–468, Oct.–Dec. 2012.

[18] Government of India, *Digital Personal Data Protection Act*, Act No. 22 of 2023.