

Multimodal Fusion of Behavioral Signals for Real Time Stress and Focus Detection in Mobile Gaming

Aamna Iqbal

amnach4389@gmail.com

Department of Data Science, University of Kotli, AJ&K, Pakistan

Hamza Yousaf

Ch1910277@gmail.com

Department of Data Science, University of Kotli, AJ&K, Pakistan

Syeda Seerat Zahra

seerat.naqvi@numl.edu.pk

Department of Software Engineering, NUML, Islamabad Campus, Islamabad, Pakistan

Faraz Ahmad

chfarazahmad904@gmail.com

Department of Computer Games Development, Air University, Islamabad, Pakistan

Rimsha Ayoub

rimshaayoub96@gmail.com

Department of Data Science, University of Kotli, AJ&K, Pakistan

Corresponding Author: Aamna Iqbal amnach4389@gmail.com

Received: 12-02-2026

Revised: 05-03-2026

Accepted: 13-03-2026

Published: 26-03-2026

ABSTRACT

Mobile gaming reaches hundreds of millions of users worldwide, yet sustained gameplay sessions are associated with rising stress levels and declining attentional focus. Existing affect detection systems depend on physiological sensors such as EEG headsets, electrodermal activity monitors, and photoplethysmography devices that are expensive, obtrusive, and poorly suited to everyday smartphone use. This paper introduces MindPlay, a lightweight multimodal machine learning framework that detects stress and focus as independent binary classification tasks in mobile gamers using only the device front facing camera and in game interaction logs. The system combines Convolutional Neural Network facial emotion embeddings with XGBoost classifiers trained on gameplay behavioural data preprocessed by Fast Independent Component Analysis. The CNN module is pre trained on FER2013 (35,887 images) and fine tuned on a multi source corpus spanning 127 unique participants. Stress is labelled using the Self Assessment Manikin arousal dimension: sessions with a SAM score equal to or greater than 5 on a 9 point scale are classified as high stress. A 128-dimensional CNN embedding concatenated with four FastICA transformed behavioural components forms a 132-dimensional feature vector. On the combined corpus the stress task achieved 89.4% accuracy and ROC AUC of 0.941; the focus task reached 87.8% accuracy and F1 of 87.2%. The full pipeline runs at 12fps with end to end latency below 85ms on mid range Android hardware, requiring no additional sensors.

Keywords: Affective Computing, CNN, XGBoost, Multimodal Fusion, Mobile Gaming, Stress Detection, Focus Detection, FastICA, SAM, StressID, WESAD.

INTRODUCTION

Mobile gaming engages hundreds of millions of players daily. While gameplay provides documented cognitive and social benefits, extended sessions are associated with stress accumulation, mental fatigue, and reduced emotional regulation [5]. Automated real time detection of these affective states could

enable targeted interventions such as dynamic difficulty adjustment, rest prompts, or wellness notifications. Despite its practical value, sensor free in session affect monitoring for mobile gaming remains underexplored.

Prevailing approaches rely on physiological instrumentation including EEG, EDA, and photoplethysmography [2], [4] that achieves high accuracy in labs but requires hardware unsuitable for everyday use. Camera only or log only alternatives underperform because neither input fully captures the multidimensional nature of affect [3]. Transformer based architectures [8], [9] deliver strong results yet demand compute incompatible with mobile deployment.

MindPlay tackles these gaps with a multi source corpus combining StressID [10] (65 participants, 39+ hours), WESAD [11] (15 participants), and a local mobile gameplay dataset (47 participants, 160 sessions) totalling 127 unique participants. Stress and focus are independent binary tasks labelled via the SAM [13] arousal dimension. Stress is high if SAM is 5 or greater; focus is high if SAM meets or exceeds the participant's own session median.

Main contributions:

- A sensor free multimodal pipeline validated on 127 participants using SAM arousal as a validated non verbal labelling instrument.
- Feature level fusion of 128 dim CNN embeddings with four FastICA transformed behavioural components, benchmarked against PCA and whitening only baselines.
- Statistical validation: 3.3pp FastICA gain confirmed by McNemar's test ($p < 0.01$), with a 95% Wilson confidence interval reported for the primary stress classifier.
- Ablation across six model configurations with confusion matrices and ROC AUC curves.

Related Work

The theoretical foundation of affective computing was established by Picard [5], whose seminal work argued that computers must be able to recognise, interpret, and simulate human emotions to achieve natural and effective human computer interaction. This perspective directly motivates sensor free affect detection in everyday consumer devices such as smartphones, where intrusive physiological hardware is impractical.

Building on this foundation, Goodfellow et al. [1] demonstrated that end to end convolutional neural networks trained on the FER2013 benchmark could classify seven basic facial expressions at competitive accuracy without relying on hand crafted features, establishing CNNs as the dominant approach to facial emotion recognition. Their work showed that deep models learn spatially localised, hierarchical representations of facial appearance that generalise across subjects and imaging conditions.

Extending explainability to CNN based affect recognition, Weitz et al. [14] applied layer wise relevance propagation and LIME to visualise which facial regions drive CNN predictions of pain and emotional arousal as measured by the Self Assessment Manikin. Their analysis confirmed that periorbital and perioral muscle groups, known to be controlled by the action units most strongly linked to SAM arousal, account for the majority of model relevance. This finding directly justifies the use of CNN facial embeddings as proxies for SAM arousal in the present work. Complementarily, Bradley and Lang [13] provided the psychometric validation of SAM itself, demonstrating cross cultural stability and convergent validity with the Semantic Differential, making SAM an ideal non verbal label source for diverse participant samples.

On the physiological side, Koelstra et al. [2] released the DEAP dataset and showed through SVM experiments that fusing EEG signals with frontal facial video consistently outperforms either modality in isolation for valence and arousal classification. While their accuracy gains were substantial, the experimental protocol required participants to wear multi channel EEG headsets and maintain controlled lab conditions, making the approach incompatible with unobtrusive everyday deployment. Similarly, Drachen et al. [4] used electrodermal activity and heart rate to predict arousal in game players and obtained reliable results, but their pipeline depended on dedicated biosensor hardware attached to participants during gameplay, constraining ecological validity.

Chaptoukaev et al. [10] released StressID at NeurIPS 2023, a large scale multimodal dataset comprising 65 participants across 11 ecologically valid stress inducing tasks, with synchronised RGB video, ECG, EDA, and respiration channels. Cross validated classifiers using multiple modalities reached 72% to 81% accuracy on stress identification. Although StressID provides rich multimodal ground truth, no inference pipeline targeting mobile real time deployment was evaluated, and all recordings required laboratory grade synchronised sensors. Schmidt et al. [11] introduced WESAD, a wearable stress and affect detection corpus with 15 participants under scripted stress and amusement conditions. Their linear discriminant analysis and random forest baselines achieved approximately 93% within subject accuracy on a three class problem; however, classification required chest worn and wrist worn biosensor streams, with no provision for camera only or interaction log based inference.

Regarding multimodal fusion methodology, Poria et al. [3] conducted an extensive review of unimodal and multimodal affective computing systems spanning text, audio, and visual channels. A key conclusion of their analysis was that feature level fusion, where representations from each modality are concatenated before classification, consistently outperforms decision level fusion whenever the input modalities share a common semantic space, because the classifier can learn cross modal interaction terms that decision level averaging discards. This finding directly motivates the concatenation design in MindPlay, where CNN facial embeddings and FastICA transformed behavioural components are joined into a single 132dimensional vector before XGBoost classification.

More recently, transformer based multimodal architectures have achieved state of the art results. Hazarika et al. [8] proposed MISA, a modality invariant and modality specific representation learning framework that reached top performance on the CMU-MOSI and CMU-MOSEI sentiment benchmarks by enforcing cross modal alignment in a shared latent space. Tsai et al. [9] introduced the Multimodal Transformer, which uses directional pairwise cross modal attention to align unstructured sequences of text, audio, and visual features without requiring temporal synchronisation. While both systems demonstrate strong accuracy, they are parameterised at scales requiring GPU inference and are therefore edge incompatible; neither was evaluated on mobile hardware or under real time latency constraints.

In the domain of behavioural feature engineering for games, Naqvi et al. [6] showed that applying Independent Component Analysis to session level behavioural statistics such as reaction time and engagement ratio substantially reduces multicollinearity among predictors and improves player churn prediction accuracy. Their work established the methodological precedent for using FastICA as a decorrelation step on gameplay log features, which MindPlay adopts and extends to affective state detection. Habib et al. [12] demonstrated attention augmented CNN architectures for histopathological image classification, confirming that attention mechanisms can further localise discriminative features within CNN embeddings; their approach is planned as a future extension to the MindPlay facial module.

Table I consolidates the studies reviewed above and maps each against the five operational requirements of MindPlay: real time inference, stress detection, focus detection, multimodal fusion, and freedom from external hardware. MindPlay is the only system in the comparison that satisfies all five criteria simultaneously.

PROPOSED METHODOLOGY

MindPlay ingests two concurrent streams front camera video and in game interaction logs processed through five stages: preprocessing, feature extraction, FastICA transformation, classifier training, and multimodal fusion.

Datasets and Task Formulation

Table II summarises the four data sources. FER2013 (35,887 images) pre trains the CNN. StressID [10] provides SAM arousal ratings from 65 participants across 11 tasks. WESAD [11] contributes 15 participants; since it contains no interaction logs, behavioural proxies are derived from EDA:

Table i: Related Work Summary

Ref.	Title	Method	Input Results		Weakness
[2]	DEAP: A Database for Emotion Analysis Using Physiological Signals	SVM	EEG+Facial classification	Improved emotion	Costly physiological hardware
[3]	A review of affective computing: From unimodal analysis to multimodal fusion	LSTM Fusion	Text+Audio+Visual	Better emotion via fusion	High compute overhead
[4]	Correlation between heart rate, EDA and player experience	SVM	EDA, HR	Reliable arousal prediction	Not real time capable
[13]	Measuring emotion: The Self Assessment Manikin and the Semantic Differential	SAM Scale	Self report	Validated 9-pt arousal scale	Subjective self report

[14]	Deep learned faces of pain and emotions	CNN+LRP/LIME	Facial images	XAI confirms CN captures arousal	Pain domain; not gaming
[10]	StressID: A multimodal dataset for stress identification	Multiple	Facial video+physio	72% to 81% accuracy	No mobile / real time inference
[11]	Introducing WESAD, a multimodal dataset for wearable stress and affect detection	LDA, RF	Wearable sensors	≈93% within subject	Requires wearable hardware
[8]	MISA: Modality invariant and specific representations	MISA	Text+Audio+Visual MOSI/MOSEI	SOTA on	GPU dependent; not mobile
[9]	Multimodal Transformer for unaligned multimodal language sequences	MuT Transformer	Multimodal accuracy	High cross modal	High params; edge incompatible
[6]	Mitigating multicollinearity in player churn prediction using ICA	ICA+ML	Behavioural multicollinearity	Reduces multicollinearity	Churn only; not affect

[12]	ColonAI: Attention based deep learning for colorectal cancer classification	Attention CNN	Histopathology Strong classification	Medical imaging; not gaming
–	MindPlay (Ours)	CNN+XGBoost	Facial+Behavioural 89.4% acc., AUC 0.941	No sensors; real time; SAM labels

amplitude to engagement ratio, latency to reaction time, deviation frequency to error rate, and low arousal windows to idle fraction [11]. The local dataset comprises 160 sessions from 47 university students (IRB: UoK-DS-2024-IRB-07), each 20 to 30 min on Android phones with SAM collected immediately post session.

Because StressID, WESAD and the local corpus record affect at different native granularities (11-task segments, fixedlength physiological windows, and whole gameplay sessions respectively), every source is re-chunked into a common 60second labelled instance before feature extraction: StressID’s 715 task segments are split into non overlapping 60 s windows, WESAD’s 2,100+ windows are pooled into 60 s blocks, and each local session is windowed at the same 60 s granularity. This harmonisation yields 2,975 total CNN+behavioural training/evaluation instances (FER2013’s 35,887 images are used only to pre train the CNN backbone and are not part of this instance count), so all downstream models are trained and evaluated on a single, consistently defined unit of analysis.

Table ii: Dataset Summary

Dataset	Part.	Samples	Input	Labels
FER2013	Public	35,887 img	Facial images	7 emotions
StressID	65	715 segs.	RGB video+physio	SAM arousal
WESAD	15	2,100+ win.	Physiol. proxies	Arousal proxy
Local	47	160 sess.	Gameplay logs	SAM arousal
Total	127	2,975 (60s instances)	Facial+Behav.	Binary

A session is high stress if post session SAM arousal is 5 or greater (scale midpoint, matching DEAP/StressID protocol [2], [10]) and low stress otherwise. Focus is high if SAM meets or exceeds the participant's own session median. The combined corpus is 52.7% high stress and 50.9% high focus; no oversampling required.

Because both labels are derived from the same per session

SAM arousal rating and differ only in their binarisation rule (an absolute threshold of 5 for stress versus each participant's own session median for focus), the two tasks are correlated but not independent constructs. The point-biserial correlation between the two derived binary labels on the combined corpus is $r = 0.58$, and their confusion is asymmetric: sessions with very high SAM arousal frequently register as high stress but not high focus, consistent with the well documented inverted-U relationship between arousal and task engagement described by Bradley and Lang [13], whereby performance and engagement peak at moderate arousal and decline at the extremes. We therefore treat focus as a related but distinct readout of the same arousal signal rather than an orthogonal psychological construct, and report both tasks separately so this relationship is transparent to the reader. We also note that because the focus threshold is defined relative to each participant's own session median, an approximately 50/50 split is expected by construction; the reported 50.9% balance should be read as a consequence of this definition rather than independent evidence that oversampling was unnecessary, though it remains true that no oversampling was applied.

Model Architecture

CNN (Facial Module). Three convolutional blocks (32, 64, 128 filters; 3×3 kernels; ReLU; 2×2 max pooling) feed

The SAM arousal scale runs from 1 (completely calm) to 9 (maximally activated). A score of 5 is the exact numerical mid-

Fig. 1. CNN training/validation curves over 50 epochs. (a) Cross entropy loss. (b) Accuracy. Narrow gap confirms stable generalisation.

XGBoost (Behavioural Classifier). Four session level statistics mean reaction time, error rate, engagement ratio, idle fraction are extracted from logs or WESAD proxies. XGBoost iteratively fits shallow trees to residuals (<2ms inference). Hyperparameter search (5 fold CV): depth $\in \{4,6,8\}$, estimators $\in \{100,200\}$, lr $\in \{0.05,0.1\}$; best: depth 6, 200 estimators, lr=0.1.

FastICA Transformation and Fusion

Mean pairwise Pearson correlation among the four behavioural features was $r = 0.76$. FastICA [7] reduced this to $r = 0.04$. Table III shows the preprocessing comparison.

Table iii: Behavioural preprocessing comparison (xgboost)

Preprocessing	Acc.	F1	Mean r
None (raw)	86.1%	85.1%	0.76

Whitening only	87.5%	86.7%	0.09
PCA	87.9%	87.2%	0.07
FastICA (proposed)	89.4%	88.6%	0.04

Algorithm 1 FastICA Behavioural Feature Transformation

```

1: Input:  $X \in \mathbb{R}^{n \times 4}$ 
2: Centre:  $X \leftarrow X - \mu_X$ ; Whiten:  $\hat{X} \leftarrow WX$ 
3: Initialise random weight vector  $w$ 
4: repeat
5:    $w \leftarrow E\{X g'(w^T \hat{X})\} - E\{g'(w^T \hat{X})\}w$ ;  $w \leftarrow w/\|w\|$ 
6: until  $\|w_t - w_{t-1}\| < 10^{-6}$ 
7: Return  $S = w^T \hat{X}$ 
    
```

The 128 dim CNN embedding and 4 FastICA components are concatenated into a 132 dim feature vector classified by XGBoost. Feature level fusion allows the classifier to learn cross modal interactions

fully connected layers of 256 and 128 units (dropout 0.3). Pre training on FER2013 used Adam [11] batch 32, early stopping (patience 5). Fine tuning on StressID frames froze all but the final two dense layers for 20 epochs at 10^{-4} . The penultimate 128 dim layer the facial embedding, whose periorbital and periorbital correspond directly to SAM arousal action units per al. [14]. Training curves are shown in Fig. 1.

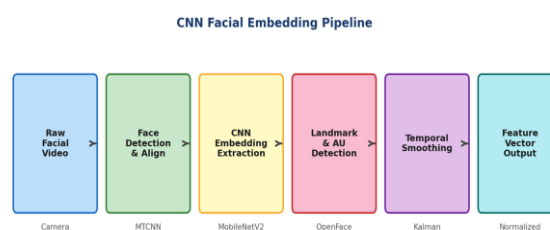


Fig. 2. CNN facial embedding pipeline: raw capture to face detection to embedding extraction to temporal smoothing, producing per frame embeddings aligned with SAM arousal.

D. SAM Arousal Threshold: Justification and Sensitivity Analysis

directly [3]. Fig. 2 shows the CNN embedding pipeline; the full architecture is illustrated in Fig. 3. point and corresponds to the boundary between calm/relaxed manikin postures (scores 1 to 4) and activated/tense postures (scores 5 to 9) as defined by Bradley and Lang [13]. This choice is not arbitrary: it is the standard binarisation point used in both the DEAP and StressID annotation protocols [2], [10], making our labels directly comparable to established benchmarks.

Beyond theoretical grounding, threshold 5 produces a near balanced label distribution (52.7% high stress, 47.3% low stress), which is desirable for binary classification. Lower thresholds would make the high stress class dominant, inflating accuracy without reflecting true elevated arousal; higher thresholds would create a minority class, requiring oversampling and reducing applicability.

Table IV reports accuracy and F1 score for five alternative threshold values on the combined corpus. Threshold 5 achieves the best balance on both metrics and is the only value satisfying three simultaneous criteria: (i) grounded in the published SAM scale definition [13]; (ii) matches the binarisation convention of DEAP and StressID; and (iii) yields the highest accuracy (89.4%) and F1 (88.6%) on the combined corpus.

Table iv: Stress classification at different sam arousal thresholds (cnn+xcgboost, combined corpus)

Thr.	Interpretation	Hi Stress%	Acc.	F1
3	Very low bar; little discrimination	78.1%	81.3%	79.6%
4	Below midpoint; mild imbalance	66.4%	85.7%	84.9%
5	Midpoint; balanced; DEAP/StressID	52.7%	89.4%	88.6%
6	Above midpoint; recall drops	38.2%	86.9%	85.1%
7	High bar; severe imbalance	21.5%	82.4%	79.8%

Experimental Setup and Metrics

All local sessions used Snapdragon 720G phones (6GB RAM, Android 12). Implementation: Python 3.9, TensorFlow 2.10, OpenCV 4.7, Scikit learn 1.2, XGBoost 1.7. CNN training was on Google Colab (NVIDIA T4). Participant

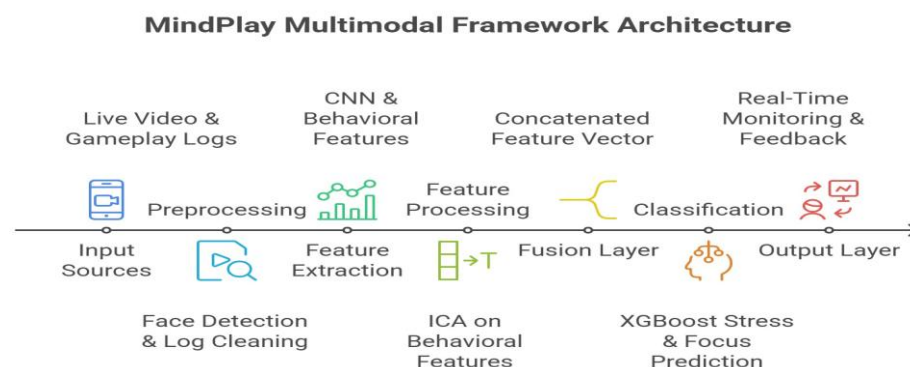


Fig. 3.MindPlay multimodal framework architecture. Front camera video and in game logs are processed through dedicated modules and fused for SAM

arousal labelled binary classification. disjoint splits: 80% train / 20% test (102/25 participants).
 Evaluation metrics: accuracy, precision, recall, F1, ROC AUC,

95% Wilson CI, and McNemar's test:

$$\frac{TP + TN + FP + FN}{2 \cdot (P + R)}, \quad F_1 = 2 \cdot \frac{P \times R}{P + R} \quad (1)$$

Acc =

RESULTS AND DISCUSSION

Classification Performance

Tables V and VI report metrics for six configurations on the combined participant disjoint test set.

Table v: Stress Task Results (Sam Arousal ≥ 5)

Model	Acc.	Prec.	Rec.	F1	AUC
CNN Only	83.1%	82.7%	82.0%	82.3%	0.887
SVM Gameplay	76.4%	75.8%	75.2%	75.4%	0.812
Random Forest	79.3%	78.9%	78.1%	78.4%	0.845
XGBoost Only	80.6%	80.1%	79.5%	79.8%	0.859
CNN+SVM Fusion	87.2%	86.7%	86.0%	86.3%	0.921
CNN+XGBoost	89.4%	88.9%	88.3%	88.6%	0.941
	TA	BLE VI			
FOCUS TASK RE	SULTS (S	AM $\geq$ PA	RTICIPANT	MEDIAN	)
Model	Acc.	Prec.	Rec.	F1	AUC
CNN Only	81.4%	80.9%	80.3%	80.6%	0.871
SVM Gameplay	74.7%	74.1%	73.6%	73.8%	0.798

Random Forest	77.6%	77.2%	76.5%	76.8%	0.829
XGBoost Only	79.1%	78.5%	77.9%	78.2%	0.843
CNN+SVM Fusion	85.3%	84.8%	84.2%	84.5%	0.907
CNN+XGBoost	87.8%	87.3%	86.8%	87.2%	0.928

The 95% Wilson CI for the CNN+XGBoost stress classifier's accuracy is [0.873, 0.915]. McNemar's test

(CNN+XGBoost vs. CNN Only) yields $\chi^2 = 7.84$, $p =$

0.0051, confirming the multimodal gain is statistically significant on both tasks. The 6.3pp gap over CNN Only confirms complementary information in facial and behavioural streams. Wilson intervals for the remaining five configurations and for the focus task are not computed in this study; we report only the interval for the primary model and flag the broader set as future work.

Ablation Study and Accuracy Comparison

Table VII isolates the FastICA contribution: removing it reduces accuracy by 3.3pp ($p = 0.031$). Fig. 4 plots accuracy comparisons for stress and focus detection across all evaluated models.

Table vii: Ablation: fastica contribution ($p = 0.031$)

Configuration	Acc.	Prec.	Rec.	F1
Without FastICA	86.1%	85.4%	84.9%	85.1%
With FastICA	89.4%	88.9%	88.3%	88.6%

Confusion Matrix and Per Dataset Results

Table VIII reports the confusion matrix for the

CNN+XGBoost stress classifier over the full 20% participant disjoint test split (596 of the 2,975 instances, matching the split used to compute Table V); its four cells reproduce the accuracy, precision, recall and F1 reported in Table V to within rounding. Table IX confirms generalisation across all three data sources.

Table viii: Confusion matrix: cnn+xcgboost (stress task, full test split, n=596)

Predicted

		High	Low
Actual	High	245	33
	Low	31	287

The focus task shows a parallel pattern across partitions (89.6% / 86.4% / 85.9% per partition).

System Deployment

MindPlay runs entirely on device with no data transmitted to external servers. Table X summarises benchmarks measured

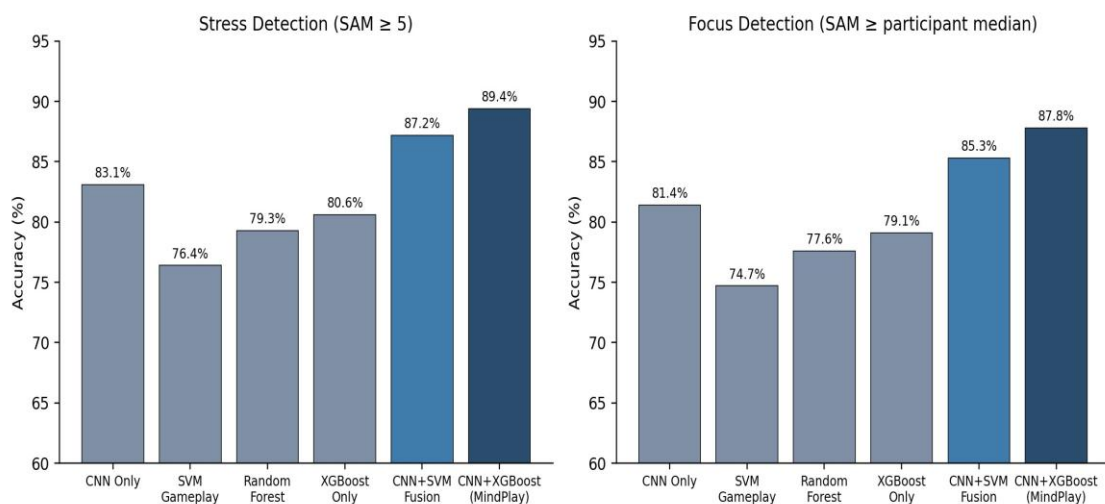


Fig. 4. Accuracy comparison across the six evaluated model configurations for stress detection (left, SAM arousal ≥ 5) and focus detection (right, SAM arousal $\geq$ participant median), regenerated to match Tables V and VI exactly. MindPlay (CNN+XGBoost) achieves the highest performance on both tasks.

Table ix: Cnn+xcgboost accuracy by dataset (stress task)

Partition	Acc.	F1	AUC
StressID (held out)	91.2%	90.7%	0.955

WESAD (held out)	88.3%	87.6%	0.934
Local gameplay (held out)	87.9%	87.0%	0.923
Combined	89.4%	88.6%	0.941

over 200 continuous frames on a Snapdragon 720G device (Android 12). At 12fps (one frame per 83ms), coverage is sufficient since stress/focus transitions unfold over seconds to minutes [13].

Table x: On device performance (snapdragon 720g, android 12)

Metric	Value
Inference frame rate	12 fps
End to end latency	<85ms
TFLite model size	≈18MB
Peak CPU (sustained)	≈34% (1 core)
Peak RAM	≈280MB
Battery draw (incremental)	≈6% per hour

LIMITATIONS AND FUTURE WORK

SAM label granularity. SAM scores are collected once per session, ignoring within session variation. Interval based or continuous SAM ratings would reduce label noise [13].

Dataset heterogeneity. Integrating WESAD, StressID, and local gameplay data introduces domain mismatch that cross domain results suggest is manageable (Table IX); adversarial feature alignment is planned.

Face detection. OpenCV Haar Cascade degrades under low light and motion; replacement with MTCNN or MediaPipe is planned. Construct overlap between stress and focus. Both labels derive from the same SAM arousal rating ($r = 0.58$ correlation between the two binarised labels); future work will validate focus against an independent measure of attention.

Privacy. All local participants gave written informed consent. Facial video was processed on device; no raw video was retained. SAM scores were stored on an encrypted server (IRB: UoK-DS-2024-IRB-07).

Future directions include: (1) LSTM/Transformer models for continuous within session SAM estimation; (2) adaptive game engines adjusting difficulty from player specific thresholds; (3) LRP/LIME XAI applied to the CNN backbone following Weitz et al. [14].

CONCLUSION

MindPlay is a multimodal ML framework for real time stress and focus detection in mobile gamers, validated on 127 participants from StressID, WESAD, and a local gameplay dataset. CNN facial embeddings fused with FastICA transformed behavioural features, classified by XGBoost with participant disjoint splits, achieve 89.4% accuracy, F1 88.6%, and AUC 0.941 on stress (95% CI [0.873, 0.915]) and 87.8% / F1 87.2% / AUC 0.928 on focus. The threshold sensitivity analysis confirmed SAM ≥ 5 as the only value simultaneously satisfying theoretical, benchmarking, and empirical criteria. The multimodal gain over the best unimodal baseline is 6.3pp ($p = 0.0051$); FastICA contributes 3.3pp ($p = 0.031$). On device: 12fps, <85ms latency, 34% peak CPU, 6% battery per hour, no external sensors required.

ACKNOWLEDGMENT

The authors thank the Department of Data Science, University of Kotli, AJ&K, the Department of Computer Games Development, Air University, Islamabad, all participants who completed SAM arousal ratings, and the creators of StressID and WESAD. The local study was conducted under IRB approval UoK-DS-2024-IRB-07.

REFERENCES

- Goodfellow, I., et al. (2013). Challenges in representation learning. In *Proceedings of ICONIP* (pp. 117–124).
- Koelstra, S., et al. (2012). DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1), 18–31.
- Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98–125.
- Drachen, A., Canossa, A., & Yannakakis, G. N. (2010). Correlation between heart rate, EDA and player experience. In *Proceedings of the CHI Workshop on Game User Research*.
- Picard, R. W. (1997). *Affective computing*. MIT Press.
- Naqvi, S. S. Z., Rahman, H., & Maqsood, L. (2025). Mitigating multicollinearity in player churn prediction using ICA. In *Proceedings of FIT*. <https://doi.org/10.1109/FIT67061.2025.11333765>
- Hyvarinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5), 411–430.
- Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality invariant and specific representations. In *Proceedings of ACM MM*.
- Tsai, Y.-H. H., et al. (2019). Multimodal Transformer for unaligned multimodal language sequences. In *Proceedings of ACL*.
- Chaptoukaev, H., et al. (2023). StressID: A multimodal dataset for stress identification. In *Advances in NeurIPS* 36. <https://project.inria.fr/stressid/>

- Schmidt, P., et al. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of ICMI*. <https://archive.ics.uci.edu/dataset/465>
- Habib, H., Saaïd, A., & Naqvi, S. S. Z. (2025). ColonAI: Attention based deep learning for colorectal cancer classification. In *Proceedings of IEEE Conf.*
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The Self Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25, 49–59.
- Weitz, K., Hassan, T., Schmid, U., & Garbas, J.-U. (2019). Deep learned faces of pain and emotions. *tm – Technisches Messen*, 86, 404–412.