

Toward Objective Mental Health Screening via Facial Expression Analysis: A Hybrid Deep Learning Proof-of-Concept

Endah Setyowati ^{1**}, Ahmad Fauzi ^{2*}, Galura Muhammad Suranegara ^{3*}, Mohd Herwan Sulaiman ^{4***}, Fauzzi Fallah Ramadhani ^{5*}, Sahra Raditiya Fadillah ^{6*}, Duke Ursula Zahra ^{7*}

* Telecommunication Systems, Universitas Pendidikan Indonesia

*** Fakulti Kejuruteraan Elektrik Elektronik, Universiti Malaysia Pahang

endahsetyowati@upi.edu ¹, ahmad.fauzi@upi.edu ², galurams@upi.edu ³, herwan@umpsa.edu.my ⁴, fauzzifallah.8@upi.edu ⁵, sahra.raditiya13@upi.edu ⁶, dukeursula.123@upi.edu ⁷

Article Info

Article history:

Received 2026-09-07

Revised 2026-10-09

Accepted 2026-10-18

Keyword:

Deep Learning,
Depression Screening,
Facial Landmark Analysis,
Grad-CAM,
Vision Transformer.

ABSTRACT

Depression remains among the largest contributors to global disease burden, yet early screening in Indonesia still relies heavily on self-report instruments prone to subjective bias. This study presents "Beyond Expression," a hybrid architecture combining a Convolutional Neural Network (ResNet-50), a Swin Transformer, and a Spatiotemporal Attention Module (TSAM) to detect depression-related facial patterns from anonymized 3D facial landmark sequences, intended as a first step toward future objective screening tools rather than a clinically validated instrument. Evaluated via 5-fold stratified cross-validation on 142 participants, the model achieved an average accuracy of $66.95\% \pm 5.49\%$ and macro F1-score of 0.5102 ± 0.0969 . Against a logistic regression baseline (F1 = 0.5091), a paired Wilcoxon signed-rank test and paired t-test found no statistically detectable difference, though the small fold count ($n = 5$) limits test power. Sensitivity remained low ($24.17\% \pm 20.48\%$), and a cross-frame Grad-CAM analysis across four representative participants showed Healthy-class activation staying relatively stable across frames (correlation up to 0.80), while Depressed-class activation was markedly less consistent (correlation near zero) and partially collapsed in one case, closely matching the model's low sensitivity. A same-dataset held-out check on withheld DAIC-WOZ test participants ($N = 47$) gave a single-run sensitivity of 57.14% and specificity of 72.73%, higher than most cross-validation folds but not yet confirmed stable across runs. The authors treat these results as genuine findings from a proof-of-concept stage, distinct from a validated clinical diagnosis, meant to guide refinements in data collection and model design for future iterations ahead of clinical validation.



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.

I. INTRODUCTION

Recent global estimates indicate that over one billion people worldwide are living with a mental health condition, with depression and anxiety standing out as the two largest contributors to Years Lived with Disability and to an estimated global economic burden of US\$1 trillion annually [1], [2]. Within Indonesia specifically, the Ministry of Health's 2018 Basic Health Research reported that 9.8% of people above the age of 15, equivalent to more than 19 million individuals, were categorized as having a mental emotional

disorder [3], while low- and middle-income countries such as Indonesia face a substantial mental health workforce shortfall, with such countries averaging only 22.3 mental health professionals per 100,000 population against the service levels needed to meet demand [4].

Widely used screening tools such as the PHQ-9 and GAD-7 rely on self-report, which leaves them exposed to subjectivity bias and to the stigma that often surrounds mental health disclosure [5]. Even trained clinicians who observe facial expression qualitatively face a similar constraint: signs such as blunted affect or fleeting micro-expressions of

sadness tend to be too brief and too subtle for consistent human detection, a pattern echoed in facial action-unit research on subthreshold depressive states [6]. These limitations underline a growing need for a screening approach that is objective, non-invasive, and scalable.

Artificial intelligence has increasingly been explored as a way to support earlier detection and intervention in mental health care [7], including conversational-agent systems that adapt their questioning based on detected psychological distress [8]. Earlier work has applied hybrid CNN–Vision Transformer architectures and related deep learning approaches to identify visual indicators of mental disorders, reporting accuracies ranging from roughly 60% to 90% across different datasets and modalities [9], [10], [11], alongside findings that depression correlates with measurable non-verbal shifts in facial expression and body posture, particularly a higher frequency of negative facial action units among affected individuals [12]. What sets the present study apart is the combination of a Spatiotemporal Attention Module (TSAM) with a CNN–Swin Transformer backbone, together with a Grad-CAM interpretability layer [13] applied directly to landmark coordinates rather than raw images. This design allows the model's decisions to be inspected without ever requiring access to the participant's actual facial images, keeping patient privacy intact.

This paper documents the outcomes of the basic-research (proof-of-concept) stage of the "Beyond Expression" architecture. It covers cross-validation performance, a clinically oriented sensitivity and specificity analysis, decision-threshold tuning experiments, and a Grad-CAM interpretability study extended to check consistency across frames, together with an initial held-out generalization check on the withheld DAIC-WOZ test subset.

II. METHODS

A. Dataset and Preprocessing

The study draws on DAIC-WOZ (Distress Analysis Interview Corpus, Wizard of Oz) [14], [15], a semi-structured clinical interview dataset collected through a Wizard-of-Oz virtual interviewer, where distress labels come from PHQ-8 scores converted into binary Healthy/Depressed categories [16], [17]. The train and dev subsets were pooled together into a combined set of 142 participants (100 Healthy, 42 Depressed) for cross-validation, while the test subset was withheld from cross-validation and later used for a separate held-out generalization check reported in Section III.E. From each frame, 68 three-dimensional facial landmark points were extracted and mapped into a 224×224-pixel pseudo-image through normalized coordinate mapping, a representation that never involves storing the original facial pixels.

Each valid frame's 68 landmarks' two-dimensional (x, y) coordinates were independently min-max normalized per frame to fill the 224×224 canvas, so that the resulting point cloud reflects the participant's facial configuration regardless of head position or camera distance within that frame. Each

landmark point was then rendered as a small filled circle on an initially blank canvas, with its color encoding two additional cues beyond spatial position: one color channel encodes the landmark's depth (z) coordinate, clipped and rescaled to a 0–255 range, and a second channel encodes the landmark's index (0–67), so that anatomically distinct facial regions receive systematically different colors across the point cloud; the third channel is held constant. It is this color-coded landmark canvas, rather than the raw facial image, that is passed to the CNN backbone, which is what allows the interpretability analysis in Section III.D to be read purely in terms of landmark geometry and depth rather than visual texture, skin tone, or facial appearance.

B. Model Architecture

"Beyond Expression" is built from four components: an ImageNet-pretrained ResNet-50 backbone that extracts spatial features from each frame; a Spatiotemporal Attention Module (TSAM) that applies both spatial and temporal attention across the resulting sequence of CNN features; a 1×1 convolutional adapter layer that reshapes channel dimensions before the Swin Transformer performs multi-scale global context modeling; and a fully connected classifier with dropout regularization. In its fully fine-tuned configuration, the architecture totals roughly 52 million parameters.

C. Training Strategy

Training follows a staged transfer-learning approach with layer-wise unfreezing across four progressive stages, using differentiated learning rates per submodule (1×10^{-5} for the CNN/Swin backbone, 1×10^{-4} for the TSAM/adapter, and 1×10^{-3} for the classifier). Gradient clipping is capped at a maximum norm of 1.0, and class imbalance is addressed through Focal Loss ($\alpha = 0.5$, $\gamma = 2.0$) [18]. Further regularization comes from weight decay (1×10^{-4}) and classifier dropout ($p = 0.5$), with early stopping triggered by validation macro F1-score after 10 epochs of no improvement within each unfreezing stage. The same validation macro F1-score also determined which epoch's checkpoint was retained; in the cross-validation runs no separate inner split was used for either purpose, so both decisions were made on the validation fold that is later reported (see Sections II.D and III.F). Learning rate scale, batch size, dropout, and weight decay were initially explored using Bayesian Optimization based on the Tree-structured Parzen Estimator, over 8 trials on a single stratified 80/20 hold-out split of the pooled 142 participants, with best validation macro F1 as the objective and shortened trials (up to 15 epochs, patience of 5). The cross-validation runs used the fixed configuration reported above (batch size 2 with two-step gradient accumulation, 16 frames per participant). Because the hold-out split for this search was drawn from the same pooled participants later used in cross-validation, the search was not independent of the validation folds. A more exhaustive search has been set aside for future work given the computational constraints of this stage.

D. Evaluation Scheme

Model evaluation used 5-fold stratified cross-validation performed at the participant level, which prevents a single participant's data from appearing in both the training and validation portions of any fold. Within each fold, however, early stopping and checkpoint selection relied on the same validation fold on which performance is reported, without a separate inner split, so the reported fold-level results are the best values reached on the evaluation fold and may be optimistically biased. Because the logistic regression and Random Forest baselines were run with fixed hyperparameters and involve no epoch-wise checkpoint selection, this bias, if anything, favours the deep model and therefore does not weaken the finding that it failed to outperform the linear baseline. The withheld test subset was not used in cross-validation, early stopping, or the preliminary hyperparameter search. Reproducibility was maintained through global seeding of weight initialization, sample ordering, and fold assignment. Beyond accuracy and macro F1-score, sensitivity and specificity were also reported, benchmarked against logistic regression and Random Forest baselines trained on statistical landmark features (mean, standard deviation, and range of motion of the normalized facial points).

E. Ablation Study Design

To isolate the individual contribution of the Spatiotemporal Attention Module (TSAM) and the Swin Transformer stage, four architectural variants were trained and evaluated under the identical 5-fold, participant-level stratified cross-validation protocol described in Section II.D: (1) *resnet_only*, a ResNet-50 backbone followed directly by the classifier head, with neither TSAM nor Swin; (2) *resnet_tsam*, ResNet-50 with TSAM but without Swin, where the temporally-attended CNN features are spatially pooled and averaged across frames before classification; (3) *resnet_swin*, ResNet-50 with the Swin Transformer stage but without TSAM; and (4) *full*, the complete architecture reported in the main results (Section III.A). All four variants share the same training configuration as the main cross-validation run described in Section II.C (Focal Loss with $\alpha = 0.5$, $\gamma = 2.0$, identical learning rates per submodule, gradient clipping, including the identical staged unfreezing schedule, batch-normalization freezing, per-participant frame sampling, and pseudo-image rendering used for the main results, so that the full variant reported here reproduces Table I). The same batch size, gradient accumulation, and full-precision training as the main run were used; a class-balanced sampler was used during training to counteract the 100:42 Healthy/Depressed imbalance. Because the ablation uses the same protocol as the main run, the full variant's figures reported here are directly comparable to Table I and should be interpreted only in relative terms across the four variants.

III. RESULTS AND DISCUSSION

A. Classification Performance Comparison

Table I compares classification performance across the hybrid "Beyond Expression" model and the two baseline models, all assessed under the same 5-fold cross-validation scheme, with accuracy and macro F1-score reported as mean $\pm$ standard deviation across the five folds. Because both baselines rely solely on statistical landmark features without any deep-learning component, they serve as a reference point for what is achievable without learning image-based representations. The hybrid model reached an accuracy of $66.95\% \pm 5.49\%$ and a macro F1-score of 0.5102 ± 0.0969 , falling short of the 90% target set for this proof-of-concept phase. More importantly, its performance stayed close to the simple logistic regression baseline (F1 = 0.5091), while Random Forest (F1 = 0.4032) did not even outperform a majority-class guessing strategy. A classifier that always predicts the dominant Healthy class would reach 70.42% accuracy (100/142) purely by chance, which is higher than the 67.64% accuracy Random Forest actually achieved. This confirms that Random Forest's seemingly decent accuracy is not just a symptom of class imbalance bias; the model performs below even a naive majority-class guess, which underscores its lack of real discriminative power.

To formally test this apparent lack of advantage rather than relying on visual overlap of means and standard deviations, a paired Wilcoxon signed-rank test and a paired t-test were conducted on the five per-fold macro F1 scores in Table II against the logistic regression baseline's per-fold scores. Two of the five folds favored Beyond Expression and three favored the baseline, with a mean paired difference of only 0.0012 macro F1. The Wilcoxon signed-rank test returned $p = 0.8125$ and the paired t-test returned $p = 0.9822$; a bootstrap 95% confidence interval on the mean difference was $[-0.0634, 0.0995]$, which includes zero. All three results are consistent with no statistically detectable difference between Beyond Expression and the logistic regression baseline at this sample size. It should be noted that with only five paired folds, the Wilcoxon signed-rank test cannot reach the conventional $\alpha = 0.05$ threshold regardless of effect size, since its smallest achievable two-tailed p-value at $n = 5$ is 0.0625; these tests should therefore be read as confirming the absence of a detectable advantage rather than as strong evidence of true equivalence, and a larger, more stable validation set, as planned for later stages of the research, would be needed to draw a more conclusive comparison.

TABLE I
ACCURACY AND MACRO F1-SCORE COMPARISON (5-FOLD CV)

Model	Accuracy	Macro F1
Logistic Regression (baseline)	0.6057 ± 0.0455	0.5091 ± 0.0682
Random Forest (baseline)	0.6764 ± 0.0426	0.4032 ± 0.0153
Beyond Expression (CNN-Swin-TSAM)	0.6695 ± 0.0549	0.5102 ± 0.0969

Table II expands the averages from Table I into per-fold figures, aiming to surface the variability that an aggregate average tends to hide. The spread between folds turns out to be substantial: accuracy ranges from 58.62% (fold 1) up to 71.43% (folds 4 and 5), and macro F1-score spans from 0.4082 (fold 2) to 0.6190 (fold 4), more than a 12 percentage point swing in accuracy and nearly 0.21 points in macro F1. This variability is consistent with the small per-fold validation size (28 to 29 participants), and it suggests that the headline numbers in Table I represent an average over fairly uneven fold conditions rather than uniformly stable performance.

TABLE II
ACCURACY AND MACRO F1-SCORE PER FOLD FOR THE BEYOND
EXPRESSION MODEL

Fold	Accuracy	Macro F1
1	0.5862	0.5167
2	0.6897	0.4082
3	0.6429	0.5906
4	0.7143	0.6190
5	0.7143	0.4167
Average	0.6695 ± 0.0549	0.5102 ± 0.0969

Positioned against prior work, the $66.95\% \pm 5.49\%$ accuracy reported here falls toward the lower end of the roughly 60% to 90% range reported for hybrid CNN-Vision Transformer and related deep-learning approaches to facial-expression-based mental-disorder detection [9], [10], [11]. This gap is broadly consistent with differences in dataset scale and modality: several of those studies draw on larger or more heterogeneous corpora, or combine facial cues with additional modalities such as audio or text, whereas the present study relies solely on 3D facial landmark sequences from a single, comparatively small cohort (142 participants). This comparison should be read cautiously, since accuracy alone does not capture the clinically relevant sensitivity gap documented in Section III.B, and none of the cited studies report sensitivity and specificity at the level of detail attempted here; a more direct, metric-matched comparison is left for future work once a larger dataset is available.

B. Clinical Analysis: Sensitivity and Specificity

Table III turns to metrics of greater clinical relevance than accuracy or macro F1, namely sensitivity (recall for the Depressed class, that is, the share of actual Depressed cases correctly detected) and specificity (recall for the Healthy class). Average sensitivity landed at just $24.17\% \pm 20.48\%$, far below the 90% clinical target, while specificity reached $85.00\% \pm 13.42\%$. This wide gap indicates that the model is considerably more dependable at recognizing Healthy participants than Depressed ones. The variability between folds is striking: two of the five folds (2 and 5) recorded a sensitivity of exactly zero, meaning the model missed every single Depressed case in those folds, while fold 3 achieved the best result at 0.5000. This suggests that detecting the minority class is highly sensitive to which participants happen to fall into each fold, a predictable consequence of training on a

minority class as small as 42 Depressed participants. The pattern lines up with the per-fold macro F1 variability seen in Table II, where the two folds with zero sensitivity also post relatively weak macro F1 scores.

TABLE III
SENSITIVITY AND SPECIFICITY PER FOLD

Fold	Sensitivity	Specificity	N (val)
1	0.3333	0.7000	29
2	0.0000	1.0000	29
3	0.5000	0.7000	28
4	0.3750	0.8500	28
5	0.0000	1.0000	28
Average	0.2417 ± 0.2048	0.8500 ± 0.1342	142

It should be noted that the standard deviations reported in Tables III and VI use a different divisor convention than those in Tables I and II: the per-fold statistics in Tables I and II use the sample standard deviation (divisor $n-1$), whereas the macro-averaged sensitivity and specificity in Table III, and all four metrics in Table VI, use the population standard deviation (divisor n) across the five folds. This is a convention difference rather than an inconsistency or computational error.

Table IV shows the combined confusion matrix built from predictions pooled across all five folds. Of 100 Healthy participants, 85 were correctly identified and 15 were misclassified as Depressed (false positives); of 42 Depressed participants, only 10 were correctly caught (true positives), while 32 were missed and classified as Healthy (false negatives). These 32 false negatives translate into a pooled sensitivity of 23.81% (10/42), which lines up with the macro-averaged sensitivity of $24.17\% \pm 20.48\%$ reported in Table III, together confirming that the large majority of genuine Depressed cases went undetected, a serious clinical concern in the context of early screening. It is worth clarifying that the sensitivity in Table III is a macro-average of per-fold values, whereas the figure implied by the pooled confusion matrix in Table IV comes from aggregated counts across all folds. The two figures are close but not identical because they reflect different aggregation methods, mean-of-ratios versus ratio-of-sums, and both are presented here for completeness.

TABLE IV
COMBINED CONFUSION MATRIX ACROSS ALL FOLDS

	Predicted Healthy	Predicted Depressed
Actual Healthy	85	15
Actual Depressed	32	10

C. Decision Threshold Adjustment Analysis

Table V reports what happens when the classification threshold on the model's Depressed-class probabilities is adjusted post hoc, without any retraining, pooled across all 142 participants from the five folds. At the default threshold of 0.50, sensitivity sat at only 23.81% alongside a high specificity of 85.00% and a macro F1 of 0.5410. Lowering the threshold to 0.40 pushed sensitivity up sharply to 71.43%, but

at the cost of specificity, which fell to 28.00%, and macro F1 actually declined to 0.4083, an unfavourable trade-off overall. At 0.35, sensitivity edged up further to 73.81% while specificity dropped again to 22.00%. At and below a threshold of 0.30, the model collapsed entirely into predicting Depressed for all 142 samples (sensitivity 100%, specificity 0%), effectively becoming a single-class guesser with no remaining discriminative value; this collapse was confirmed to be a stable plateau rather than a single threshold artifact, as identical results (sensitivity 100%, specificity 0%, macro F1 0.2283) were observed at every threshold tested down to 0.10, shown in Table V. Notably, every one of the 142 participants received a Depressed-class probability of at least 0.30, and 117 of them fell between 0.30 and 0.50, indicating that the predicted probabilities are compressed into a narrow band and separate the two classes only weakly. While no threshold reached the 90% sensitivity target without severely sacrificing specificity, these results suggest threshold tuning could still serve as a useful complementary strategy in clinical screening settings where minimizing false negatives is the priority. The probabilities in Table V are pooled from five independently trained fold models whose checkpoints were selected on the same folds, and no threshold was selected as an operating point; any operating threshold for screening use would need to be chosen on separate validation data and confirmed on unseen participants. It is also worth noting that the macro F1 of 0.5410 at the default threshold in Table V is computed by pooling predictions across all folds before scoring (a ratio-of-sums calculation), whereas the 0.5102 ± 0.0969 reported in Table I is the mean of per-fold macro F1 values (a mean-of-ratios calculation); the two are close but not identical for the same reason already noted for sensitivity in Section III.B, and both are reported here for completeness rather than as competing estimates.

TABLE V
DECISION THRESHOLD ANALYSIS (POOLED ACROSS ALL FOLDS, N=142)

Threshold	Sensitivity	Specificity	Macro F1
0.50	0.2381	0.8500	0.5410
0.40	0.7143	0.2800	0.4083
0.35	0.7381	0.2200	0.3707
0.30	1.0000	0.0000	0.2283
0.25	1.0000	0.0000	0.2283
0.20	1.0000	0.0000	0.2283
0.15	1.0000	0.0000	0.2283
0.10	1.0000	0.0000	0.2283

D. Model Interpretability (Grad-CAM)

Grad-CAM was applied to the final spatial layer of the CNN backbone to produce activation maps overlaid on the pseudo-image landmark coordinates. Each figure (Figs. 1 to 4) shows three panels per frame: the original landmark representation, the Grad-CAM heatmap (with red/yellow marking high activation and dark blue marking low), and an overlay combining the two, arranged as a 2×2 grid across frames 0, 5, 10, and 15. To check whether the observed activation was a stable pattern or simply a coincidence tied to

one frame, this analysis was carried out across two Healthy samples (P303, P302) and two Depressed samples (P335, P319).

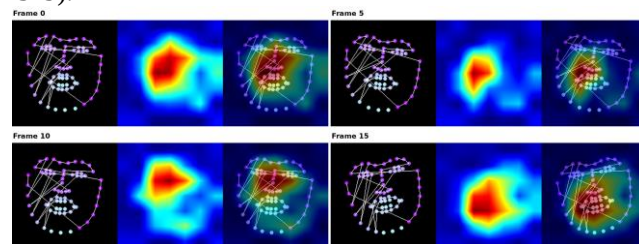


Fig. 1. Grad-CAM sequence for Healthy sample (P303), frames 0/5/10/15: [original landmarks | heatmap | overlay]. Activation is consistently formed in every frame despite location shifts.

For Healthy sample P303 (Fig. 1), activation forms reliably and stays reasonably well-defined across all four frames. Its center does shift, moving from the mid-left side of the face early on to an area overlapping the perioral landmark cluster mid-sequence, but the heatmap never flattens out entirely. A comparable pattern emerges for Healthy sample P302 (Fig. 2): elongated activation appears in the perioral region (around the mouth and nose) at frame 0, followed by a strong, blob-shaped signal at frame 5, and activation that widens but remains clearly visible through frames 10 and 15. This tendency for activation to persist across every observed frame in both samples, even as its location shifts, is suggestive of a pattern associated with the Healthy class in this model, consistent with the relatively high specificity of $85.00\% \pm 13.42\%$ reported in Table III; with only two Healthy and two Depressed participants examined in this qualitative analysis, however, this observation should be read as an illustrative pattern rather than a statistically established class signature.

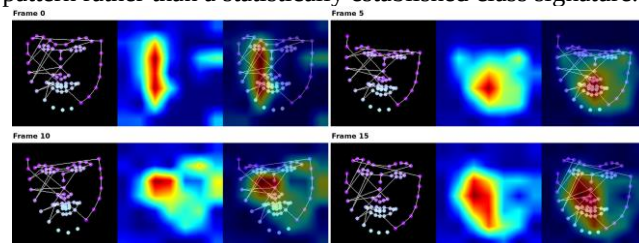


Fig. 2. Grad-CAM sequence for Healthy sample (P302), frames 0/5/10/15. Activation consistently appears in the perioral area and its surroundings across all frames.

Depressed sample P335 (Fig. 3), by contrast, shows a far more erratic pattern: two weak blobs in the lower face at frame 0, a single strong blob in the upper face at frame 5, a shift to a single strong blob lower down at frame 10, and a split into two unequal blobs at frame 15, so that no single location or shape repeats across the four frames. An even more pronounced instability shows up in Depressed sample P319 (Fig. 4): a focused activation blob appears in the upper face at frame 0, but frames 5 and 10, half of the observed sequence, collapse into a completely flat map with no discernible spatial focus at all, before activation resurfaces at a different (lower) location in frame 15.

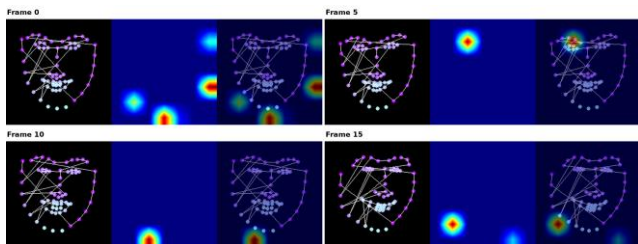


Fig. 3. Grad-CAM sequence for Depressed sample (P335), frames 0/5/10/15. Activation location and shape vary between frames with no stable spatial locus.

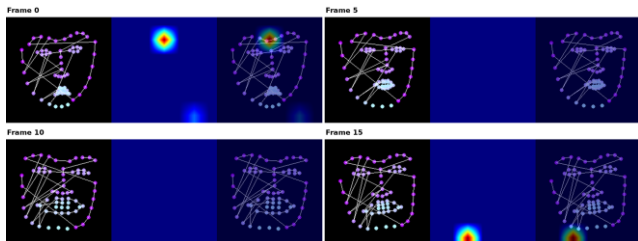


Fig. 4. Grad-CAM sequence for Depressed sample (P319), frames 0/5/10/15. Activation collapses completely (flat) in 2 of the 4 observed frames.

The activation collapse seen in half of P319's observed frames offers a direct visual explanation for the low sensitivity reported in Table III ($24.17\% \pm 20.48\%$) and for the large number of false negatives in the combined confusion matrix (32 of 42 Depressed cases missed, Table IV): when the model cannot locate a reliable spatial signal in some frames, its final decision tends to default toward the majority Healthy class. This pattern at the individual-frame level mirrors what happens at the fold level, where the two folds with zero sensitivity (folds 2 and 5) are also the ones with comparatively low macro F1 scores in Table II, indicating that the Grad-CAM instability observed frame by frame corresponds to instability in classification performance at both the participant and fold levels. Taken together, this contrast between classes suggests the model has not yet learned a stable, clinically meaningful spatial signature for the Depressed class, relying instead on cues that shift unpredictably over time, further evidence that its learned representation has not generalized well enough given the current data scale.

To move from qualitative description to quantitative validation, cross-frame consistency metrics were computed directly from the fold-4 checkpoint used to generate Figs. 1-4, at the same four frames (0, 5, 10, and 15) for all four illustrated participants. Pairwise pixel correlation between frames averaged 0.803 for P303 and 0.559 for P302 (Healthy), compared to 0.028 for P335 and -0.009 for P319 (Depressed), and mean centroid displacement (on a normalized 0-1 scale) was correspondingly lower for the Healthy samples (0.058 and 0.118) than for P335 (0.443) and P319 (0.304). For P319, two of the four analyzed frames (frames 5 and 10, activation standard deviation below 0.05) were flagged as collapsed, exactly the frames visually identified as flat in Fig. 4, yielding a collapse fraction of 0.50 against 0.00 for the other three participants; because this figure is a mean across all four

frames, including the two collapsed ones, it should be read alongside that collapse fraction rather than as a stand-alone indicator of localization instability. Pairwise Intersection-over-Union (IoU) on percentile-thresholded activation masks was also computed but did not track class membership as cleanly as correlation and centroid displacement, since percentile-based thresholding always selects a fixed proportion of pixels as active regardless of whether the underlying map carries a genuine signal; IoU can therefore be inflated for frame pairs that include a collapsed map, and is reported here only for completeness rather than as primary evidence. Cross-checking model predictions for these four participants further revealed that P335, although genuinely Depressed, was misclassified as Healthy by the fold-4 model (predicted probability for the Depressed class of only 0.418), offering a plausible explanation for why its Grad-CAM activation lacked a stable spatial signature. It should be noted that all four heatmaps in Figs. 1-4 are computed with respect to each participant's ground-truth class (the target logit backpropagated is Healthy for P303/P302 and Depressed for P335/P319, following standard practice); for P335, this means the Depressed-class logit being backpropagated is one the model itself assigned low confidence to (0.418), so part of the observed activation instability for P335 may reflect the mechanics of computing Grad-CAM from a low-confidence logit rather than purely representational instability, a distinction this analysis does not disentangle. Taken together, these metrics provide quantitative support for the pattern described above, while the IoU result illustrates a limitation of threshold-based overlap metrics under activation collapse.

E. Ablation Study

TABLE VI
ABLATION STUDY: CONTRIBUTION OF TSAM AND SWIN TRANSFORMER (5-FOLD CV)

Variant	Acc.	Macro F1	Sens.	Spec.
ResNet-50 only	0.6268 ± 0.0741	0.5338 ± 0.0347	0.3611 ± 0.1741	0.7400 ± 0.1715
ResNet-50 + TSAM	0.6345 ± 0.0925	0.5102 ± 0.0629	0.3306 ± 0.2546	0.7600 ± 0.2354
ResNet-50 + Swin	0.6759 ± 0.0735	0.4580 ± 0.0579	0.1500 ± 0.2424	0.9000 ± 0.2000
Full (ResNet-50 + TSAM + Swin)	0.6695 ± 0.0491	0.5102 ± 0.0866	0.2417 ± 0.2048	0.8500 ± 0.1342

Table VI isolates the contribution of each architectural component, all four variants trained under the identical protocol used for the main results. ResNet-50 alone achieved the highest macro F1 (0.5338 ± 0.0347) and the highest sensitivity ($36.11\% \pm 17.41\%$) among the four variants, ahead of ResNet-50 + TSAM ($F1 = 0.5102 \pm 0.0629$, sensitivity $33.06\% \pm 25.46\%$), the full model ($F1 = 0.5102 \pm 0.0866$, matching Table I's mean; sensitivity $24.17\% \pm 20.48\%$, matching Table III), and ResNet-50 + Swin, which recorded both the lowest macro F1 (0.4580 ± 0.0579) and the lowest

sensitivity ($15.00\% \pm 24.24\%$) of the four. Adding the Swin Transformer stage, whether alone or combined with TSAM, did not improve sensitivity over the ResNet-50 backbone by itself, and in the Swin-only variant reduced it by more than half. TSAM alone remained highly unstable across folds: it ranged from complete collapse toward the Healthy class in one fold (sensitivity 0%, specificity 100%) to a near-complete swing toward the Depressed class in another (sensitivity 77.78%, specificity 35%). The full model no longer shows the complete sensitivity collapse observed under the earlier simplified ablation protocol; two of its five folds still recorded zero sensitivity (consistent with Table III), but its overall ranking is comparable to ResNet-50 + TSAM rather than the worst of the four variants. Taken together, these results suggest that, at the current dataset scale, the Swin Transformer stage is not adding discriminative value for the Depressed class over the CNN backbone alone, and that TSAM's contribution, while real, is inconsistent across folds; simplifying the architecture or revisiting how Swin is integrated (for example, staged introduction after TSAM has stabilized, or evaluation at a larger data scale) may be a more productive direction than the current combination.

As a partial step toward the generalization evidence requested during review, the full model was additionally trained on the pooled 142-participant train+dev set, using 120 participants for fitting and a stratified internal slice of 22 participants solely for early stopping, with the same training protocol as the main cross-validation run, and evaluated once on DAIC-WOZ's held-out test split ($N = 47$ participants, entirely unseen during training, early stopping, or model selection), yielding an accuracy of 68.09%, a macro F1-score of 0.6390, a sensitivity of 57.14%, and a specificity of 72.73% (of 33 Healthy test participants, 24 were correctly identified and 9 misclassified as Depressed; of 14 Depressed participants, 8 were correctly identified and 6 missed). This sensitivity is higher than the mean sensitivity obtained across the five cross-validation folds in Table III (24.17%), though it comes from a single training run on a different train/validation split than any individual fold, so it is not directly comparable to a specific fold's result. Unlike the complete sensitivity collapse observed in two of the five cross-validation folds (Table III) and in the full variant's aggregate ablation figures reported here (Table VI), this single held-out run did not collapse entirely toward the majority class, though its sensitivity and specificity should still be read with caution given that it reflects one training run rather than a distribution of outcomes. It should be emphasized that this evaluation uses held-out participants from the same DAIC-WOZ corpus rather than an independent dataset, and therefore constitutes a same-dataset generalization check rather than genuine cross-dataset validation. Cross-dataset validation was not conducted at this stage due to the lack of current access to a compatible independent corpus (e.g., E-DAIC) with comparable facial landmark annotations; evaluation on such an independent corpus remains necessary before any claim of practical

screening capability can be supported, and is planned as a distinct milestone within the research roadmap described in the Limitations section.

F. Limitations

Six limitations stand out. First, the scale of the dataset (142 participants, only 42 labeled Depressed) constrains how well a model of this size (52 million parameters) can form a generalized representation, as reflected in its performance landing roughly on par with a simple linear baseline. Second, while the full model was evaluated once on DAIC-WOZ's held-out test subset (Section III.E), achieving a sensitivity higher than the cross-validation mean but from a single training run, this remains a same-dataset held-out-participant check rather than genuine cross-dataset validation. A fully independent corpus (e.g., E-DAIC or another depression-screening dataset with comparable facial landmark annotations) has not yet been evaluated, primarily because access to such a corpus was not available within the scope of this proof-of-concept stage; independent cross-dataset validation is planned as a discrete step preceding the clinical validation phase of the 2027 to 2029 research roadmap referenced in the following point. Third, expert validation against assessments from senior clinical psychologists, originally planned in the research proposal, has not yet taken place; together with the cross-dataset validation described above, it remains an agenda item for the 2027 to 2029 research roadmap, which sequences independent-corpus validation ahead of clinical expert validation. Fourth, the ablation study reported in Section III.E was conducted under the same protocol as the main results (Section II.C); while all four variants were compared fairly under this identical protocol, each variant, and the held-out check, reflects a single training run rather than repeated seeds, so the relative ranking of variants in Table VI and the specific sensitivity and specificity obtained in the held-out check could shift under a different random initialization or data split. Fifth, several lower-level implementation parameters (e.g., the exact PHQ-8 cutoff convention, per-participant frame count, landmark marker size, batch size, epoch count, optimizer, and Swin Transformer variant) are not exhaustively enumerated in the main text; these were held fixed across all reported comparisons and are available from the corresponding author upon reasonable request to support independent reproduction. Sixth, early stopping and checkpoint selection in the cross-validation, as well as the preliminary Bayesian hyperparameter search, were not separated from the validation data on which results are reported (Sections II.C and II.D): no inner split or nested cross-validation was used, and the hold-out split used in the preliminary search was drawn from the same pooled participants as the validation folds. The cross-validated figures, including those of the four ablation variants that share this protocol, should therefore be read as optimistic estimates, and an inner-split or nested protocol is planned for subsequent iterations. The single evaluation on the withheld test subset (Section III.E) is not

affected by this issue, because early stopping there used an internal slice of the pooled participants and the test participants played no role in training, early stopping, or model selection. Finally, it is important to distinguish what this study has demonstrated from what it has not: the model classifies statistical patterns in 3D facial landmark motion that correlate with PHQ-8-derived Healthy/Depressed labels within the DAIC-WOZ cohort, and the modest and fold-variable sensitivity observed across cross-validation, the four ablation variants, and the held-out check indicates this correlation is not yet reliable enough to be read as evidence of practical screening capability; the results reported here should accordingly be read as evidence of a learnable, but not yet clinically actionable, association between facial landmark dynamics and self-reported depressive symptom scores, not as evidence that the system performs clinical depression screening.

IV. CONCLUSION

This study successfully implemented and functionally validated a hybrid CNN–Swin Transformer–TSAM architecture for detecting depression-related facial landmark patterns, meeting the structural milestones expected at the proof-of-concept stage, while stopping short of a clinically validated screening capability. That said, classification performance ($66.95\% \pm 5.49\%$ accuracy, 24.17% sensitivity) fell short of the proposal's targets, a gap that shows up consistently across the per-fold breakdowns and is further illustrated through the cross-frame Grad-CAM analysis, which revealed spatial activation instability specific to the Depressed class; a held-out evaluation on the withheld test subset (Section III.E) produced a higher single-run sensitivity (57.14%) than the cross-validation mean, but as a single evaluation it does not establish that this gap has closed. These outcomes are in line with the study's standing as first-year basic research within a five-year roadmap, and they offer direct guidance for refining data acquisition, exploring augmentation strategies such as SMOTE, and optimizing the architecture in the next iteration, ultimately working toward a multimodal model and eventual prospective clinical trials.

ACKNOWLEDGMENT

This community service program is funded by the Universitas Pendidikan Indonesia through the community service program (contract No. 301/UN40.F7/AM.02.01/2026).

REFERENCES

- [1] Khubchandani et al., "The greatest global public health threats in the post-COVID-19 pandemic era," *Journal of Medicine, Surgery, and Public Health*, vol. 9, p. 100245, Aug. 2026.
- [2] O. Thornton and W. Li, "Global Trends and Correlations in Mental Health Disorders: A Comprehensive Analysis from 1990 to 2019," *Int. Neuropsychiatr. Dis. J.*, vol. 21, pp. 36–53, Mar. 2024, doi: 10.9734/INDJ/2024/v21i2425.
- [3] A. Z. Zhaifarah and S. P. P. Hastono, "Analisis Disparitas Geospasial dan Determinan Sosio-Demografis Masalah Kesehatan Jiwa di Indonesia: Analisis Data Survei Kesehatan Indonesia (SKI) 2023," *J. Biostat. Kependud. Inform. Kesehat.*, vol. 6, no. 2, p. 1, 2026, doi: 10.7454/bikfokes.v6i2.1135.
- [4] T. A. Bruckner, R. M. Scheffler, G. Shen, J. Yoon, D. Chisholm, J. Morris, B. D. Fulton, M. R. Dal Poz, and S. Saxena, "The mental health workforce gap in low- and middle-income countries: A needs-based approach," *Bull. World Health Organ.*, vol. 89, no. 3, pp. 184–194, Mar. 2011, doi: 10.2471/BLT.10.082784.
- [5] B. Levis, A. Benedetti, and B. D. Thombs, "Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis," *The BMJ*, vol. 365, p. 11476, Apr. 2019, doi: 10.1136/bmj.11476.
- [6] E. Sugimori and M. Yamaguchi, "Subthreshold depression is associated with altered facial expression and impression formation via subjective ratings and action unit analysis," *Sci. Rep.*, vol. 15, no. 1, p. 30761, Aug. 2025, doi: 10.1038/s41598-025-15874-0.
- [7] D. B. Olawade, O. Z. Wada, A. Odetayo, A. C. David-Olawade, F. Asaolu, and J. Eberhardt, "Enhancing mental health with Artificial Intelligence: Current trends and future prospects," *J. Med. Surg. Public Health*, vol. 3, p. 100099, Aug. 2024, doi: 10.1016/j.glmedi.2024.100099.
- [8] K. Shidara, H. Tanaka, H. Adachi, D. Kanayama, T. Kudo, and S. Nakamura, "Adapting the Number of Questions Based on Detected Psychological Distress for Cognitive Behavioral Therapy With an Embodied Conversational Agent: Comparative Study," *JMIR Form. Res.*, vol. 8, no. 1, p. e50056, Mar. 2024, doi: 10.2196/50056.
- [9] J. Aina, O. Akinniyi, MD. M. Rahman, V. Odero-Marrah, and F. Khalifa, "A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition," *IEEE Access Pract. Innov. Open Solut.*, vol. 12, pp. 91410–91425, 2024, doi: 10.1109/access.2024.3421376.
- [10] M. Mujiyanto, A. Setyanto, K. Kusri, and E. Utami, "Swin Transformer with Enhanced Dropout and Layer-wise Unfreezing for Facial Expression Recognition in Mental Health Detection," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 6, pp. 19016–19023, Dec. 2024, doi: 10.48084/etasr.9139.
- [11] M. Narigina, A. Romanovs, and Y. Merkuryev, "Convolutional neural network-based digital diagnostic tool for the identification of psychosomatic illnesses," *Algorithms*, vol. 17, no. 8, art. no. 329, Jul. 2024, doi: 10.3390/a17080329.
- [12] Y. Liu et al., "Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli," *J. Affect. Disord.*, vol. 381, pp. 44–54, Jul. 2025, doi: 10.1016/j.jad.2025.03.155.
- [13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [14] J. Gratch et al., "The Distress Analysis Interview Corpus of human and computer interviews," presented at the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 2014, pp. 3123–3128. Accessed: Sep. 04, 2026. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf
- [15] D. DeVault et al., "SimSensei kiosk: a virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, in AAMAS '14. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, May 2014, pp. 1061–1068. Accessed: Sep. 04, 2026. [Online]. Available: <https://dl.acm.org/doi/10.5555/2615731.2617415>
- [16] A. Jallaglag, M. A. Sabri, A. Yahyaouy, and A. Aarab, "Lightweight Depression Detection Using 3D Facial Landmark Pseudo-Images and CNN-LSTM on DAIC-WOZ and E-DAIC," *BioMedInformatics*, vol. 6, no. 1, p. 8, Feb. 2026, doi: 10.3390/biomedinformatics6010008.
- [17] J. Jung, C. Kang, J. Yoon, S. Kim, and J. Han, "HiQuE: Hierarchical Question Embedding Network for Multimodal Depression Detection,"

- in Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Oct. 2024, pp. 1049–1059. doi: 10.1145/3627673.3679797.
- [18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” in 2017 IEEE International Conference on Computer Vision (ICCV), Oct. 2017, pp. 2999–3007. doi: 10.1109/ICCV.2017.324.