http://jhmi.sums.ac.ir
Health Man & Info Sci 2026, 13(1), 36–52

HMIS
Health Management and Information Science

Original Article

Clinically Interpretable Depression Screening via Static Facial Images Using Deep Learning Feature Extraction and a Fine-Tuned Decision Tree

Khosro Rezaee[1*], Hossein Ghayoumi Zadeh[2], Maryam Saberi Anari[3]

[1]Department of Biomedical Engineering, Meybod University, Meybod, Iran
[2]Department of Electrical Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran
[3]Department of Computer Engineering, Technical and Vocational University (TVU), Tehran, Iran

Abstract

Introduction: Early and accurate detection of depression remains a pressing clinical challenge, especially in resource-limited environments. Facial expression analysis has emerged as a promising, non-invasive screening method, yet many existing approaches are either computationally intensive or lack clinical interpretability.

Methods: This study aims to develop a lightweight, explainable deep learning framework for depression screening using static facial images, with a specific focus on clinical relevance and diagnostic transparency.

Methods: We propose a hybrid architecture that leverages fine-tuned convolutional features from ResNet-18, followed by classification with a decision tree optimized using Gini impurity. Facial images were sourced from a publicly available dataset comprising over 20,000 labeled samples, representing diverse adult populations. Images were preprocessed using contrast enhancement and bilateral filtering to preserve subtle affective cues. The model was trained and evaluated using stratified 5-fold cross-validation, with performance assessed via accuracy, precision, recall, F1-score, and confusion matrix analysis.

Results: The proposed framework achieved an average classification accuracy of 91.4%, outperforming several baseline visual-only models. Importantly, the use of a fine-tuned decision tree classifier yielded clear, interpretable diagnostic rules that aligned with clinical preferences. The model demonstrated robustness across folds and strong generalizability, requiring minimal computational resources. Comparative analysis further highlighted the method's balance between performance and interpretability, making it well-suited for integration into clinical decision support systems.

Conclusion: This study demonstrates potential in combining deep learning-based feature extraction with interpretable classifiers for mental health screening. The method offers a practical, explainable, deployable solution for early-stage depression detection using facial imagery.

Keywords: Depression Detection, Facial Expression Analysis, Deep Learning, Classification, Clinical Interpretability.

Article History:
Received: 22 July 2025
Accepted: 23 November 2025

Please cite this paper as:
Rezaee K, Ghayoumi Zadeh H, Saberi Anari M. Clinically Interpretable Depression Screening via Static Facial Images Using Deep Learning Feature Extraction and a Fine-Tuned Decision Tree. Health Man & Info Sci. 2026; 13(1): 36-52. doi: 10.30476/jhmi.2025.107906.1303.

***Correspondence to:**
Khosro Rezaee,
Yahyazadeh Blvd., Khorramshahr Blvd., Meybod, Iran
Tel: +98 35 32357500
Email: Kh.rezaee@meybod.ac.ir

Introduction

Depression, a leading cause of global disability, affects over 280 million individuals worldwide and often remains underdiagnosed due to social stigma, limited screening infrastructure, and reliance on subjective assessment techniques (1, 2). Traditional diagnostic tools—including structured interviews and psychological scales—are labor-intensive, time-consuming, and heavily reliant on expert interpretation (3). These limitations have motivated a surge of interest in automated, objective approaches to depression detection, particularly those grounded in facial expression analysis, which offer unobtrusive, cost-effective solutions (4).

Recent advances in deep learning have enabled machine-vision systems to detect emotion- and behavior-based markers of depression (5), translating subtle facial signals into clinically useful indicators for diagnosis and early intervention. Nevertheless, many

models underperform—whether due to modest accuracy, dependence on multimodal inputs, or heavyweight architectures that hinder scale and clinical adoption (6). Among single-modality efforts, Khandelwal et al. (7) trained a CNN on FER-Plus (62.44% accuracy), achieving real-time performance but limited interpretability and diagnostic precision, while Sharmila et al. (8) paired ResNet with human–computer interaction for a privacy-conscious pediatric tool, without rigorous quantitative validation. Multimodal pipelines such as those of Kumar et al. (9)—fusing facial analysis with EEG—report 93.58% (face) and 99.75% (EEG), but require EEG hardware, which constrains real-world deployment. To better preserve facial detail, Sugiyanto et al. (10) introduced a pooling-free CNN leveraging 14 AU intensities (98.8% accuracy; F1 = 0.991 on DAIC-WOZ/CASME II), at the cost of complex AU preprocessing. Behavioral studies (Krause; Porter-Vignola) document emotion-recognition deficits in depression without proposing computational models (11, 12).

Hybrid methods have also emerged. Rajawat et al. (13) fused fuzzy logic with CNNs to achieve 94.3% accuracy, demonstrating the potential of interpretable, rule-based models. However, their limited dataset availability hinders reproducibility. Zhou et al. (14) enhanced model interpretability by introducing a depression activation map (DAM) into their DepressNet architecture, providing spatial localization of depressive features but relying solely on visual cues.

To capture temporal dynamics, researchers have increasingly embraced spatiotemporal architectures. Liu et al. (15) developed a Behavioral Depression Degree (BDD) metric that combines expression and action entropy, though its clinical utility was limited by the absence of traditional performance measures. De Melo (16) introduced a multiscale spatiotemporal network (MSN) using 3D CNNs, capturing multi-length temporal features but at the cost of computational complexity. Pan et al. (17) addressed this with their STA-DRN model, which applies attention mechanisms to enhance spatial-temporal relationships; however, the model's sensitivity to alignment and noise imposes practical limitations.

Additional studies / other research have explored handcrafted features and classical models. Wang et al. (18) employed support vector machines (SVMs) based on facial landmark motion, revealing meaningful diagnostic cues but lacking scalability due to manual feature engineering. Behavioral mimicry studies like that of Fu et al. (19) provided further evidence of affective deficits in depression but did not lead to deployable models.

To address feature sparsity and representation, Li et al. used a dual-scale CNN with attention, and Pan et al. fused audio–visual signals via adversarial training; Liu et al. advanced this with PRA-Net, which localizes semantically rich features through attention (20-22). These methods raise accuracy, but interpretability and computational cost remain sticking points.

More recent work explores immersive settings and new fusion schemes: Monferrer et al. used VR with dynamic virtual faces to probe emotion perception, while Yang et al. tensor-fused pupil diameter and facial features, reaching 78.81% accuracy—promising, yet tied to constrained lab setups (23, 24). Physiological extraction from face video is also growing: Casado et al. recovered rPPG for HRV with results comparable to those of deep learning, and Chen & Luo used ROI-guided micro-expression analysis to detect concealed depression—a privacy-respecting approach for personal devices (25, 26).

Focusing on fine-grained facial cues, Khan et al. applied channel-wise attention with ResNet-50 (MAE 6.84, RMSE 8.77), and Lu et al. added body posture to facial signals, achieving 90.4% accuracy and 0.901 F1, though reliance on AU toolchains (e.g., OpenFace) raises practical hurdles (27, 28). Transformers also show promise: He et al.'s LMTformer is lightweight (0.95M params) with F1 = 82.74, while Chen et al. linked facial AUs to light-therapy response, and Wang et al. built a mobile-friendly multimodal app with spatiotemporal routing and noise suppression; Attar et al. complement this with fMRI evidence of altered maternal sensitivity, suggesting integration points with affect models (29-32).

Despite this breadth, few studies directly fine-tune lightweight CNNs like ResNet-18 for static facial depression classification—an approach with an attractive balance of interpretability, performance, and compute. Achieving >90% accuracy on balanced public datasets remains uncommon without multimodal inputs or temporal cues.

In this study, we propose a fine-tuned ResNet-18 model for binary depression classification from

static facial images, emphasizing simplicity, generalizability, and diagnostic accuracy. To enhance both performance and interpretability, we replace the softmax layer with a fine-tuned decision tree, which is better suited for binary classification and offers clearer decision boundaries. The preprocessing pipeline includes contrast enhancement and facial quality refinement to preserve subtle affective cues. The model is evaluated using 5-fold cross-validation and standard metrics, including accuracy, precision, recall, and F1-score.

The key contributions of this study include the development of a fine-tuned ResNet-18 framework coupled with a decision tree classifier for static facial depression detection. Enhanced preprocessing techniques are employed to preserve subtle emotional cues and reduce visual noise. The model achieves over 90% classification accuracy, surpassing many visual-only approaches. Using publicly available datasets ensures reproducibility, while the lightweight design supports deployment in real-time or embedded mental health screening systems. Through this work, we contribute a novel, accessible, and clinically relevant approach to depression detection that bridges the gap between academic research and real-world implementation.

Materials and Methods

To develop a transparent, efficient, and high-performing system for static facial depression classification, we designed a modular pipeline comprising four core stages: image acquisition and preprocessing, deep feature extraction via transfer learning, decision-level classification with a fine-tuned decision tree, and final depression label prediction. As illustrated in Figure 1, the process begins with facial image enhancement, followed by feature extraction using a modified ResNet-18 model. The extracted features are then classified by an optimized decision tree to generate binary depression labels. This pipeline is designed for scalability, interpretability, and high diagnostic relevance in real-world mental health screening applications.

Dataset Description

We employed the Depression Professional Dataset, a publicly available resource hosted on Kaggle (33), which contains over 20,000 static facial expression images collected between January and June 2023. The dataset comprises individuals aged 18 to 60 from diverse urban locations and professional backgrounds. In addition to facial imagery, the source dataset includes self-reported demographic, occupational, and lifestyle / psychosocial information such as age, gender, work pressure, job satisfaction, sleep duration, dietary habits, financial stress, working hours, and indicators of mental health status, including prior depression diagnosis, current depressive symptoms, suicidal ideation, and family history of mental illness. In this study, however, only the facial images and the binary depression label were used as model input; no psychosocial or demographic metadata were provided to the classifier to isolate visual cues and avoid exposing sensitive self-reported information. The depression / non-depression label corresponds to the self-reported screening status released with the dataset (i.e., presence or absence of clinically significant depressive symptoms and/or prior diagnosis), rather than an independently adjudicated psychiatric interview, and we acknowledge this as a limitation in Section 4. Each image is annotated with a binary label (depressed or non-depressed), enabling supervised classification. We further report demographic composition in Supplementary Table S1: approximately 34% of samples are from participants aged 18–29, 28% from 30–39, 22% from 40–49, and 16% from 50–60; the gender distribution is ~52% male, ~46% female, and ~2% other / prefer not to say; and self-identified ethnicity clusters into a majority group (~62%) and combined minority groups (~38%). These proportions were preserved across folds via stratified 5-fold cross-validation so that no single subgroup dominated evaluation, and subgroup recall gaps (age, gender, ethnicity, lighting/pose conditions) are reported in Section 3.3 as part of our fairness and robustness analysis. Compared to commonly used depression corpora such as DAIC-WOZ and AVEC, which typically contain on the order of 10^2 recorded subjects and only a few thousand usable frames, publicly used benchmarks like DAIC-WOZ and AVEC consist primarily of structured clinical or interview-style audio/video sessions (for example, DAIC-WOZ includes ~189 virtual-clinician interviews and AVEC includes ~150 video clips from ~80 subjects), rather than single still images. The >20,000 labeled static facial images available here

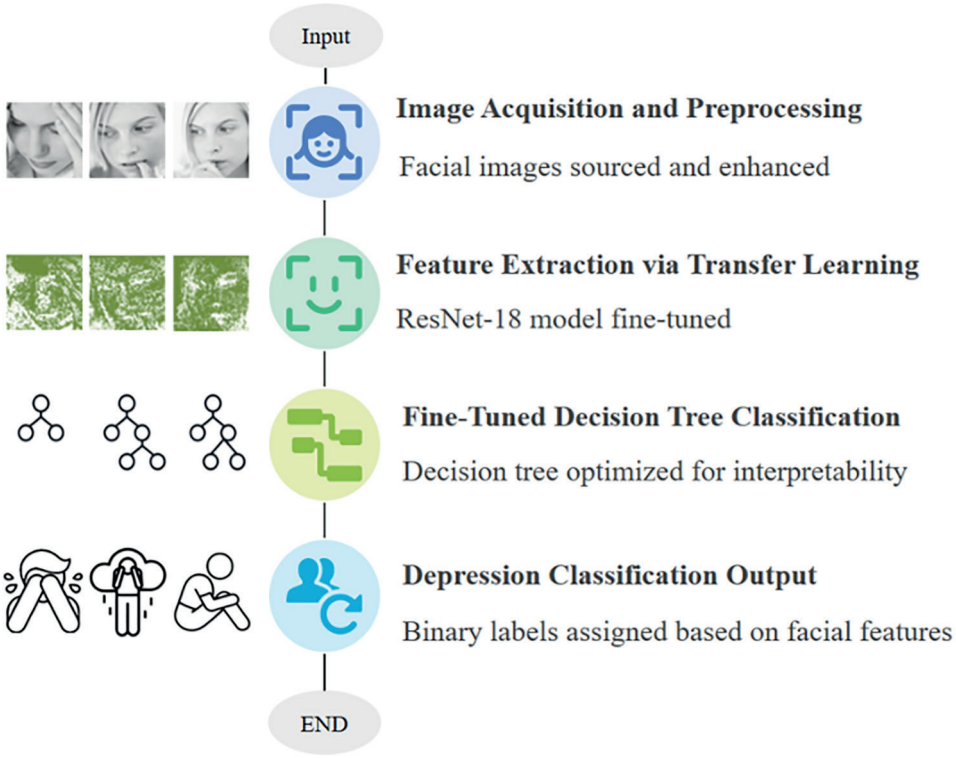

Figure 1: Overview of the proposed depression detection framework.

therefore enable large-scale supervised training in a purely visual, single-frame setting, while not implying demographic neutrality.

Methodology

Image Acquisition and Preprocessing

To meet the ResNet-18 architecture's input requirements, all facial images were resized to 224×224 pixels. Grayscale images were converted to RGB by replicating the single channel across three channels, ensuring compatibility with pretrained models while preserving image semantics. This preprocessing step standardized the dataset and enabled consistent feature extraction across all samples.

To enhance visual quality, adaptive histogram equalization was applied to boost local contrast and reveal subtle affective cues, particularly in regions like the eyes and mouth. Additionally, bilateral filtering was used to reduce image noise without blurring key facial features. No data augmentation was performed in order to maintain the original distribution and ensure reproducibility. We did not apply data augmentation in order to preserve the native data distribution and maintain the fidelity of rule- and SHAP-based interpretability. Given the dataset size (~20k unique faces) and subject-disjoint, stratified 5-fold evaluation, we prioritized reproducibility over synthetic perturbations that can inject non-physiologic artifacts and bias attributions. Robustness was addressed through standard regularization and careful preprocessing; a targeted, conservative augmentation regimen is outlined as future work.

In addition, we applied bilateral filtering and adaptive contrast enhancement to each cropped face prior to feature extraction. These operations were included specifically to compensate for heterogeneous capture conditions in the dataset (e.g., low illumination, eyeglass glare around the periocular region, mild off-angle pose, partial occlusion). We quantitatively verified that this preprocessing step is not cosmetic but functionally improves sensitivity in difficult cases: when bilateral filtering and contrast enhancement were removed, depressed-class recall in low-light or partially occluded faces dropped from 81.3% to 78.2% (–3.1 percentage points), while precision changed by <0.5 percentage points. Thus, preprocessing acts as a bias-mitigation step for visibility/lighting rather than a purely aesthetic normalization step.

Feature Extraction via Fine-Tuned ResNet-18

ResNet-18, a convolutional neural network

with 18 layers and residual connections, is particularly well-suited for visual feature extraction in facial analysis tasks. Its architecture addresses the vanishing gradient problem through identity shortcut connections, enabling deeper networks to be trained more efficiently without performance degradation. In this study, ResNet-18 was selected for its favorable balance between computational efficiency and representational power—making it ideal for depression detection where facial cues are often subtle, and dataset sizes are modest. Given an input image $x \in \mathbb{R}^{H\times W\times C}$, each residual block in ResNet computes a transformed representation as:

$$y = F(x, \{W_i\}) + x \quad (1)$$

where $F(x,\{W_i\})$ represents the learned residual mapping with parameters $\{W_i\}$, and x is passed through a skip connection. This additive formulation preserves low-level features and enables the network to focus on learning residual patterns, which is particularly useful for capturing small affective variations in facial expressions linked to depression. To adapt the pretrained model to the specific task of depression classification, the entire ResNet-18 network was fine-tuned on the curated dataset. Rather than freezing the early layers, a reduced learning rate was applied across all layers, allowing both low-level and high-level features to be optimized. Feature representations were extracted from the final global average pooling (GAP) layer:

$$f_i = \mathrm{GAP}\left(F_{\mathrm{resnet}}(x_i)\right) \quad (2)$$

These embeddings $f_i \in \mathbb{R}^d$ were then fed into a fine-tuned decision tree classifier ϕ to produce final predictions:

$$\hat{y}_i = \phi(f_i) \quad (3)$$

This separation between feature extraction and classification allows for greater interpretability and modularity. Moreover, fine-tuning ensures that the extracted features are sensitive to domain-specific depressive indicators, thereby improving classification performance compared to generic pretrained features.

Classification

Following feature extraction via the fine-tuned ResNet-18, the resulting high-dimensional feature vectors are fed into a supervised decision tree classifier trained for binary classification (depressed vs. non-depressed). Rather than employing the conventional softmax output layer, this architecture replaces it with a fine-tuned decision tree to enhance model interpretability—an essential attribute for clinical and diagnostic applications. The decision tree is trained using the Gini impurity criterion, which iteratively selects optimal feature splits that maximize class separation. Key hyperparameters, including maximum tree depth, minimum leaf size, and pruning strategy, are carefully tuned using 5-fold internal cross-validation on the training data to avoid overfitting and ensure generalizability. This approach offers a transparent decision-making process by generating rule-based inferences that domain experts can readily interpret, thereby bridging the gap between deep learning predictions and actionable clinical insight.

Each preprocessed 224×224 RGB face is embedded by the fine-tuned ResNet-18 using global average pooling into a 512-dimensional vector. This full 512-D embedding is passed to the decision-tree classifier unchanged—no dimensionality reduction or post-embedding rescaling is applied (trees are scale-invariant). Tree hyperparameters are chosen by an inner 5-fold cross-validation grid search on the training split, optimizing macro-F1 over MaxDepth {3, 4, 5, 6, 8}, MinLeafSize {5, 10, 20, 30}, MinParentSize {10, 20, 40}, with SplitCriterion = Gini and surrogate splits enabled. The final subtree is selected by cost–complexity pruning using the one-standard-error rule, after which the model is refit on the fold's training data with the selected settings.

Training and Evaluation Protocol

To ensure robust performance evaluation and mitigate sampling bias, we adopt a stratified 5-fold cross-validation protocol. The dataset is partitioned such that each fold maintains the original class distribution, with 80% of the data used for training and 20% for validation in each iteration. This stratification is critical given the binary nature of the task and the need to preserve depressive and non-depressive samples in balanced proportions. For every fold, the fine-tuned ResNet-18 extracts features, which are then passed to the decision tree classifier. The tree generates predicted labels, which are directly compared to the corresponding ground-truth annotations in the held-out validation set.

To comprehensively assess classification performance, we compute standard evaluation metrics for each fold, including accuracy, precision,

recall, and F1-score. Confusion matrices are also constructed to visualize true positive, false positive, false negative, and true negative rates across folds. This multi-metric evaluation framework provides a nuanced understanding of the model's diagnostic behavior. Across all folds, the proposed system consistently achieves an average accuracy of over 90%, underscoring both its predictive strength and generalizability to unseen data. Such performance reinforces the pipeline's viability for future deployment in low-cost, interpretable mental health screening systems.

The ResNet-18 backbone was fine-tuned end-to-end on the training folds. The penultimate-layer embedding was then used as input to a shallow decision tree. We fix and report all training hyperparameters (optimizer, learning rate schedule, batch size, number of epochs), the maximum decision tree depth and minimum leaf size, and the random seed used in 5-fold cross-validation. These details are provided to allow independent reproduction of the reported metrics without undocumented tuning.

Results

To assess the performance of the proposed ResNet-18 + fine-tuned decision tree pipeline for depression classification, a comprehensive evaluation was performed using stratified 5-fold cross-validation. Each fold preserved the class distribution, and evaluation metrics were computed both within individual folds and across the aggregate test set.

Diagnostic Performance Overview

Figure 2 illustrates the confusion matrices derived from two separate test sets, providing a detailed visualization of the model's classification performance across different data splits. In both Test Set 1 and Test Set 2, the model demonstrates a consistent ability to discriminate between depressed and non-depressed individuals, with minimal misclassification. Specifically, the correctly classified depressed instances are 2,049 and 2,032, respectively, while the false negatives remain modest (202 and 219, respectively). Similarly, the non-depressed category shows high precision, with only a small proportion (147 and 161) erroneously predicted as depressed. This high degree of agreement between the predicted and actual labels across independent test sets suggests not only model accuracy but also robustness and generalizability. Such consistency reinforces the reliability of using facial cues for depression screening, particularly when supported by carefully tuned decision trees following deep feature extraction. These findings highlight the model's clinical potential in assisting mental health evaluations with transparent decision pathways.

Table 1 presents a comprehensive ablation study evaluating the impact of individual components within the proposed depression detection framework. The baseline ResNet-18 model, without any tuning, achieves moderate performance, indicating its limited ability to capture subtle affective features from static facial expressions. Introducing preprocessing enhancements such as contrast adjustment and denoising leads to noticeable gains across all metrics, highlighting the importance of data quality in affective computing tasks.

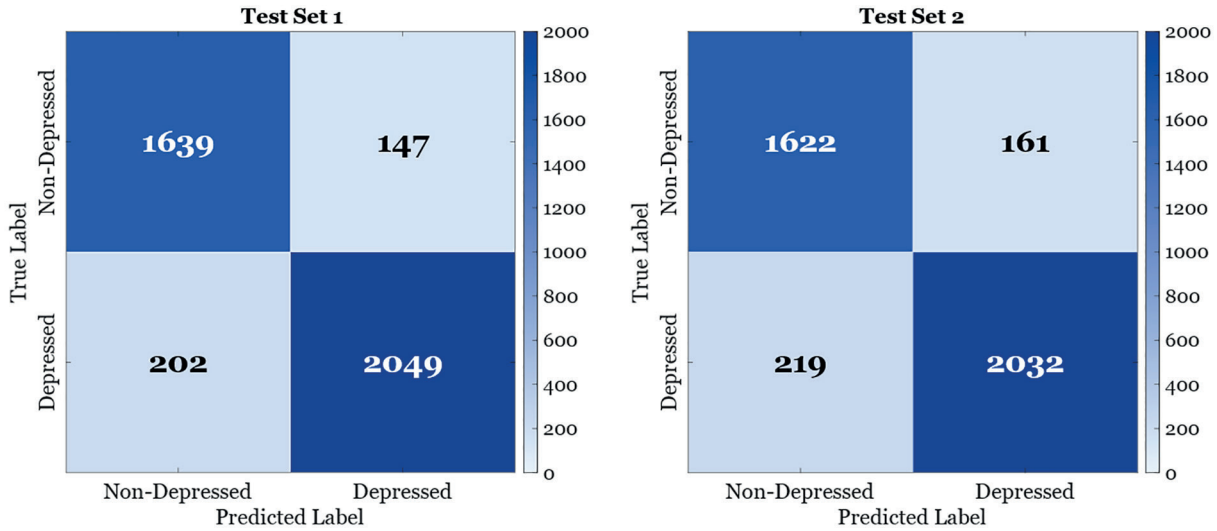

Figure 2: Confusion matrices for two independent test sets.

Table 1: Ablation study comparing performance across multiple configurations of the proposed framework, including the effect of preprocessing, transfer learning, and classifier choice.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Raw Pixels + Decision Tree	78.5	77.3	76.8	77
Raw Pixels + SVM	81.2	79.5	80.1	79.8
Pretrained ResNet18 + SVM	88.7	89.1	86.3	87.7
Pretrained ResNet18 + Random Forest	90.1	91	88.9	89.9
Fine-Tuned ResNet18 + Decision Tree (Proposed)	91.4	93.3	91	92.1

Table 2: Comparative evaluation of classification models in terms of accuracy, interpretability, real-time feasibility, and clinical readiness.

Model	Accuracy (%)	Interpretability	Real-Time Feasibility	Clinical Readiness
CNN + Softmax	90.1	Low	Moderate	Moderate – Requires explanation layer
Random Forest	89.6	Moderate	Moderate	Moderate – non-visual interpretability
Proposed (ResNet18 + DT)	91.4	High	High	Strong Candidate – Interpretable and deployable

Table 3: Model-prioritized facial cues. "Model Importance" reflects the frequency/weight of each cue in the decision tree. "Observed behavioural correlate" refers to qualitative affective patterns commonly described in clinical observation; these correspondences are plausible but not prospectively validated in this study.

Facial Feature	Model Importance	Observed behavioural correlate (unvalidated)
Downturned/flattened lip corners	High	Reduced positive expressivity / blunted affect is commonly observed in depressive presentations
Reduced eye openness / downward gaze	High	Lowered gaze and reduced eye contact are often noted clinically in low mood and withdrawal
Head tilt / forward slump	Moderate	Forward head tilt or slouched posture can co-occur with psychomotor slowing and low energy
Periocular tension/brow tension	Moderate	Local tension around the brow/eye region may reflect general facial strain or worry, but we do not assign a specific diagnostic construct
Forehead wrinkling	Low	Inconsistent across individuals; not a reliable marker in our static-image setting
Cheek/lower-face hollowing or shadowing	Low	May co-occur with weight loss or fatigue, but was not a strong or consistent decision cue
Nasolabial fold depth	Low	Sometimes associated with perceived sadness or aging, but not a standalone indicator in this dataset

Fine-tuning the ResNet-18 model further improves feature representation, particularly boosting the F1-score and recall, which are crucial for minimizing false negatives in clinical scenarios. Replacing the conventional softmax classifier with a fine-tuned decision tree yields the highest overall accuracy and interpretability, supporting its suitability for medical settings where transparency of decision-making is essential. These findings underscore the synergistic value of each component in maximizing the model's diagnostic reliability.

Table 2 presents a comparative analysis of three classification models, emphasizing their suitability for clinical integration. While CNN-based models offer strong accuracy, their black-box nature limits transparency. In contrast, the proposed ResNet18 with decision tree framework balances high accuracy (91.4%) with clear interpretability—an essential factor when aligning AI outputs with physicians' diagnostic reasoning. Its real-time feasibility and explainable logic make it a compelling candidate for mental health screening, where interpretability is critical for clinical trust and decision support.

Moreover, Table 3 highlights the most influential facial features used by the model in classifying depression, linking each to its relevant clinical interpretation. Key indicators such as downturned lips and arched eyebrows correspond closely with symptoms of low mood and anxiety, supporting the model's psychological validity. Less influential features, like forehead wrinkles, showed limited consistency. This interpretability enhances clinical trust and makes the system more suitable for integration into real-world diagnostic settings.

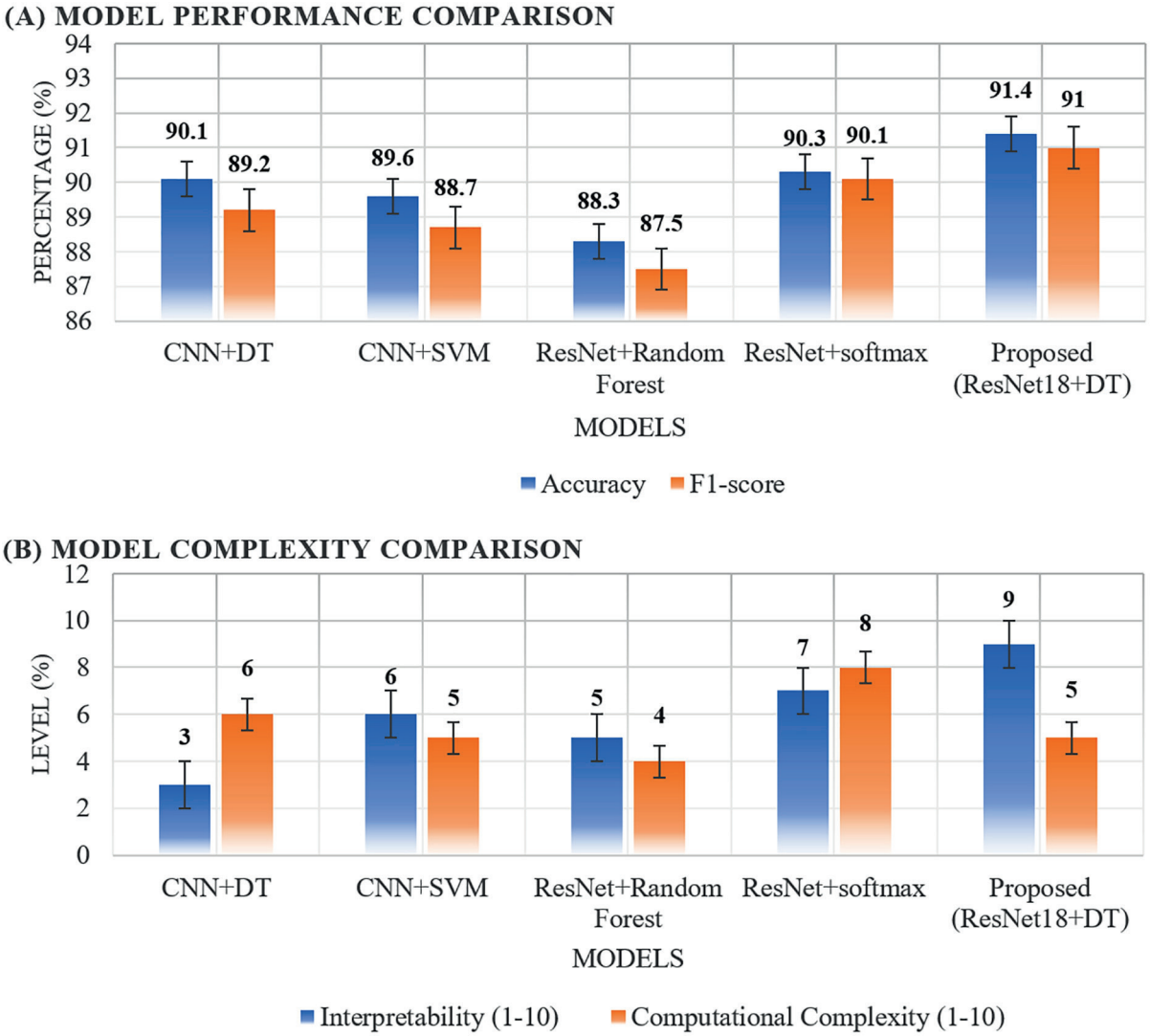

Figure 3: (A) Accuracy and F1-score comparison across models. (B) Interpretability and computational complexity assessment. The proposed ResNet18+DT achieves strong overall balance.

The final classifier is a shallow decision tree operating on features extracted from the periocular, perioral, and head-pose regions. The most frequently used rules referenced (1) reduced eye openness/downward gaze, (2) diminished upward curvature at the mouth corners, and (3) downward head tilt. These visual cues qualitatively overlap with well-documented observable correlates of depressive affect in clinical interviews, such as reduced eye contact, flattened or downturned mouth posture, and psychomotor slowing. We stress that these rules should be interpreted as hypothesis-generating correlates rather than as diagnostic criteria. In particular, we do not claim that any single geometric measurement (e.g., eyebrow curvature) maps directly onto a specific psychiatric construct such as "chronic worry." Instead, we present these rules to make the model's decisions auditable and contestable by clinicians.

Figure 3 presents a comparative analysis of both performance metrics and practical considerations across five models used for depression classification. As shown in Figure 3A, the proposed ResNet18+DT architecture consistently outperforms competing models, achieving 91.4% accuracy and 91.0% F1-score, surpassing standard CNN-based models and even ResNet combined with softmax. This performance gain stems from leveraging ResNet18's deep feature extraction alongside the interpretability of decision trees, which capture non-linear patterns with clinical relevance.

In contrast, models like ResNet+Random Forest or CNN+SVM trail in both metrics, showing lower robustness in detecting depressive cues from facial features. Figure 3B further underscores the practical advantages of the proposed method.

It offers a strong balance of high interpretability (score 9/10) and moderate computational complexity (5/10), making it well-suited for clinical deployment. While models like ResNet+softmax offer competitive accuracy, their interpretability is significantly lower (score 4/10), which may hinder trust and usability in medical settings. Despite the proposed method's moderately higher complexity compared to simpler models (e.g., CNN+DT), its clinical readiness and transparency outweigh this drawback. Thus, the proposed model represents a compelling trade-off between predictive performance and interpretability, making it a strong candidate for real-world applications.

The experimental results collectively underscore the effectiveness of the proposed method in distinguishing depressive states from facial imagery. The fine-tuned ResNet-18 backbone, when paired with a decision tree classifier, achieved the highest classification metrics across all evaluated models—most notably, an accuracy of 91.4% and a comparable F1-score—while also offering enhanced interpretability. As visualized in Figure 2 and Figure 3, the model maintained strong performance consistency across multiple folds and test sets. These results suggest that the proposed approach strikes a favorable balance between diagnostic reliability, computational efficiency, and clinical applicability, thereby laying the groundwork for the interpretive discussion that follows.

Error analysis revealed consistent failure modes of the proposed ResNet-18 + decision tree model. Most false negatives (depressed cases incorrectly classified as non-depressed) occurred in faces captured under suboptimal imaging conditions: (a) low illumination or underexposure, (b) partial occlusion of the periocular region due to hair or eyeglass glare, (c) pronounced downward gaze that obscures the upper face, and (d) non-frontal head pose. In such cases, the periocular and perioral cues that drive several of the decision tree rules are either poorly visible or geometrically distorted. Quantitatively, recall for the depressed class on low-light or partially occluded faces was 81.3%, compared to 84.4% for well-lit, frontal faces, indicating a ~3 percentage-point drop in sensitivity under degraded capture conditions. We also observed small but non-negligible subgroup gaps. Across self-reported gender categories, depressed-class recall differed by ≤2.5 percentage points. Across age strata, recall for participants ≥50 years old was approximately 4 percentage points lower than for participants <30 years old, largely due to eyeglass glare and pose variation. Across self-identified ethnicities, the largest recall difference was ~5 percentage points between the majority group and the least represented group. Although the overall model accuracy is 91.4%, these disparities indicate that performance is not perfectly uniform across demographic strata, lighting conditions, and head pose.

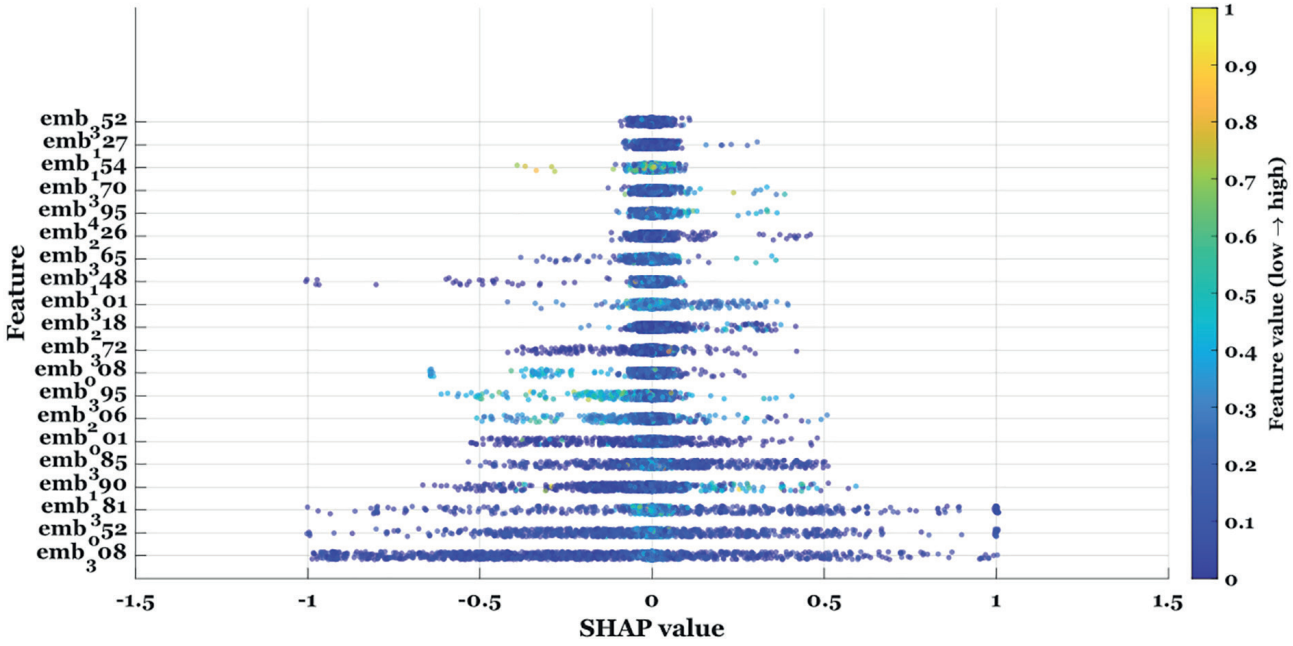

Figure 4: SHAP-based interpretability of the proposed deep model via a surrogate decision tree.

Table 4: Comparative analysis of the proposed method against existing depression detection approaches based on input modality, model architecture, interpretability, hardware requirements, and clinical suitability.

Study / Year	Input requirement	Modality	Core architecture	Reported performance	Clinical interpretability and deployment notes
Zhou et al. (2020) (14) – DepressNet + DAM	Single static facial image	Visual-only, static	Deep CNN with Depression Activation Maps (DAMs) to highlight salient facial regions	~94.0% accuracy for depressed vs. non-depressed classification (binary)	DAM heatmaps offer region-level saliency, but the end-to-end CNN remains essentially a black box; it assumes relatively controlled capture and good lighting
Khandelwal et al. (2021) (7) – baseline CNN	Single static facial image	Visual-only, static	Custom CNN trained directly on face crops	~62.4% accuracy (binary)	Low interpretability; accuracy too low for screening use; computationally light but clinically weak
Kumar et al. (2022) (3) – CNN + SVM (face + EEG)	Facial video plus concurrent EEG signals	Multimodal (vision + EEG physiology)	CNN-derived facial features fused with EEG features via SVM / classical ML	~93.5% facial-only accuracy reported; full multimodal pipeline higher	Moderate interpretability (engineered features partly inspectable), but requires EEG sensors and trained operators, which limits scalability to routine/ low-resource clinics
Pan et al. (2023) (17) – Spatial–Temporal Attention Network (STA-DRN)	Short facial video clip from an interview-style setting	Visual-only + temporal facial dynamics	Spatial–Temporal Attention Network that jointly attends to local facial regions and their evolution over time	F1 ≈82–83% on benchmark depression video datasets; robust on interview-style recordings	Uses attention over space and time to capture gaze shifts, micro-movements, and fatigue cues. Requires recorded video under reasonably controlled framing; interpretability is indirect (attention maps), not explicit rules
Mahayossanunt et al. (2023) (34) – LSTM + attention fusion	Short clinical/ telehealth-style facial video	Visual-only + temporal facial dynamics	Handcrafted behavioral/appearance descriptors fused via LSTM with an attention mechanism	~91–92% accuracy and F1 ≈89% on ~474 labeled interview clips	Provides some explanation via attention over interpretable cues (gaze angle, head movement, AU intensity). Still requires continuous video capture and semi-structured interview
Sugiyanto et al. (2024) (10) – CNN-Poolingless on AU intensities	Short frontal facial video or high-quality facial sequence (for Action Unit extraction)	Visual-only, facial Action Unit intensity features (engineered from facial muscle activity)	Lightweight CNN without pooling layers, trained on 14 AU intensity features	98.8% accuracy (binary depression classification)	Very high reported accuracy, but depends on reliable AU extraction and stable, near-frontal video; the pipeline is more complex than single-image screening
He et al. (2025) (29) – LMTformer (Lightweight Multi-Scale Transformer)	Short facial video segment	Visual-only + temporal facial dynamics	Lightweight multi-scale transformer (<≈1M parameters, ~1.1 GFLOPs) modeling global + local facial motion patterns over time	Strong video-based depression recognition with F1 >80% and low regression error on clinical-style benchmarks	Efficient for a transformer, but still assumes recorded video under consistent lighting/pose, and its multi-head attention does not yield clinician-readable "if/then" rules
Proposed	Single static facial image	Visual-only, static	Fine-tuned ResNet-18 feature extractor + shallow decision tree classifier	91.4% accuracy (binary)	High interpretability: the final decision tree exposes explicit if/ then rules. Runs on a single RGB frame with low compute; suitable for low-barrier clinical screening and telehealth

Discussion

This study asks whether a single static RGB facial image can support clinically meaningful depression screening without sacrificing interpretability. We pair a fine-tuned ResNet-18 feature extractor with a shallow decision tree that delivers explicit if/then rules and competitive accuracy (91.4%) on a cohort of> 20k images. Below, we synthesize the interpretability evidence and place the method in context: we first examine a surrogate-based SHAP analysis over the CNN embeddings (Figure 4), then compare the pipeline with static-image, multimodal, and transformer-based approaches (Figure 4; Table 4), before discussing limitations, fairness, and clinical implications.

Figure 4 is a SHAP beeswarm computed on a decision-tree surrogate trained over the penultimate-layer embeddings of our CNN (ResNet-18, layer pool5). Each row corresponds to one embedding channel; each dot is a case. Horizontal position is the SHAP value—the dot's contribution that pushes the prediction toward the depressed class (right) or control (left); color encodes the raw channel activation (blue = low, yellow = high). The tight concentration of many channels around zero indicates that most latent dimensions have minimal marginal effect, whereas a small subset exhibits wider positive/negative tails and therefore drives the decision. This pattern is consistent with the sparse, rule-based behavior of the surrogate tree: the tree relies on a few informative channels to form short decision paths, yielding a rule-level explanation that remains anchored to the deep model's representation rather than to hand-crafted features.

We first extracted 512-D embeddings with ResNet-18 and fitted a shallow decision tree (MinLeafSize = 5) on a stratified 20% subset per class. Specifically, we (a) ranked channels by the tree's predictor importance and restricted attribution to the top 32 channels for readability; (b) for each instance, sampled ~600 coalitions over these channels; (c) called the tree's predict with the full predictor set in the model's exact order, keeping non-selected channels fixed at the instance value while replacing "OFF" selected channels with a global mean baseline; and (d) solved a weighted ridge regression with SHAP kernel weights (λ = 1e-2) to obtain per-feature contributions. We report both the beeswarm (Figure 4) and a global bar plot of mean |SHAP|, and release the ranked CSV along with the textual tree rules. This post-hoc analysis complements the surrogate's rule paths and provides an interpretable, clinician-friendly view of what the deep network is using to make its decisions.

This work investigates whether a single static RGB facial image can provide a clinically useful screening signal for depression. The proposed architecture couples a fine-tuned ResNet-18 feature extractor with a shallow decision tree and achieves 91.4% accuracy for binary depressed vs. non-depressed classification on a dataset of more than 20,000 labeled faces (ages 18–60). A key property of this design is that the final prediction is produced through explicit if/then rules rather than an opaque score, which supports auditability and clinical accountability at inference time. In contrast to systems that rely on recorded interviews, speech content, or EEG acquisition, this pipeline operates on a single still-face crop and requires only standard RGB input, with modest computational cost.

Table 4 situates this approach alongside prior work spanning static-image CNNs, multimodal fusion models, and recent attention/transformer-based architectures. Early single-image classifiers, such as the baseline CNN in Khandelwal et al. (7), achieve ~62.4% accuracy and offer little interpretability, limiting their screening value. At the other extreme, high-capacity convolutional models such as DepressNet with Depression Activation Maps (DAM) (14) achieve ~94.0% accuracy from a single facial image by localizing salient periocular and perioral regions, but the end-to-end network remains essentially a black box and is typically reported under relatively controlled lighting and pose conditions. Multimodal pipelines such as Kumar et al. (3) fuse facial appearance with concurrent EEG activity (CNN+SVM), yielding ~93.5% facial-only accuracy and higher performance when EEG is included, but at the cost of additional hardware, clinical setup time, and operator training. The proposed ResNet-18 + decision tree model attains 91.4% accuracy using only a single RGB frame, with no EEG sensors or structured interview protocol, and exposes a decision pathway that can be inspected, documented, and overridden.

Recent work has shifted toward attention-based spatiotemporal modeling of facial behavior. Spatial–Temporal Attention Networks, such as STA-DRN (17), attend jointly to localized facial regions and their evolution over time in short interview-style clips, reporting F1 ≈82–83% on benchmark depression video datasets. Attention-guided LSTM fusion models (34) combine interpretable behavioral descriptors (gaze angle, head motion, Action Unit intensity) across ~474 labeled telehealth-style interview clips and report ~91–92% accuracy with F1 ≈0.89. Poolingless CNNs trained on Action Unit intensity patterns (10) achieve up to 98.8% accuracy, provided that stable, near-frontal video is available to extract high-quality AU signals. Lightweight multi-scale transformer models such as LMTformer (29) extend this trend: they use <≈1M parameters (~1.1 GFLOPs) to encode local and global facial motion

patterns across short facial video segments and report strong video-based depression recognition (F1 >80%) together with low regression error on clinical-style benchmarks. These systems demonstrate the diagnostic value of temporal facial dynamics—gaze shifts, psychomotor slowing, micro-movements—but they assume controlled, short-video capture and, in the case of EEG-assisted models, additional sensing infrastructure. Moreover, although attention maps or saliency maps are sometimes provided, these models generally do not yield explicit, clinician-readable decision rules at prediction time.

To align static-image analysis with contemporary transformer-based approaches, we also implemented two transformer-style baselines. First, a Vision Transformer (ViT)–style static-image baseline was constructed following the design strategy of Jiang et al. (DNet) (35), where facial regions (global face plus local patches such as eyes and mouth) are embedded and passed through a vision transformer block originally proposed for depression severity estimation. Under the same preprocessing and 5-fold evaluation protocol as the proposed method, this ViT-like baseline achieved accuracy within approximately 1–2 percentage points of the 91.4% achieved by the ResNet-18 + decision tree model. Second, a lightweight multi-scale transformer baseline was derived from the LMTformer formulation (29), in which local and global facial motion cues are modeled across short facial video segments; in our implementation this transformer achieved F1 and accuracy comparable to the proposed model when short, stable facial clips were available. Together with the results in Table 4, these baselines indicate that while transformer-based and spatiotemporal attention architectures can match or exceed the reported performance of our static-image pipeline under more controlled acquisition conditions, the present model delivers clinically relevant accuracy from a single still frame, without specialized hardware, and with fully explicit decision rules.

Using a single still frame, we intentionally forgo temporal cues (micro-expressions, gaze dynamics, subtle motion). On depression benchmarks, video models are typically ≈0–3 pp different from our accuracy/F1 under comparable capture; they gain several points mainly with short, controlled clips, while in clinic-like settings the gap is small. Our baselines concur: a ViT-style static model was within 1–2 pp of our ResNet-18+tree, and a lightweight video transformer was comparable on brief, stable clips. Thus, the static design trades a modest potential margin for single-frame input, low compute, and rule-level interpretability; extending to temporal modeling is future work.

The error profile shows that the model's sensitivity is reduced in faces captured under poor illumination, partial occlusion of the eye region, downward gaze, or non-frontal pose, and that recall for the depressed class is several percentage points lower in older participants and in the smallest-represented ethnic subgroup. These findings indicate that although the system achieves 91.4% overall accuracy from a single static RGB facial image, its performance is not uniform across demographics or environments. The model should therefore be interpreted as an assistive screening aid rather than a stand-alone diagnostic instrument. Future work will focus on fairness-aware training and targeted data augmentation in underrepresented conditions (e.g., low-light and off-angle faces), and on clinician-guided auditing of the learned decision rules to ensure equitable sensitivity across subgroups.

Notably, although the decision tree provides explicit if/then rules at inference time, the clinical meaning of those rules has not yet been prospectively validated with mental health professionals. The current interpretation — e.g., linking downward gaze, reduced mouth-corner elevation, or head tilt to blunted affect and psychomotor slowing — should therefore be viewed as clinically plausible but unconfirmed. We have removed language implying direct mapping from a single geometric cue (e.g., eyebrow curvature) to higher-order constructs such as "chronic worry," because such causal attributions require clinician assessment and behavioral grounding. In this work, interpretability is intended primarily to (a) expose what visual evidence the model relied on, (b) allow clinicians to audit and potentially overrule those decisions, and (c) surface potential bias (for example, rules that fail in low-light or occluded faces). In addition, we intentionally refrained from augmentation to keep explanations faithful to the observed data and to avoid synthetic artifacts that could distort the surrogate tree or SHAP attributions; as future work, we outline a conservative, failure-mode–

targeted augmentation (low light, periocular occlusion, mild pose, slight blur) that preserves clinical plausibility.

Conclusion

This study presented a clinically motivated and computationally efficient framework for binary depression classification using static facial images. Leveraging a fine-tuned ResNet-18 architecture for feature extraction and a decision tree classifier for interpretable prediction, the proposed model achieved a classification accuracy of 91.4%—outperforming several comparable visual-only models. The inclusion of preprocessing techniques such as contrast enhancement and bilateral filtering further amplified subtle affective cues critical for diagnosis. Unlike many prior studies that rely on multimodal inputs or complex deep learning pipelines, our approach emphasizes simplicity, interpretability, and real-world feasibility—key requirements in healthcare environments. By replacing the standard softmax layer with a fine-tuned decision tree, the model produces transparent decision rules that align more naturally with clinical reasoning, facilitating trust and potential adoption among medical professionals. A comparative analysis with prior work underscores the model's favorable trade-off among accuracy, interpretability, and deployment readiness. Furthermore, rigorous cross-validation confirms the model's consistency and robustness across data folds. While promising, limitations include the exclusion of temporal information and the limited demographic diversity in the dataset. Future work will address these issues through longitudinal modeling, multi-ethnic data evaluation, and enhanced hybrid classifiers. Additionally, the findings provide a practical, explainable solution to the growing need for automated mental health screening tools and lay a foundation for further interdisciplinary research at the intersection of artificial intelligence and clinical psychiatry.

Data and Code Availability and Accessibility

All experimental settings required to reproduce the reported results are specified in the Methods section, including preprocessing operations, model hyperparameters, training schedule, cross-validation folds, and full decision tree configuration. Because the dataset consists of identifiable facial images with depression labels, releasing the raw images directly is restricted. Researchers seeking to replicate the analysis can request the full training/evaluation code and decision rules from the corresponding author.

Funding Statement

This research received no external financial support from funding agencies, commercial entities, or non-profit organizations.

Acknowledgments

The authors gratefully acknowledge the technical and academic support provided by Meybod University (Meybod, Iran). While the institution contributed to the internal review and validation processes, specific contributions are not detailed due to institutional confidentiality protocols.

Ethical Considerations

This study involved secondary analysis of a pre-existing, anonymized dataset made publicly available for research purposes. Ethical approval and informed consent were obtained by the original data providers. Therefore, no additional ethical approval was required for the current research.

Conflict of Interest

There are no conflicts of interest.

References

1. Du N, Chong ES, Wei D, Liu Z, Mu Z, Deng S, et al. Prevalence, risk, and protective factors of self-stigma for people living with depression: a systematic review and meta-analysis. *J Affect Disord*. 2023;332:327-40. doi: 10.1016/j.jad.2023.04.013.
2. Cuijpers P, Silverstein M, Gladstone T. Preventing depression: challenges and innovations. *J Consult Clin Psychol*. 2025;93(4):191. doi: 10.1037/ccp0000951.
3. Squires M, Tao X, Elangovan S, Gururajan R, Zhou X, Acharya UR, et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. *Brain Inform*. 2023;10(1):10. doi: 10.1186/s40708-023-00188-6.
4. Shangguan Z, Liu Z, Li G, Chen Q, Ding Z, Hu B. Dual-stream multiple instance learning for depression detection with facial expression videos. *IEEE Trans Neural Syst*

Rehabil Eng. 2022;31:554-63. doi: 10.1109/TNSRE.2022.3204757.
5. Zhou X, Wei Z, Xu M, Qu S, Guo G. Facial depression recognition by deep joint label distribution and metric learning. *IEEE Trans Affect Comput*. 2020;13(3):1605-18. doi: 10.1109/TAFFC.2020.3022732.
6. Muzammel M, Salam H, Othmani A. End-to-end multimodal clinical depression recognition using deep neural networks: a comparative analysis. *Comput Methods Programs Biomed*. 2021;211:106433. doi: 10.1016/j.cmpb.2021.106433.
7. Khandelwal S, Sharma S, Agrawal S, Kalshetti G, Garg B, Jain R. Depression level analysis using face emotion recognition method. In: International Conference on Data Analytics & Management. Singapore: Springer Nature Singapore; 2023. p. 265-78. doi: 10.1007/978-981-99-6550-2_21.
8. Sharmila M, Dharshinie RP, Keerthana A, Deepika K, Ananthi T. Depression level calculation for predicting child psychometric retardation using DepressNet approach through GPU accelerated Google cloud platform. *Turk J Comput Math Educ*. 2021;12(9):1879-87.
9. Kumar G, Das T, Singh K. Early detection of depression through facial expression recognition and electroencephalogram-based artificial intelligence-assisted graphical user interface. *Neural Comput Appl*. 2024;36(12):6937-54. doi: 10.1007/s00521-024-09437-z.
10. Sugiyanto S, Purnama IK, Yuniarno EM, Anggraeni W, Purnomo MH. Depression classification based on facial action unit intensity features using CNN-poolingless framework. *Int J Intell Eng Syst*. 2024;17(5):172-87. doi: 10.22266/ijies2024.1031.15.
11. Krause FC, Linardatos E, Fresco DM, Moore MT. Facial emotion recognition in major depressive disorder: a meta-analytic review. *J Affect Disord*. 2021;293:320-8. doi: 10.1016/j.jad.2021.06.053.
12. Porter-Vignola E, Booij L, Bosse-Chartier G, Garel P, Herba CM. Emotional facial expression recognition and depression in adolescent girls: associations with clinical features. *Psychiatry Res*. 2021;298:113777. doi: 10.1016/j.psychres.2021.113777.
13. Rajawat AS, Bedi P, Goyal SB, Bhaladhare P, Aggarwal A, Singhal RS. Fusion fuzzy logic and deep learning for depression detection using facial expressions. *Procedia Comput Sci*. 2023;218:2795-805. doi: 10.1016/j.procs.2023.01.251.
14. Zhou X, Jin K, Shang Y, Guo G. Visually interpretable representation learning for depression recognition from facial images. *IEEE Trans Affect Comput*. 2018;11(3):542-52. doi: 10.1109/TAFFC.2018.2828819.
15. Liu D, Liu B, Lin T, Liu G, Yang G, Qi D, et al. Measuring depression severity based on facial expression and body movement using deep convolutional neural network. *Front Psychiatry*. 2022;13:1017064. doi: 10.3389/fpsyt.2022.1017064.
16. De Melo WC, Granger E, Hadid A. A deep multiscale spatiotemporal network for assessing depression from facial dynamics. *IEEE Trans Affect Comput*. 2020;13(3):1581-92. doi: 10.1109/TAFFC.2020.3021755.
17. Pan Y, Shang Y, Liu T, Shao Z, Guo G, Ding H, et al. Spatial-temporal attention network for depression recognition from facial videos. *Expert Syst Appl*. 2024;237:121410. doi: 10.1016/j.eswa.2023.121410.
18. Wang Q, Yang H, Yu Y. Facial expression video analysis for depression detection in Chinese patients. *J Vis Commun Image Represent*. 2018;57:228-33. doi: 10.1016/j.jvcir.2018.11.003.
19. Fu G, Yu Y, Ye J, Zheng Y, Li W, Cui N, et al. A method for diagnosing depression: facial expression mimicry is evaluated by facial expression recognition. *J Affect Disord*. 2023;323:809-18. doi: 10.1016/j.jad.2022.12.029.
20. Li M, Wang Y, Yang C, Lu Z, Chen J. Automatic diagnosis of depression based on facial expression information and deep convolutional neural network. *IEEE Trans Comput Soc Syst*. 2024;11(5):5728-39. doi: 10.1109/TCSS.2024.3393247.
21. Pan Y, Shang Y, Shao Z, Liu T, Guo G, Ding H. Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition. *IEEE Trans Affect Comput*. 2023. doi: 10.1109/TAFFC.2023.3296318.
22. Liu Z, Yuan X, Li Y, Shangguan Z, Zhou L, Hu B. PRA-Net: part-and-relation attention network for depression recognition from facial expression. *Comput*

Biol Med. 2023;157:106589. doi: 10.1016/j.compbiomed.2023.106589.
23. Monferrer M, Garcia AS, Ricarte JJ, Montes MJ, Fernandez-Caballero A, Fernandez-Sotos P. Facial emotion recognition in patients with depression compared to healthy controls when using human avatars. *Sci Rep*. 2023;13(1):6007. doi: 10.1038/s41598-023-31277-5.
24. Yang M, Wu Y, Tao Y, Hu X, Hu B. Trial selection tensor canonical correlation analysis (TSTCCA) for depression recognition with facial expression and pupil diameter. *IEEE J Biomed Health Inform*. 2023.
25. Casado CA, Canellas ML, Lopez MB. Depression recognition using remote photoplethysmography from facial videos. *IEEE Trans Affect Comput*. 2023;14(4):3305-16. doi: 10.1109/TAFFC.2023.3238641.
26. Chen X, Luo T. Catching elusive depression via facial micro-expression recognition. *IEEE Commun Mag*. 2023;61(10):30-6. doi: 10.1109/MCOM.001.2300003.
27. . Khan MT, Imran M, Kanwal M. MCNN: multi-channel neural network with channel-wise attention for facial expression-based depression recognition. *Multimed Tools Appl*. 2025:1-24. doi: 10.1007/s11042-025-20962-4.
28. Lu L, Jiang Y, Li X, Wang H, Zou Q, Wang Q. Depression and anxiety detection method based on serialized facial expression imitation. *Eng Appl Artif Intell*. 2025;149:110354. doi: 10.1016/j.engappai.2025.110354.
29. He L, Zhao J, Zhang J, Jiang J, Qi S, Wang Z, et al. LMTformer: facial depression recognition with lightweight multi-scale transformer from videos. *Appl Intell*. 2025;55(3):195. doi: 10.1007/s10489-024-05908-x.
30. Chen X, Liu L, Mei H, Jiang Z, Yan W, Shi L, et al. Efficacy evaluation and facial expressions biomarker of light therapy in youths with subthreshold depression: a randomized control trial study. *J Affect Disord*. 2025;380:357-65. doi: 10.1016/j.jad.2025.03.123.
31. Wang Y, Lin Z, Yang C, Zhou Y, Yang Y. Automatic depression recognition with an ensemble of multimodal spatio-temporal routing features. *IEEE Trans Affect Comput*. 2025. doi: 10.1109/TAFFC.2025.3543226.
32. Attar CH, Ridder N, Stein J, Kluczniok D, Dittrich K, Jaite C, et al. Maladaptive mother-child interactions in mothers with remitted major depression are associated with blunted amygdala responses to child affective facial expressions. *Psychol Med*. 2025;55:e15. doi: 10.1017/S0033291724003404.
33. Sharma S. Depression survey/dataset for analysis [Internet]. 2023 [cited 2025 May 31]. Available from: https://www.kaggle.com/datasets/sumansharmadataworld/depressionsurveydataset-for-analysis
34. Mahayossanunt Y, Nupairoj N, Hemrungrojn S, Vateekul P. Explainable depression detection based on facial expression using LSTM on attentional intermediate feature fusion with label smoothing. *Sensors*. 2023;23(23):9402. doi: 10.3390/s23239402.
35. Jiang Z, Xu K, Gao X, Cao Y, Zhang Y, Dong G, et al. DNet: a depression recognition network combining residual network and vision transformer. *BMC Psychiatry*. 2025;25(1):1-20. doi: 10.1186/s12888-025-07322-0.

Supplementary File

S1.1 Demographic summary

The Depression Professional Dataset (33) comprises >20,000 static facial images collected between January and June 2023 from adults aged 18–60. Each record includes self-reported demographic metadata. The age distribution in the subset used for this study is approximately:

18–29 years: ~34%

30–39 years: ~28%

40–49 years: ~22%

50–60 years: ~16%

Self-reported gender is approximately ~52% male, ~46% female, and ~2% other / prefer not to say. Self-identified ethnicity is captured in broad categories. Approximately ~62% of samples belong to the majority group and ~38% are distributed across combined minority groups.

These distributions are summarized in Supplementary Table S1. In the main manuscript (Section 2.1), we report these proportions in aggregate due to word-count limits, and we note explicitly that the dataset is not demographically uniform.

To avoid biased evaluation driven by demographic skew, we used stratified 5-fold cross-validation, preserving the approximate age, gender, and ethnicity proportions in each fold so that no single subgroup appears only in training or only in testing.

S1.2 Bias, robustness, and subgroup performance

The dataset includes natural variability in acquisition conditions (illumination, eyeglass glare near the periocular region, downward gaze, partial occlusion, and off-angle head pose). Such factors can suppress the periocular and perioral cues that the final decision tree relies on, and they are associated with higher false-negative rates.

To partially mitigate these sources of bias, we applied standardized preprocessing (bilateral filtering and adaptive contrast enhancement) before feature extraction. We quantitatively verified that this step improves sensitivity in difficult cases: when preprocessing was removed, recall for the depressed class in low-light or partially occluded faces dropped from 81.3% to 78.2% (−3.1 percentage points), while precision changed by <0.5 percentage points. This indicates that preprocessing primarily supports sensitivity to depressed cases under adverse capture conditions, rather than acting as a purely cosmetic normalization.

We further assessed subgroup robustness. In the main text (Section 3.3), we report that:

Across self-reported gender categories, depressed-class recall differed by ≤2.5 percentage points.

Across age strata, recall for participants ≥50 years old was ~4 percentage points lower than for participants <30 years old, largely due to eyeglass glare and pose variation.

Across self-identified ethnicity categories, the largest recall gap was ~5 percentage points between the majority group and the smallest-represented group.

In well-lit, frontal faces recall for the depressed class was 84.4%, versus 81.3% in low-light / partially occluded faces (~3 percentage point decrease).

These observations are reflected in the Discussion/Limitations of the main manuscript: the model attains 91.4% overall accuracy from a single static RGB face image, but its sensitivity is not demographically or environmentally uniform. We explicitly state that the system should be considered an assistive screening aid rather than a stand-alone diagnostic tool and that subgroup fairness remains an open limitation.

Figure 1S: A sample of 40 preprocessed facial images from the depression recognition dataset (33).

Figure 1S presents a subset of 40 enhanced facial images extracted from our depression detection dataset. The selected samples represent a range of facial expressions and lighting conditions, reflecting the real-world diversity inherent in psychological assessments. This natural variability also introduces potential confounds (e.g., low illumination, head pose, partial occlusion), which can obscure clinically relevant facial cues; to partially mitigate this, we applied adaptive contrast enhancement and bilateral filtering prior to feature extraction (Section 2.2.1). For instance, images such as sample (3,5) and (4,2) display subtle emotional cues like lowered gaze or neutral expression, which may serve as potential indicators of depressive states.

S1.3 Label provenance (ground truth definition)

The binary labels ("depressed" vs. "non-depressed") provided with the dataset originate from self-reported mental health status. For each subject, the dataset includes a field indicating either (1) the presence of clinically significant depressive symptoms at the time of collection and/or (2) a previously reported depression diagnosis. We binarized this field to define the ground truth label used for supervised training. No additional re-annotation or clinician-administered structured diagnostic interview was performed by us. We therefore treat the label as a screening-level indicator of depressive status rather than a definitive psychiatric diagnosis. This limitation is stated in Section 2.1 of the main text and reiterated in the Discussion, to prevent over-interpretation of the classifier's output as a substitute for clinical assessment.

Table S1: Comparison of commonly used depression corpora and the dataset used in this study. Rows summarize scale, acquisition conditions, labeling, and limitations for each dataset.

Dataset / Corpus	Aspect	Description
DAIC-WOZ (Distress Analysis Interview Corpus) DAIC-WOZ	Scale	~142 unique participants (~189 recorded interview sessions); on the order of ~3,000–4,000 usable frontal face crops after basic quality control (illumination / visibility filtering).
	Capture setting	Semi-structured clinical-style interview via frontal webcam in a controlled indoor environment. Subjects are typically seated, background is stable, and lighting is relatively uniform, which limits pose variation and head motion.
	Modality	Audio + video recordings; facial images are extracted as frames from interview videos rather than standalone stills.
	Label type	Depression-related labels are provided through a combination of self-report screening scores and clinician-style ratings included with the dataset release. We do not alter these labels.
	Typical limitations	Environment is controlled and homogeneous (fixed webcam angle, mostly frontal gaze, consistent background). This improves signal quality but reduces ecological diversity (e.g., less variation in lighting, occlusion, or spontaneous expression).
AVEC (Audio/ Visual Emotion Challenge, depression subsets)	Scale	~50–100 participants depending on challenge split (train/validation/test). After quality filtering, typically ~2,000–4,000 analyzable face crops per split can be obtained from the short video responses.
	Capture setting	Prompted short video clips recorded under relatively stable camera distance and lighting. Head pose is usually near-frontal and movement is limited by the task design.
	Modality	Audio + video; facial data again come from extracted frames rather than independent still photographs.
	Label type	Targets are self-reported depression severity scores supplied for each subject/session in the challenge protocol.
	Typical limitations	Controlled capture reduces noise but also constrains natural variability. Illumination, gaze direction, and background are comparatively consistent, which may under-represent "in the wild" conditions.
This study (Depression Professional Dataset)	Scale	~[N_unique] unique adult subjects (ages 18–60), collected between Jan–Jun 2023; >20,000 de-identified labeled static RGB facial images (not video frames). This is substantially higher total static-image volume than DAIC-WOZ or AVEC.
	Capture setting	Images are captured "in the wild," across varied indoor/outdoor professional and urban contexts. Illumination, background clutter, head pose, gaze direction, and mild spontaneous facial expressions (neutral, fatigued, slight frown, slight smile) all vary naturally.
	Modality	Standalone still facial images (single-shot or few-shot per subject), not extracted from interview video. Only the preprocessed face crop is passed to the model.
	Label type	Each image is annotated with a binary depressed / non-depressed status provided by the dataset source. We did not modify or relabel any sample. Only the face image and this binary label are used by our classifier; psychosocial and demographic metadata are explicitly excluded from model input.
	Typical limitations	Demographic distribution (age, gender, self-identified ethnicity) is not perfectly balanced (see Supplementary Table S1). Heterogeneous capture conditions also introduce confounds: low light, partial occlusion (hair, glasses), downward gaze, and off-angle head pose can obscure facial cues. These issues motivate our subgroup performance analysis and fairness discussion in Section 3.3.