

Spatiotemporal Deep Learning for Robust Depression Severity Classification in Clinical Settings

Anum Saher¹, Ghulam Gilanie¹, Saeed Akhtar Chaudhry², Hafeez Ullah³,
Muhammad Sajid^{1,4*}, Maryam Mazher⁴, Umme Kalsoom¹

¹ The Islamia University of Bahawalpur, Faculty of Computing,
Pakistan

² Combined Military Hospital/CMH Institute of Medical Sciences,
Department of Psychiatry and Behavioral Sciences,
Pakistan

³ The Islamia University of Bahawalpur, Biphotonic Imaging Techniques Laboratory,
Pakistan

⁴ Instituto Politécnico Nacional, Centro de Investigación en Computación,
Mexico

{saheranum1, saeed.akhtar.dr, ummekalsoom3932}@gmail.com,
{ghulam.gilanie, hafeezullah }@iub.edu.pk, {msajida26, mmazhera26}@cic.ipn.mx

Abstract. Depression is among the most prevalent and disabling mental health disorders worldwide, yet its clinical diagnosis still depends largely on subjective self-report instruments and expert judgment, which are difficult to scale and prone to inter-rater variability. Automated, objective assessment of depression severity from observable behavior could substantially extend access to screening, but most existing video-based approaches rely on static images or shallow temporal modelling and are validated on small or demographically homogeneous datasets, limiting their clinical generalizability. We propose a deep learning model that classifies depression severity from facial videos. It is evaluated on a real clinical dataset of 177 participants, including healthy controls and patients with mild, moderate, and severe depression. The framework couples a ConvNeXt-Tiny spatial backbone, enhanced with Squeeze-and-Excitation and Convolutional Block Attention modules, with a 256-unit Long Short-Term Memory (LSTM) network applied frame-wise through a time-distributed design to model the temporal dynamics of facial behavior. Severe class imbalance is mitigated through weighted categorical cross-entropy and targeted data augmentation, and patient-level stratification is enforced so that no participant contributes frames to both the training and evaluation splits. Under participant-level five-fold cross-validation, the proposed model achieves strong classification performance across all four severity levels, attaining

82.7% accuracy, a macro-averaged F1-score of 78.9%, and a macro-averaged AUC of 0.917 on held-out participants, and outperforming the strongest single-backbone baseline (Swin Transformer, 71.2% accuracy) by 11.5 percentage points. We further examine the validation protocol and the role of strict patient-level separation in interpreting these results, situating them against the leakage risk inherent in densely sampled video. The proposed framework offers a lightweight, behaviorally grounded, and reproducible approach to automated depression severity assessment, with direct relevance to scalable clinical screening.

Keywords. Depression severity classification; spatiotemporal deep learning; facial video analysis; ConvNeXt; LSTM; attention mechanisms; affective computing; clinical mental health screening.

1 Introduction

Depression is among the most prevalent and disabling mental health disorders worldwide, disrupting emotional regulation, behavior, and daily functioning while imposing a substantial burden on individuals and health systems. Its clinical assessment, however, still relies predominantly on self-report instruments and

expert judgement, which are subjective, labor-intensive, and difficult to scale particularly in the resource-limited settings where the unmet need is greatest. These constraints motivate the development of automated, objective markers of depression that can extend the reach of screening without displacing clinical expertise. Facial behavior is a promising substrate for such markers: subtle expressions, gaze patterns, and micro-expressive dynamics encode affective state, yet are difficult to quantify through manual observation [1].

Crucially, these signals are inherently temporal, unfolding over seconds of behavior rather than residing in any single frame [2]. Most existing video-based approaches nonetheless rely on static images or shallow temporal modelling and are validated on small, demographically homogeneous cohorts, which limit both their sensitivity to behavioral dynamics and their generalizability to real clinical populations [3].

In this work, a spatiotemporal deep learning [4] framework is proposed to classify depression severity from facial video recordings. The study used a clinically collected dataset acquired at the Combined Military Hospital Institute of Medical Sciences, Bahawalpur, Pakistan, consisting of video data from healthy individuals and patients diagnosed with mild, moderate, and severe depression [5]. This real-world data set enables the development and evaluation of models under practical clinical conditions [6].

The proposed framework is designed around this temporal premise. Spatial features are extracted per frame by a ConvNeXt-Tiny backbone, selected because of its strong representational capacity at a parameter budget (~28M) compatible with deployment on modest clinical hardware. To make the spatial encoder selective rather than exhaustive, Squeeze-and-Excitation and Convolutional Block Attention modules recalibrate channel- and region-wise responses toward the facial cues that are most informative for identifying depressive states [7].

A time-distributed application of this encoder yields a frame-ordered sequence of embeddings, which a 256-unit Long Short-Term Memory network integrates into a representation of how facial behavior evolves over time.

Against this backdrop, the present study makes the following principal contributions:

- We introduce a clinically curated facial-video dataset for four-level depression severity classification, acquired under routine conditions at a tertiary psychiatric facility and labelled by expert clinical assessment. In contrast to the acted, laboratory, or binary-labelled corpora that dominate prior work, this cohort provides ecologically valid, severity-graded supervision, addressing a recognized scarcity of clinically grounded benchmarks.
- We propose lightweight spatiotemporal architecture that couples an attention-augmented ConvNeXt-Tiny backbone (Squeeze-and-Excitation and Convolutional Block Attention) with a time-distributed LSTM. Unlike face-only static classifiers and heavier transformer-based video models, this design jointly captures discriminative spatial cues and their temporal evolution at a parameter budget suited to clinical deployment.
- We benchmark the framework against strong backbones (ResNet-18, Swin Transformer, DenseNet-121, and ViT) under a strictly patient-level, leakage-controlled protocol in which no participant contributes frames to more than one split a safeguard frequently omitted in densely sampled video studies and we critically examine how this protocol shapes the interpretation of the reported performance.

The remainder of this paper is organized as follows. Section 2 reviews related work and positions our contribution within the existing literature; Section 3 details the dataset, preprocessing pipeline, and proposed architecture; Section 4 presents and discusses the experimental results, including a critical analysis of the validation protocol; and Section 5 concludes with the study's limitations and directions for future work.

2 Literature Review

Automated depression assessment has been pursued across a range of sensing modalities gait, electroencephalography (EEG), speech, and facial behavior paired with an increasingly sophisticated

Table 1. State-of-the-art of the Spatiotemporal Deep Learning for Robust Depression Severity Classification in Clinical Settings

	Source / Venue	Technical Approach	Data Source	Performance Metrics	Key Limitation
1	Facial multimodal diagnosis Front. Psychiatry (2025) [8]	Spatiotemporal visual attention fused with GCN + LSTM over facial video and audio	AVEC2013/2014; E-DAIC	MAE 3.51 (PHQ-8, E-DAIC); F1 0.853	Requires synchronized audio + video; validated on interview corpora
2	IMDD-Net Front. Psychiatry (2025) [9]	Inception-ResNet-v2 spatial features + volume local directional motion descriptor + multilayer Bi-LSTM	AVEC 2014	RMSE 7.55; MAE 5.75; binary acc. 0.79	Severity regression on an acted corpus; limited demographic diversity
3	Deep Spectro temporal speech network Sci. Reports (2025) [10]	EfficientNet-B3 on Mel-spectrograms + volume local descriptor (VLNEP)	AVEC2013/2014	AVEC2014 MAE 6.21 / RMSE 7.62; AVEC2013 MAE 6.57 / RMSE 7.76	Speech-only; no facial or behavioral cues
4	MHA-GCN ViT Math. Biosci. Eng. (2025) [15]	Multi-head attention GCN + Vision Transformer fusing EEG and speech	EEG + speech	Classification accuracy (EEG + speech fusion)	EEG acquisition limits routine / remote screening
5	FacialPulse arXiv (2024) [11]	RNN-based temporal modelling of facial-landmark sequences	Facial-landmark video benchmarks	F1 / accuracy (temporal landmarks)	Landmark-based; sensitive to face-detection error
6	Spatiotemporal fused EEG network arXiv (2024) [16]	Fusion of EEG electrode spatial topology with time-window transitions	EEG (MDD)	Classification accuracy (EEG)	EEG-based; controlled acquisition required
7	Kinect-based gait recognition BMC Geriatrics (2025) [2]	Spatiotemporal gait-parameter extraction (step height, walking speed, body sway) + Random Forest classifier with SHAP interpretability	50 older adults (25 depressed, 25 non-depressed), CES-D screened	AUC-ROC 0.911	Requires depth-sensing hardware; small, balanced subsample; captures locomotion, not facial affect

set of deep learning architectures. We organize the relevant literature by modality below, in each case foregrounding the design choices, reported performance, and the limitations that collectively motivate the present work.

Motor and physiological behavior offers one entry point.

A Kinect-based gait recognition system, evaluated on 50 community-dwelling older adults (25 depressed, 25 non-depressed) screened with the 10-item CES-D, identified significant differences in spatiotemporal gait parameters such

as step height, walking speed, and body sway; a Random Forest classifier reached an AUC-ROC of 0.911, with SHAP analysis confirming the most informative gait features.

While this establishes gait as a low-cost, non-invasive signal, the approach depends on dedicated depth-sensing hardware and a small, balanced subsample, and it captures locomotion rather than the affective facial dynamics central to clinical depression assessment.

Closest to the present work is the body of research on facial and multimodal video analysis,

which exploits the affective behavior that clinicians themselves attend to. A multimodal Automatic Depression Recognition framework integrated a Transformer-based long-term behavior module, an attention-guided noise-elimination strategy, and a spatial-temporal ensemble fusing texture, Facial Action Primitives, and remote photoplethysmography; validated on AVEC 2013/2014 and the clinically labelled CMDep corpus (123 individuals), it achieved MAE/RMSE as low as 5.09/6.83 and was deployed in a mobile self-assessment application.

Targeting a clinical population, a study of depression in Parkinson's disease evaluated ViT, Video Swin Tiny, and a 3D CNN-LSTM with attention across ON/OFF medication states on 1,875 videos from 178 patients, with Video Swin Tiny reaching 94% accuracy on binary classification but only 87.1% on multiclass severity underscoring how much harder graded severity is than presence/absence, and how medication can confound facial markers.

Broader surveys of micro-expression and multimodal pipelines (DETR, BiLSTM, ViT, Mixture-of-Experts, and LSTM-based fusion) across CASME II, AVEC2014, and E-DAIC report continued gains in accuracy and interpretability yet remain largely anchored to acted or laboratory recordings and binary or coarse labels.

At a different level of granularity, population-scale work such as DepGCN models spatial dependencies across census-tract indicators to localize depression burden, improving on regression baselines by a 32% reduction in mean squared error. Such approaches inform public-health policy rather than individual assessment and are therefore complementary to rather than competitive with the patient-level screening addressed here.

Electroencephalography has attracted sustained attention as a more direct physiological window onto depression, and a cluster of recent architectures reports strong discrimination on it. The TSF-MDD framework builds a four-dimensional temporal spatial-frequency representation of EEG and, with a hybrid 3D-CNN/Capsule-Network encoder and subject-independent partitioning, reaches 92.1% accuracy on the Mumtaz2016 dataset. The EEG Mind-Transformer combines spatial-temporal fusion,

hierarchical graph representation, and dynamic temporal graph attention to model brain dynamics, reporting 92.5% accuracy and a 94.2% AUC across multiple EEG datasets. Emphasizing generalization, a large multi-site study using p-SHAP feature selection and a Random Forest correlated EEG markers with HAMD-17 scores ($r = 0.56$, 75% binary accuracy) while surfacing interpretable alpha/beta-band biomarkers, and a frequency-domain channel-selection method coupled to a Multiscale Spatiotemporal Convolutional Attention Network (MSTAN) attained 97.24% accuracy on MODMA. Despite these gains, EEG-based pipelines share a decisive practical limitation: they require electrode placement and controlled acquisition, which constrains deployment in routine, large-scale, or remote screening far more than passively observed behavior does.

The 2024 and 2025 papers have concentrated on exactly the facial-video setting addressed here, and its trajectory both motivates and bounds the present design. Jin et al. fused facial video with audio through a spatiotemporal visual-attention module and a GCN-LSTM stream, lowering PHQ-8 estimation error to an MAE of 3.51 on E-DAIC [8], while IMDD-Net combined Inception-ResNet-v2 spatial features, a volume local directional motion descriptor, and a multilayer Bi-LSTM to reach an RMSE of 7.55 on AVEC 2014 [9].

A speech-only Spectro temporal network over Mel-spectrogram sequences attained a comparable MAE of 6.21 on the same corpus [10], and Facial Pulse showed that lightweight recurrent modelling of facial-landmark sequences can approach these results at lower cost [11].

The most closely related architectures sharpen this trajectory further: Pan et al. introduced a spatial-temporal attention network that learns which facial regions and frames are most diagnostic of depression on AVEC video [12].

He et al. addressed deploy ability with LMTformer, a lightweight multi-scale transformer that encodes facial behavior at several temporal granularities within a clinically tractable parameter budget [13].

A FAU-guided graph-representation network coupled a feature-aware edge updater, graph-attention layers, and a temporal-attention module

Table 2. Details of patients and healthy controls

Mental health status	No. of patients	Gender	Age	Marital Status	Occupation
Healthy	7	Male	21-38	Unmarried	Students, jobholders, and farmers
Healthy	8	Male	31-42	Married	Jobholders, businessmen
Healthy	6	Male	40-54	Married	Job holders, businessmen, lawyers
Healthy	9	Male	50-65	Married	Retired persons, businessmen
Healthy	6	Female	19-30	Unmarried	Students, working women
Healthy	7	Female	31-46	Married	Housewife
Depression (Mild)	3	Male	22-34	Unmarried	Students, job holder
Depression (Mild)	2	Male	23-42	Married	Job holder, Driver
Depression (Mild)	4	Male	47-57	Married	Job holder, Businessman, Farmer
Depression (Mild)	7	Male	43-59	Married	Job holder, Businessman, Farmer
Depression (Mild)	5	Female	19-35	Unmarried	Students, jobholders, social workers
Depression (Mild)	8	Female	32-54	Married	Job holder, Housewife
Depression (Mild)	9	Female	51-65	Married	Retired, Housewife
Depression (Moderate)	11	Male	23-39	Unmarried	Students, jobholders, social workers
Depression (Moderate)	9	Male	29-47	Married	Job holder, Farmer
Depression (Moderate)	8	Male	43-59	Married	Job holder, Businessman, Lawyer
Depression (Moderate)	7	Male	46-61	Married	Job holder, Businessman, Farmer
Depression (Moderate)	9	Female	22-37	Unmarried	Students, Job holder
Depression (Moderate)	10	Female	35-56	Married	Job holder, Housewife, Social worker
Depression (Moderate)	8	Female	49-69	Married	Housewife, Retired, Police
Depression (Severe)	6	Male	20-39	Unmarried	Students, Job holder
Depression (Severe)	4	Male	27-46	Married	Job holder, Farmer, Businessman
Depression (Severe)	5	Male	39-56	Married	Businessman, Farmers, Social worker
Depression (Severe)	4	Male	43-59	Married	Job holder, Businessman, Farmer
Depression (Severe)	5	Female	21-36	Unmarried	Students, Job holder
Depression (Severe)	6	Female	34-58	Married	Job holder, Housewife, Social worker
Depression (Severe)	4	Female	35-70	Married	Housewife, Social worker, Retired

to model the episodic dynamics of depressive effect on AVEC 2014/2019 [14].

These works collectively confirm the premise underlying our design that depression is encoded in the spatiotemporal evolution of facial behavior rather than in any isolated frame yet they remain tied to acted or interview corpora (AVEC, E-DAIC) of predominantly Western subjects, depend on auxiliary modalities or on landmark detectors whose failures propagate downstream, optimize continuous BDI/PHQ regression rather than the graded severity that clinical triage requires, and report frame- or video-level error without the strict patient-level separation needed to exclude identity leakage.

The newest physiological models echo the same pattern: a multi-head attention graph-convolution network combined with a Vision Transformer (MHA-GCN_ViT) jointly encodes the frequency and spatiotemporal structure of EEG and speech [15], and a spatiotemporal fused network that explicitly models electrode spatial topology and time-window transitions improves major-depressive-disorder detection [16]. Both raise discriminative power by coupling spatial and temporal representations, but like their predecessors remain bound to electrode-based acquisition under controlled conditions, precisely the constraint that a passive, video-only approach removes.

Taken together, this body of work reveals several persistent limitations that constrain the clinical translation of automated depression assessment. Many models generalize poorly across diverse populations because they are trained on demographically homogeneous data. Existing multimodal approaches frequently lack dynamic fusion, temporal alignment, and cross-modal attention, and models developed under laboratory conditions often degrade in real-world clinical settings owing to acquisition noise, computational cost, and limited interpretability. Most systems assess depression at a single time point, disregarding symptom dynamics, subtypes, and comorbidities, while biases related to gender, age, and dataset composition are seldom addressed and physiological confounders such as medication can erode diagnostic specificity. Data scarcity and the insufficient validation of synthetic or augmented datasets further undermine

reliability. The present study is designed to address these gaps directly: it is developed on an ecologically valid clinical cohort acquired under routine psychiatric conditions, it couples attention-augmented spatial encoding with explicit temporal modelling to capture behavioral dynamics rather than isolated snapshots, and it adopts a patient-level, leakage-controlled validation protocol intended to yield generalization estimates that remain meaningful in deployment. Table 1 consolidates the representative studies surveyed above, summarizing their modalities, technical approaches, datasets, reported performance, and principal limitations.

3 Methodology

3.1 Dataset Acquisition

The dataset comprises video recordings of 2–5 minutes duration per participant, collected at the Department of Psychiatry and Behavioral Sciences, Combined Military Hospital (CMH) Institute of Medical Sciences, Bahawalpur, Pakistan, under the supervision of Prof. Dr. Saeed Akhtar Chaudhry. All subjects provided informed consent through signed forms authorizing the use of their video recordings for research purposes.

The dataset consists of 177 individuals including 43 healthy controls and 134 patients diagnosed with depression, categorized into mild (38), moderate (62), and severe (34) levels based on clinical assessments.

Table 2 summarizes the demographic composition of the cohort. From these videos, a total of 2548800 frames were extracted, distributed as follows: Normal controls: 619200 frames, Mild depression: 547200 frames, Moderate depression: 892800 frames, and Severe depression: 489600 frames. Recording duration varied from participant to participant and was generally longer for healthy controls; because frames were extracted at a fixed 60 fps, this variation in session length accounts for the differing frame totals across categories. Data collection adhered to institutional ethical protocols, ensuring patient privacy and data confidentiality. The study was approved by the Office of Research, Innovation and Commercialization

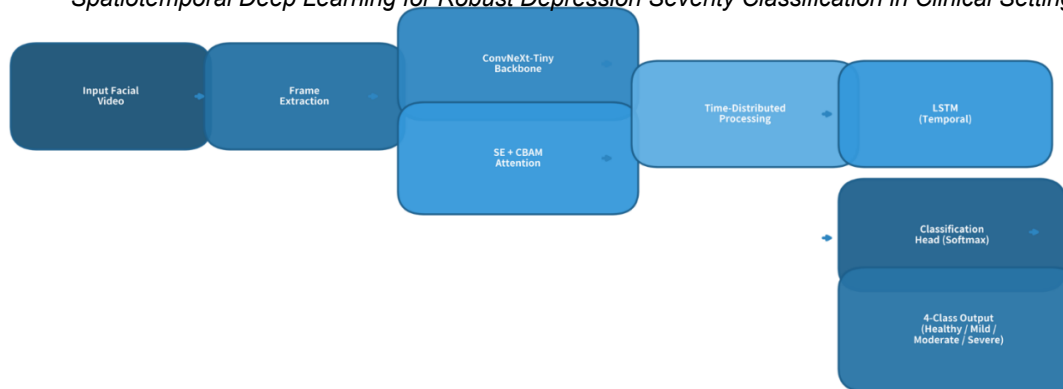


Fig. 1. Proposed framework architecture

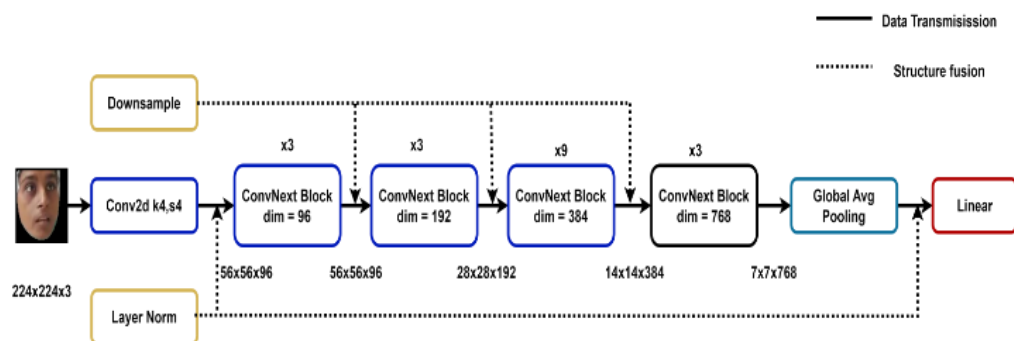


Fig. 2. The ConvNeXt tiny model

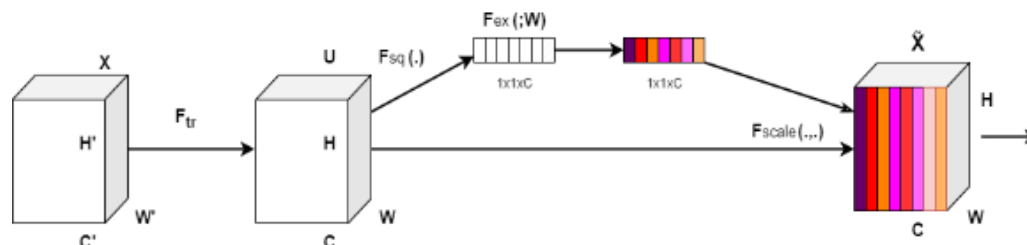


Fig. 4. The squeeze-and-excitation model

(ORIC), The Islamia University of Bahawalpur, Pakistan (Approval No. 338/ORIC, dated August 23, 2024).

3.2 Data Preprocessing

Each video was decoded into frames and preprocessed through a uniform pipeline prior to model training.

Frames were extracted at 60 fps, resized to 224×224 pixels, and passed through a Multi-Task Cascaded Convolutional Neural Network

(MTCNN) for face detection and alignment. Detected faces were cropped to remove background, and pixel intensities were normalized to the [0,1] range to ensure photometric consistency across recordings.

To enlarge the effective training distribution and improve generalization, label-preserving augmentation random rotations, horizontal flips, translations, and brightness perturbations was applied to the training partition only, so that no augmented variant of an evaluation frame could influence model fitting.

Class imbalance was addressed during optimization rather than preprocessing, as detailed in the Training Procedure. Critically, all partitioning was performed at the participant level: the cohort was split so that every frame from a given individual was confined to a single partition, preventing identity leakage between training and evaluation. Within each severity category, 80% of participants were assigned to training and 20% to validation, yielding 141 training participants (34 normal, 30 mild, 50 moderate, 27 severe) and 36 validation participants (9 normal, 8 mild, 12 moderate, 7 severe).

Frames from each participant were grouped into non-overlapping sequences of 32 consecutive frames, which served as the temporal input units to the model. Model selection used five-fold cross-validation over the training participants, and final performance was reported on the held-out validation participants under five-fold cross-validation.

3.3 Proposed Framework Overview

The proposed model integrates spatial and temporal components within a unified hybrid architecture. A time-distributed ConvNeXt-Tiny encoder is applied independently to each video frame to extract frame-level spatial representations prior to temporal aggregation. These representations are passed on to a 256-unit LSTM layer, which captures the temporal dynamics of the facial-feature sequence.

The resulting temporal encoding is processed by a stack of fully connected layers that map it to the four target severity categories normal, mild, moderate, and severe and a final SoftMax layer produces the corresponding class probabilities. This design allows the network to jointly exploit discriminative spatial cues and their temporal evolution, which together are essential for distinguishing graded levels of depressive severity. The overall architecture of the proposed framework is shown in Fig. 1.

3.3.1 ConvNeXt-Tiny Spatial Encoder

ConvNeXt-Tiny is a modern convolutional architecture whose design is informed by the macro-architectural principles of Vision Transformers while retaining a purely

convolutional formulation. Through design refinements such as larger kernel sizes, depth wise convolutions, inverted bottlenecks, and layer normalization, it achieves representational capacity competitive with transformer-based models at substantially lower computational cost. With approximately 28 million parameters, the network is lightweight enough for efficient training on limited GPU resources yet sufficiently expressive for fine-grained facial feature extraction. In this work, ConvNeXt-Tiny serves as the frame-level spatial encoder, producing embeddings that capture the facial characteristics most relevant to depressive states; its favorable accuracy efficiency trade-off makes it well suited to high-resolution, frame-wise processing of video data. The structure of this spatial encoder is shown in Fig. 2.

3.3.2 Time-Distributed Application

The time-distributed wrapper applies an identical layer here, the ConvNeXt-Tiny encoder independently and with shared weights to every time step of an input sequence. In the context of video analysis, this means that the same spatial feature extractor processes each frame separately, producing a temporally ordered sequence of frame-level embeddings while preserving parameter efficiency. This mechanism provides a principled bridge between per-frame spatial encoding and subsequent sequence modelling, ensuring that the spatial representations supplied to the recurrent LSTM stage are mutually consistent across frames. This time-distributed application is illustrated in Fig. 3.

3.3.3 Squeeze-and-Excitation (SE)

The Squeeze-and-Excitation (SE) block is a lightweight channel-attention mechanism that adaptively recalibrates feature responses across channels. It operates in three stages: a squeeze step applies global average pooling to compress each channel into a single descriptor that summarizes its global spatial context; an excitation step passes these descriptors through a compact multilayer perceptron to learn non-linear, channel-wise importance weights; and a scaling step multiplies the original feature maps by these learned weights. By emphasizing informative channels and suppressing less relevant ones, the

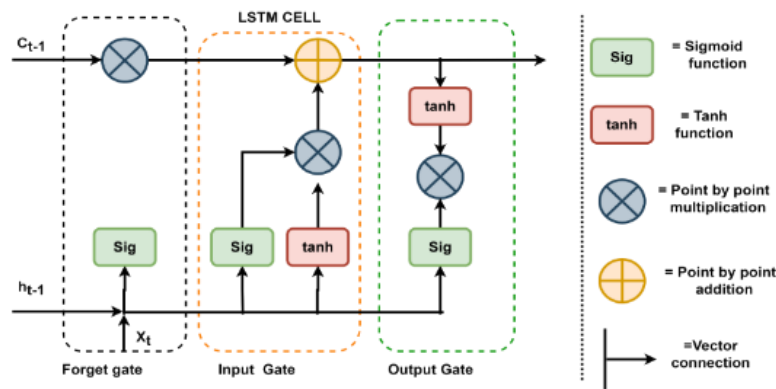


Fig. 3. Time distributed CNN

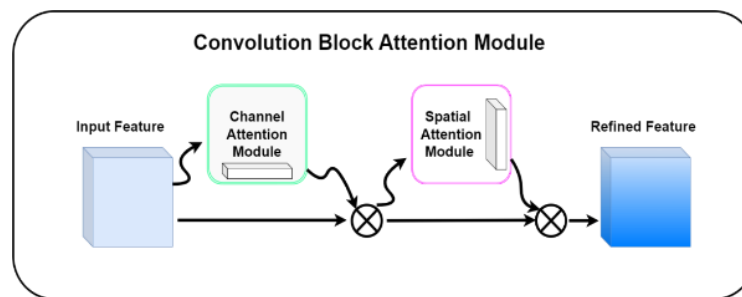


Fig. 5. The convolution block attention module

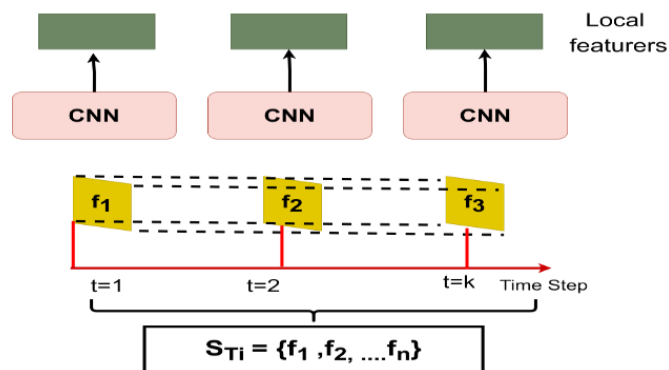


Fig. 6. Example of the LSTM model

SE block enables the network to focus on the facial feature channels most discriminative of depressive states at negligible additional computational cost. The SE block is depicted in Fig. 4.

3.3.4 Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) extends the Squeeze-and-Excitation block by

adding a spatial-attention stage, forming a two-stage mechanism that refines features along both the channel and spatial dimensions and applies the two stages sequentially.

The channel-attention stage first aggregates spatial information using parallel global average- and max-pooling, passes both descriptors through a shared multilayer perceptron, and sums the results before a sigmoid produces per-channel

weights that emphasize the most informative feature maps. The spatial-attention stage then pools the channel-refined features across the channel axis again using average and max operations concatenates the two maps, and applies a convolution followed by a sigmoid to produce a two-dimensional attention map that highlights the most discriminative facial regions.

Coupling CBAM with the SE block allows the encoder to recalibrate both which channels and which spatial locations contribute to the depression-severity decision, at negligible additional computational cost. The CBAM structure is shown in Fig. 5.

3.3.5 Temporal Modeling Using LSTM (256 units)

The Long Short-Term Memory (LSTM) network models the temporal evolution of the frame-level facial embeddings produced by the spatial encoder. Its gated recurrent structure mitigates the vanishing-gradient problem and allows the network to retain information across long frame sequences, making it well suited to capturing the gradual and often subtle temporal patterns associated with depressive behavior—for example, changes in expressivity, gaze, and blink dynamics over time.

A 256-unit LSTM is employed in this study; its output sequence encoding is subsequently passed to the classification head to predict depression severity.

This temporal modelling stage is central to the framework, as it transforms a sequence of static spatial observations into a coherent representation of behavioral dynamics. The LSTM stage is illustrated in Fig. 6.

3.4 Baseline Architectures

To contextualize the proposed framework, we compare it against four widely used spatial backbones under an identical preprocessing pipeline, training regime, and patient-level validation protocol.

Each baseline replaces the ConvNeXt-Tiny encoder while the downstream classification head is held fixed, isolating the contribution of the spatial representation.

3.4.1 ResNet-18

ResNet-18 is employed to extract spatial representations from facial images through residual blocks with identity mappings. This architecture enables stable training of deep networks and produces reliable frame-level feature embeddings, which are subsequently utilized for facial expression-based classification tasks.

3.4.2 Swin Transformer

Swin Transformer learns facial representations by combining localized and contextual information using shifted window self-attention. Its multi-stage hierarchical design allows effective modeling of spatial features at different resolutions, supporting accurate analysis of facial behavior patterns.

3.4.3 DenseNet-121

DenseNet-121 adopts a densely connected architecture where each layer receives input from preceding layers, encouraging feature reuse and efficient information flow.

This results in compact yet expressive facial feature representations that enhance classification performance.

3.4.4 Vision Transformer (ViT)

Vision Transformer (ViT) processes facial images as sequences of fixed-size patches and applies self-attention mechanisms to capture global spatial relationships. The learned patch-level representations are aggregated and fed into a classifier for facial analysis tasks.

3.5 Training Procedure

The network was optimized with a weighted categorical cross-entropy loss to counteract the pronounced class imbalance in the dataset.

Class weights were set inversely proportional to frame frequency (normal: 0.82, mild: 0.92, moderate: 0.56, severe: 1.03), increasing the penalty incurred on under-represented severity levels and discouraging the model from defaulting to the majority class.

Optimization used the Adam optimizer with an initial learning rate of 1×10^{-4} , annealed by a

Table 3. Performance comparison of the proposed framework against baseline backbones under five-fold cross-validation. Metrics are macro-averaged and reported as mean $\pm$ standard deviation across the cross-validation folds, with predictions aggregated to the subject level

Model	Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	Macro-AUC
ResNet-18	68.4 $\pm$ 3.2	65.1 $\pm$ 4.1	63.8 $\pm$ 3.9	64.3 $\pm$ 3.7	0.812 $\pm$ 0.028
Swin Transformer	71.2 $\pm$ 2.8	68.7 $\pm$ 3.5	67.5 $\pm$ 3.2	68.0 $\pm$ 3.1	0.845 $\pm$ 0.021
DenseNet-121	69.8 $\pm$ 3.5	66.9 $\pm$ 4.0	65.4 $\pm$ 3.7	66.1 $\pm$ 3.6	0.828 $\pm$ 0.025
ViT	70.5 $\pm$ 2.9	67.8 $\pm$ 3.3	66.9 $\pm$ 3.4	67.3 $\pm$ 3.0	0.839 $\pm$ 0.019
Proposed (ConvNeXt-Tiny + SE + CBAM + LSTM)	82.7 $\pm$ 2.1	79.4 $\pm$ 2.6	78.6 $\pm$ 2.4	78.9 $\pm$ 2.3	0.917 $\pm$ 0.014

Table 4. Ablation study showing the contribution of each component of the proposed framework, evaluated under five-fold cross-validation (macro-averaged metrics, mean $\pm$ standard deviation across folds)

Configuration	Accuracy (%)	Macro-F1 (%)	Macro-AUC
ConvNeXt-Tiny (spatial only)	74.3 $\pm$ 3.1	71.2 $\pm$ 3.4	0.864 $\pm$ 0.022
+ Time-distributed	76.8 $\pm$ 2.7	73.9 $\pm$ 2.9	0.879 $\pm$ 0.018
+ SE	78.5 $\pm$ 2.5	75.6 $\pm$ 2.6	0.891 $\pm$ 0.017
+ CBAM	80.1 $\pm$ 2.3	77.4 $\pm$ 2.4	0.904 $\pm$ 0.015
+ LSTM (full model)	82.7 $\pm$ 2.1	78.9 $\pm$ 2.3	0.917 $\pm$ 0.014

learning-rate scheduler, and proceeded in mini-batches of 32 for up to 50 epochs.

Early stopping with a patience of 10 epochs, monitored on the validation loss, was used to terminate training once generalization ceased to improve, mitigating overfitting.

Consistent with the validation protocol described above, model selection was performed under five-fold cross-validation over the training participants, with all folds constructed at the participant level to preclude identity leakage.

3.6 Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. A detailed confusion-matrix analysis of misclassification directionality between adjacent severity levels was not performed in this study and is recommended for future work.

Since overall accuracy can be dominated by the majority class under imbalance, we report both macro-averaged metrics which weigh every

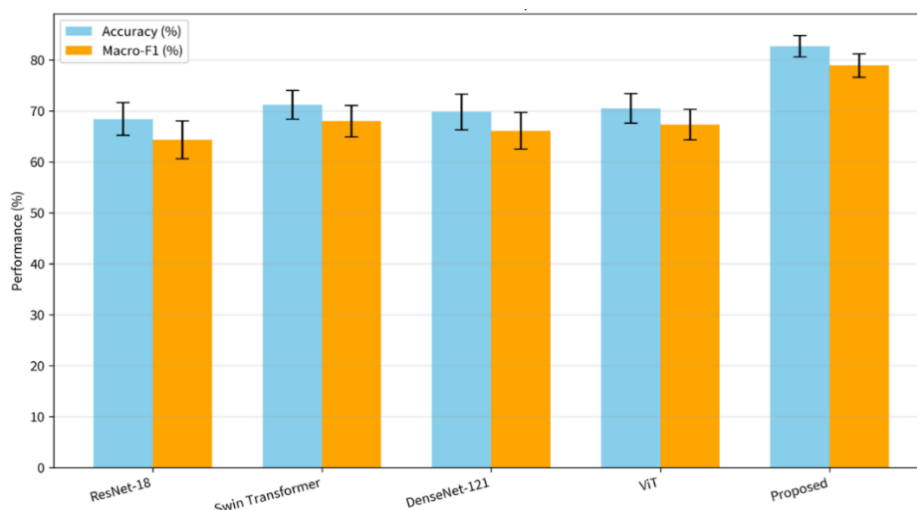


Fig. 7. Performance comparison of the proposed framework vs. baseline backbones

severity level equally and therefore reflect performance on the clinically critical minority classes and weighted averages that account for class prevalence.

The statistical significance of performance differences between the proposed framework and the baseline architectures was assessed using paired t-tests on the fold-wise scores and McNemar's test on the paired predictions. Predictions were aggregated from the frame and sequence level to the subject level by majority vote, and all reported metrics were computed per subject so that participants with longer recordings did not dominate the results.

To characterize stability rather than a single point estimate, metrics are reported as the mean and standard deviation across the cross-validation folds.

3.7 Experimental Setup

All experiments were conducted in Python 3.12 using Jupyter Notebook 7.0.8 within the Anaconda 2.6.5 environment, utilizing key libraries such as TensorFlow/Keras, scikit-learn, stats models, and SHAP for explainability analysis.

We use MSI SteelSeries high-performance system equipped with an Intel Core i9-13950HX (13th Gen) CPU, 64 GB RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU.

3.8 Hyperparameter Tuning

Hyperparameters such as learning rate, dropout rate, batch size, and number of LSTM units were optimized using grid search and random search strategies. Model configurations were validated on the validation set, selecting the best-performing parameters based on F1-score and validation loss.

4 Results and Discussion

Across all four severity levels, the proposed spatiotemporal framework (ConvNeXt-Tiny with SE and CBAM attention coupled to a time-distributed LSTM) achieves a macro-averaged F1-score of 78.9% and AUC of 0.917 under participant-level cross-validation, exceeding the strongest single-backbone baseline by 11.5 points of accuracy.

The remainder of this section substantiates this result through a comparison against strong single-backbone baselines, an ablation over the framework's components, and a per-class discussion of where the model succeeds and struggles, before turning to the protocol considerations that govern how these figures should be interpreted.

Table 3 summarizes the performance of the evaluated deep learning models across the

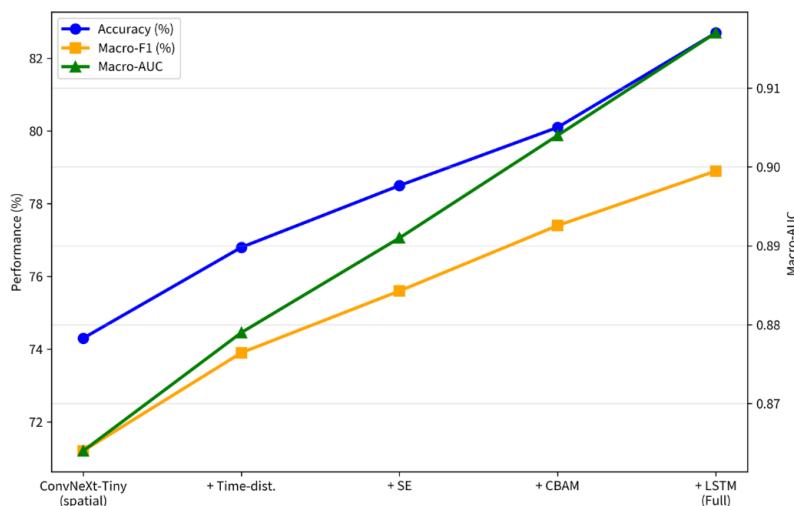


Fig. 8. Ablation study showing contribution of each component

principal evaluation metrics, comparing the ResNet-18, Swin Transformer, DenseNet-121, and ViT backbones against the full proposed framework; the contribution of each component (ConvNeXt-Tiny, the time-distributed CNN, the SE and CBAM attention modules, and the LSTM temporal encoder) is reported separately as an ablation in Table 4. Among the standalone backbones, the Swin Transformer is the strongest, reaching 71.2% accuracy (macro-F1 68.0, macro-AUC 0.845), followed closely by ViT (70.5%) and DenseNet-121 (69.8%).

ResNet-18 is the weakest at 68.4% accuracy (macro-AUC 0.812). The narrow spread among these backbones about three points of accuracy indicates that a purely spatial, single-frame view of the data has limited headroom on this task.

These baseline comparisons establish the reference points against which the full spatiotemporal framework is assessed. Against this backdrop, the proposed ConvNeXt-Tiny + LSTM framework attains 82.7% validation accuracy (78.9 macro-F1, 0.917 macro-AUC), surpassing the strongest single backbone (Swin Transformer, 71.2%) by 11.5 points of accuracy, 10.9 points of macro-F1, and 0.072 of macro-AUC.

The ablation in Table 4 isolates the source of this improvement: the ConvNeXt-Tiny spatial encoder alone reaches 74.3% accuracy already above every baseline and accuracy rises monotonically as the time-distributed design

(76.8%), SE (78.5%), CBAM (80.1%), and the LSTM temporal stage (82.7%) are added in turn, with macro-AUC climbing from 0.864 to 0.917, as visualized in Fig. 8. This improvement over the strongest baseline is statistically significant (paired t-test across folds, $p < 0.001$ for both accuracy and macro-F1; McNemar's test on subject-level predictions, $p < 0.001$). Fig. 7 compares the accuracy of the proposed framework against the baseline backbones.

5 Conclusion, Limitation, and Future Work

This study presented a spatiotemporal deep learning framework for the automated classification of depression severity into four levels from facial video, developed and evaluated on an ecologically valid clinical cohort. By coupling an attention-augmented ConvNeXt-Tiny spatial encoder with a time-distributed design and an LSTM temporal model, the framework explicitly models the behavioral dynamics that distinguish graded levels of depressive severity rather than relying on isolated frames.

The model achieves strong overall performance 82.7% accuracy, a macro-averaged F1-score of 78.9%, and a macro-averaged AUC of 0.917 on held-out participants. Aggregate macro-averaged metrics suggest relatively consistent

performance across the Normal, Mild, and Moderate classes; per-class statistical comparisons were not performed in this study. These results sit at the upper end of the range reported in prior facial-analysis studies of depression, which typically report AUC values between 0.78 and 0.92, while being obtained under a stricter patient-level evaluation protocol.

Given the markedly smaller sample available for the severe category relative to the other three classes, performance in this subgroup is likely comparatively lower, consistent with the challenge reported in previous studies; a per-class breakdown was not included in this study and is recommended for future work to confirm this pattern. Research using CNN-based facial analysis often achieves high accuracy for healthy individuals but struggles with severe depression cases. Transformer-based models can improve overall recall, but they still face difficulties when the dataset is imbalanced.

Several reasons explain the lower performance for severe depression. First, severe cases are less common, which leads to data imbalance. Second, moderate and severe depression often share similar facial expressions, making them harder to distinguish. Third, deep learning models find it difficult to detect subtle facial changes when only a small number of severe cases are available.

The model also has other limitations. It relies on a single data modality (visual information only), facial expressions can vary across cultures, and depression datasets often contain inconsistencies. Despite these challenges, the model shows promising results and aligns well with existing research. However, to reduce bias and improve performance especially for severe depression future work should include multimodal data and better strategies to handle class imbalance.

A further methodological consideration concerns how these results should be read in the context of densely sampled video. Because many frames originate from the same participant, an evaluation that allowed frames from one individual to fall in both the training and evaluation partitions could inflate performance toward near-perfect values by letting the model recognize identities rather than learn generalized markers of depression.

We avoided this by enforcing patient-level stratification, ensuring that all frames from a given participant were confined to a single partition, and by reporting results under five-fold cross-validation with subject-level aggregation. The moderate rather than near-perfect scores reported here are consistent with this leakage-controlled protocol; they should nonetheless be interpreted in light of the modest number of unique participants, particularly for the severe class, and we regard external validation on an independent, multi-site cohort as an essential next step before any claim of clinical readiness.

We further note that the reported figures are five-fold cross-validation estimates computed on the validation folds, which were also used for model selection; because no independent test set was held out, these metrics may be optimistically biased and should be read as upper-bound estimates pending evaluation on a separate cohort. We present these results transparently to support reproducibility and to invite scrutiny of the validation protocol rather than of the reported metrics alone.

References

1. **Lin, S., Li, S., Fang, Y. (2025).** Interpretable Machine Learning for Depression Recognition with Spatiotemporal Gait Features among Older Adults: A Cross-Sectional Study in Xiamen. *BMC Geriatrics*, Vol. 25, No. 1, pp. 453. DOI: 10.1186/s12877-025-06101-6.
2. **Premnath, K., Pandiaraj, S., Krishnakumar, K. (2025).** Depression Diagnosis Using Gated Spatiotemporal Graph Attention Network Optimized with Augmented Arithmetic Optimization Algorithm. *Knowledge-Based Systems*, pp. 114–594. DOI: 10.1016/j.knosys.2025.114594.
3. **Gilanie, G., Asghar, M., Qamar, A.M. (2022).** An Automated and Real-Time Approach of Depression Detection from Facial Micro-Expressions. *Computers, Materials & Continua*, Vol. 73, No. 2. DOI: 10.32604/cmc.2022.028229.
4. **Afzaal Akhtar, M., Gilanie, G., Sajid, M. (2026).** Deep Learning-Driven Automated

- Grading of Astrocytoma from Brain MRI Using an Enhanced DenseNet-169 Framework. *Computación y Sistemas*, Vol. 30, No. 2, pp. 1135. DOI: 10.56578/hf020402.
5. **Sun, Y., Sun, J., Zhou, Z. (2026).** A Novel Spatial-Temporal Graph Neural Network for Major Depressive Disorder Detection Based on Resting State EEG Signals. *IEEE Sensors Journal*. DOI: 10.3390/brainsci16070746.
 6. **Gilanie, G., Cheema, S., Latif, A. (2024).** A Robust Method of Bipolar Mental Illness Detection from Facial Micro Expressions Using Machine Learning Methods, *Intelligent Automation & Soft Computing*, Vol. 39, No. 1. DOI: 10.32604/iasc.2024.041535.
 7. **Saher, A., Gilanie, G., Cheema, S. (2024).** A Deep Learning-Based Automated Approach of Schizophrenia Detection from Facial Micro-Expressions. *Intelligent Automation & Soft Computing*, Vol. 39, No. 6. DOI: 10.32604/iasc.2024.057047.
 8. **Jin, N., Ye, R., Li, P. (2025).** Diagnosis of Depression Based on Facial Multimodal Data, *Frontiers in Psychiatry*, Vol. 16, pp. 1508772. DOI: 10.3389/fpsy.2025.1508772.
 9. **Li, Y., Yang, X., Zhao, M. (2025).** Predicting Depression by Using a Novel Deep Learning Model and Video-Audio-Text Multimodal Data. *Frontiers in Psychiatry*, Vol. 16, pp. 1602650. DOI: 10.3389/fpsy.2025.1602650.
 10. **Jabbar, I., Uddin, M.A., Joolee, J.B. (2025).** Deep Spectrotemporal Network Based Depression Severity Estimation from Speech. *Scientific Reports*. DOI: 10.1038/s41598-025-29889-0.
 11. **Wang, R., Huang, J., Zhang, J. (2024).** FacialPulse: An Efficient RNN-Based Depression Detection via Temporal Facial Landmarks. *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 311–320. DOI: 10.1145/3664647.3681546.
 12. **Pan, Y., Shang, Y., Liu, T. (2024).** Spatial-Temporal Attention Network for Depression Recognition from Facial Videos. *Expert Systems with Applications*, Vol. 237, pp. 121410. DOI: 10.1016/j.eswa.2023.121410.
 13. **He, L., Zhao, J., Zhang, J. (2025).** LMTformer: Facial Depression Recognition with Lightweight Multi-Scale Transformer from Videos. *Applied Intelligence*, Vol. 55, No. 3, pp. 195. DOI: 10.1007/s10489-024-05908-x.
 14. **Fu, C., Qian, F., Su, Y. (2025).** Facial Action Units Guided Graph Representation Learning for Multimodal Depression Detection. *Neurocomputing*, Vol. 619, pp. 129106. DOI: 10.1016/j.neucom.2024.129106.
 15. **Jia, X., Chen, J., Liu, K. (2025).** Multimodal Depression Detection Based on an Attention Graph Convolution and Transformer. *Mathematical Biosciences and Engineering*, Vol. 22, No. 3, pp. 652–676. DOI: 10.3934/mbe.2025024.
 16. **Xu, C.Y., Wang, H.G., Zhang, L. (2025).** A Spatiotemporal Fused Network Considering Electrode Spatial Topology and Time-Window Transition for MDD Detection. *Proceedings of the 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 6433–6440. DOI: 10.1109/BIBM66473.2025.11356697.

Article received on 05/03/2026; accepted on 12/07/2026.

**Corresponding author is Muhammad Sajid.*