Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

A Motion Aware Transformer based Latent Modeling Framework for Pain Induced Mental Stress Prediction

Soumalya De

Department of Information Technology, Techno International New Town, Kolkata, India,

Saiyed Umer

Department of Computer Science and Engineering, Aliah University, Kolkata, India,

Abstract - Mental health conditions are often accompanied by physiological stress responses that are expressed through involuntary facial dynamics. Automated analysis of such facial motion patterns offers a promising, non-invasive pathway for continu-ous mental health assessment. In this work, we propose a latent variable model-ing framework for mental health–oriented prediction that uses pain-induced facial expressions as a monotonic proxy for stress-related affective states. The frame-work explicitly models facial motion using dense optical flow and spatiotemporal motion features, enabling the capture of fine-grained temporal variations that are difficult to infer from static appearance alone. To effectively learn long-range tem-poral dependencies in facial motion sequences, we introduce a Transformer-based temporal encoder driven by self-attention mechanisms. Given short facial video sequences, optical flow magnitude maps are extracted between consecutive frames and structured as temporal motion representations, which are then processed by the Transformer to emphasize psychologically salient motion patterns. The pro-posed model is evaluated on the BioVid heat pain dataset and compared against a motion-based baseline and multiple deep spatiotemporal learning architectures. Experimental results demonstrate that self-attention based temporal modeling of facial motion leads to consistent performance improvements, underscoring the relevance of explicit motion dynamics for mental health–related inference. This study highlights the potential of motion-aware, attention-driven frameworks for pain-induced mental health assessment.

Keywords: **Digital health-informatics, Monotonic Latent Mapping, Multi-algorithm, Transformer, Optical flow**

1. Introduction

Mental health disorders is one of the major global health issue which greatly increase disability, worse quality of life, and socioeconomic burden globally [1]. Due to their subjective character, societal stigma, and reliance on self-reported as-sessments, conditions like depression, anxiety, and stress-related illnesses are fre-quently underdiagnosed. Due to these constraints, interest in digital informatics systems that employ objective behavioral and physiological cues for early men-tal health monitoring and prediction has grown [2]. Due to similar neurobiolog-ical and affective processes, there is a strong relation between pain perception and mental health. Emotions like anxiety, depression, and stress-related disor-ders are linked to chronic pain responses [3] [4]. According to neuroimaging and psychophysiological research [5], brain areas like the anterior cingulate cor-tex, insula, and prefrontal cortex are involved simultaneously in pain processing and emotion recognition, which forms a core component for diagnosis of mental health state. Thus, facial dynamics and other pain-induced behavioral expressions present a feasible pathway for indirect mental health inference. Recent advance-ment in computer vision and affective computing have enabled automatic analysis of facial expressions for emotion, stress, and mental state recognition [6]. Among these approaches, optical flow based temporal modeling has gained attention due to its ability to capture precise motion patterns that static image features often fail to represent [7] [8]. Temporal facial feature, such as muscle activation speed, asymmetry, and persistence, are particularly relevant for identifying affective re-sponses linked to pain and emotional distress.

The BioVid Heat Pain Dataset provides a controlled experimental framework for studying physiological and behavioral responses [9] that triggers due to pain stimuli. The dataset includes synchronized facial video recordings and five dis-crete pain intensity levels, ranging from baseline (BL1) to high pain. While BioVid does not contain explicit mental health labels, prior research has estab-lished that graded pain responses correlate strongly with stress, anxiety, and emo-tional regulation mechanisms [10]. This makes BioVid particularly suitable for modeling mental health–relevant affective states through pain intensity progres-sion. Existing studies using BioVid have primarily focused on pain intensity recognition through handcrafted features, deep learning–based physiological anal-ysis, or multimodal fusion [11]. However, limited attention has been given to the temporal evolution of facial motion as a standalone indicator for mental health pre-diction. Furthermore, most pain recognition systems treat each frame neglecting long-range temporal dependencies that are crucial for modeling emotional escalation, sustained distress which are the key indicators of mental health vulnerability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

To address these gaps, this work proposes a digital informatics system for mental health prediction that exploits temporal information encoded in optical flow derived from facial videos. By modeling motion trajectories across five pain levels, the system captures dynamic facial responses that reflect stress intensity, emotional regulation capacity, and affective transitions. Rather than aiming for clinical diagnosis, the proposed framework focuses on mental health–relevant state prediction inferred from pain-induced facial expression. Thus, this study contributes in the formulation of pain-level–driven facial motion analysis as a proxy for mental health prediction; the systematic exploitation of optical flow–based temporal features to capture fine-grained affective dynamics; and the development of a digital informatics system suitable for real-world mental health monitoring by bridging affective computing, pain analysis, and digital mental health informatics. The flow of the paper is as follows - section 2 discusses about the existing mod-els and related work, section 3 describes the datasets that we have used for our experiments, section 4 illustrates the methodology, section 5 provides an insights about how we have performed the experiment, section 6 discusses the results and findings that we have obtained, followed by the conclusion in section 7.

2. Related Work and Contribution

Recent advances in digital mental health informatics have increasingly em-phasized objective, data-driven approaches that uses behavioral, physiological, and visual signals to overcome the limitations of self-reported assessments. It has shown a clear transition toward multimodal and temporal modeling frameworks, particularly for stress, pain, and affective state recognition, which are closely linked to mental health outcomes.

2.1. Multimodal Stress and Mental Health State Recognition

Multimodal learning has become a dominant paradigm for mental health–related affect recognition. Recent study proposes a large-scale multimodal stress detec-tion dataset integrating facial expressions with physiological signals such as heart rate variability and electrodermal activity, demonstrating that multimodal fusion significantly outperforms unimodal approaches for stress detection and emotional state classification [12]. Systematic reviews published during this period further highlight the increasing clinical relevance of automated emotion recognition sys-tems. These reviews emphasize the role of facial dynamics and temporal infor-mation in identifying emotional states linked to anxiety, depression, and stress [13].

2.2. Pain Recognition and the BioVid Dataset

Pain recognition research has advanced significantly with the adoption of deep learning models. Several studies between 2021 and 2025 have used the BioVid Heat Pain Dataset to estimate pain intensity from facial videos and physiolog-ical signals. A recent 2025 study systematically evaluated deep convolutional and transformer-based models for facial-only pain recognition on BioVid, establishing new performance benchmarks and confirming that facial motion pat-terns alone carry strong discriminative information across graded pain levels [14]. Multimodal approaches combining ECG, EDA, and facial expressions have also been explored, showing improved performance through temporal fusion strategies [15]. However, these studies primarily focus on pain estimation rather than mental health inference and often underexploit ingrainedtemporal facial motion cues.

2.3. Temporal Modeling and Optical Flow in Affective Computing

Temporal modeling has gained traction as a means to capture affective dynam-ics rather than static expressions. Optical flow-based representations have proven particularly effective in encoding subtle facial muscle movements and micro-expressions. Recent work proposed optical flow-based “driven hierarchical deep learning architectures for psychological state prediction, demonstrating that motion-based facial representations significantly improve the classification of mental and emo-tional states compared to static image features [16]. In parallel, researchers have explored visual encodings of temporal physiological signals such as converting time-series data into image representations to enhance stress and emotion classification further underscoring the value of temporal dynamics in affective comput-ing [17]. A comparative study on recent work on Biovid dataset is explored in Table 1.

However, the contribution of the existing work not only comply us to meet our research goal but to motivate us in proposing a novel temporal optical flow-based architecture for mental health prediction to address the conceptual and method-ological gap. This work addresses an identified gap by using temporal optical flow-based facial motion analysis to model graded pain responses from the BioVid dataset as proxies for mental health-relevant affective states within a digital infor-matics framework.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Table 1: Existing work on Biovid Dataset

Study (Year)	Dataset	Facial Representation	Key Limitation
Kächele et al. (2021)	BioVid	CNN-based spatial features	Motion dynamics underexplored
Werner et al. (2022)	BioVid	Handcrafted + CNN features	Limited long-range temporal modeling
Thiam et al. (2023) [15]	BioVid	Deep visual embeddings	Focused on pain, not mental health
Alshamsi et al. (2024) [13]	Multiple datasets (Review)	Facial expressions	Identifies lack of temporal facial modeling
Li et al. (2025) [16]	Mental health video dataset	Optical flow + deep networks	Not validated on pain datasets
Kächele et al. (2025) [14]	BioVid	CNN / Transformer spatial features	Static bias in facial modeling

3. Dataset

3.1. Biovid Heat Pain Dataset

In this study, we employ the BioVid Heat Pain Database [18], a widely recog-nized benchmark for automatic pain recognition research. The dataset was devel-oped jointly by the University of Ulm (Medical Psychology) and the University of Magdeburg (Neuro-Information Technology) with the objective of providing a controlled, multimodal resource for modeling and analyzing human pain re-sponses. It comprises recordings from approximately 90 healthy adult participants (aged 20–65 years). Heat pain was induced using a thermode applied to the inner forearm. For each subject, four distinct pain intensity levels (PA1-PA4) were in-dividually calibrated between their pain threshold and tolerance, in addition to a baseline (BL1) no pain condition. Each condition was repeated 20 times, result-ing in highly controlled and balanced experimental data. Thus, each participant contributes 100 samples (5 conditions × 20 trials), yielding approximately 8,700 video instances.

Although the BioVid Heat Pain Database was originally designed for auto-matic pain recognition, its multimodal structure makes it highly relevant for men-tal health research. There is a strong correlation between pain and mental health since stress, worry, and depression are known to change how pain is perceived and to cause different behavioral and physiological reactions. In stress and emotional computing research, the dataset offers synchronized recordings of electromyogra-phy (EMG), galvanic skin response (GSR), electrocardiogram (ECG), and facial expressions all of which are recognized biomarkers. A prominent correlation be-tween Biovid modalities and its relevance to mental health prediction is illustrated in Table 2.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Table 2: Mapping BioVid modalities to their relevance in mental health prediction.

Modality (BioVid)	Extracted Features	Mental Health Relevance
Facial Video	Micro-expressions, facial action units (e.g., brow furrow, lip press)	Negative affect recognition, depression symptom detection, anxiety-related facial tension
ECG	Heart rate (HR), heart rate variability (HRV: LF/HF ratio, RMSSD, pNN50)	Stress and anxiety biomarkers; dysregulation in depression
GSR (EDA)	Skin conductance level (SCL), phasic peaks, rise time	Autonomic arousal indicator; heightened responses linked to stress and anxiety
EMG (Facial Muscles)	Muscle activation intensity, frequency bands	Muscle tension associated with stress; reduced expressiveness linked to depression
Pain Intensity Labels (BL1,PA1–PA4)	Stimulus-based affective states	Can be reinterpreted as proxy levels for stress/negative affect burden

4. Proposed Methodology

This study proposes a motion-aware spatiotemporal deep learning framework for mental health state prediction by utilizing pain-induced facial dynamics from the BioVid Heat Pain dataset. The central hypothesis is that temporal facial mo-tion patterns corresponding to different pain intensities are indicative of underly-ing mental stress and affective states, and can be effectively modeled using op-

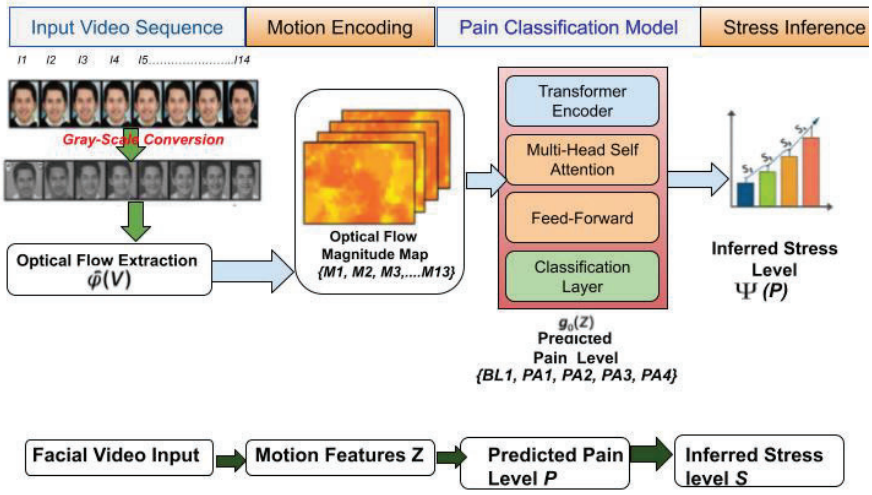


Figure 1: Proposed Framework.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

tical flow–based representations combined with modern temporal learning archi-tectures. Figure 1 represents the proposed framework for stress detection.

4.1. Problem Formulation

Let $V = \{I_1, I_2, \ldots, I_{14}\}$ denote a facial image sequence of 14 consecutive frames extracted from a video segment corresponding to a specific pain stimu-lus. Each sequence is associated with a discrete pain level label

$$y \in \{\mathrm{BL}_1, \mathrm{PA}_1, \mathrm{PA}_2, \mathrm{PA}_3, \mathrm{PA}_4\},$$

where increasing pain levels are treated as indicators for escalating mental stress intensity. The objective is to learn a function $f : V \to y$. More specifically, the input video sequences V are collection of several image frames I_n, mathematically can be written as $V = \{I_1, I_2, \ldots, I_{14}\}$. Before classification, V is transformed into a motion representation:

$$\phi(V) \;=\; \{M_1, M_2, \ldots, M_{13}\}$$

where M_t is the optical flow magnitude map between frames I_t and I_{t+1} and Op-tical flow is computed using Farnebäck's dense optical flow algorithm. Thus, $f : V \to y$ can be written as $f(V) = g(\phi(V))$, where $g()$ is learned from data across different experiments.

4.2. Motion Representation Using Dense Optical Flow

To explicitly encode facial dynamics, dense optical flow is computed between consecutive frames using the Farnebäck algorithm. Given two adjacent grayscale

frames I_t and I_{t+1} dense optical flow estimates a motion field: $F_t = \big(u_t(x,y),\, v_t(x,y)\big)$, where u_t and v_t represent the horizontal and vertical displacement vectors at pixel location (x,y). The optical flow magnitude is computed as:

$$M_t(x,y) = \sqrt{u_t(x,y)^2 + v_t(x,y)^2} \text{ ----- [8]}$$

For each 14-frame sequence, this yields 13 optical flow magnitude maps, cap-turing the intensity and distribution of facial muscle movements over time. This motion-centric representation is critical, as pain expressions are dominated by subtle non-rigid facial movements that are poorly captured by static appearance features alone.

4.3. Spatiotemporal Feature Construction

Depending on the experimental configuration, the optical flow information is represented in two forms: 1. The mean optical flow magnitude is computed for each frame pair, producing a compact 13-dimensional temporal motion vector can be called as Statistical Motion Descriptor (Baseline). 2. Full Spatiotemporal Mo-tion Tensor where optical flow magnitude maps are stacked temporally to form a 5D tensor: $X \in \mathbb{R}^{N\times 13\times H\times W\times 1}$ preserving both spatial motion patterns and tem-poral evolution. This second representation forms the foundation of the proposed deep learning models.

4.4. Deep Learning Architectures

To systematically study the role of motion and temporal modeling, multiple architectures were implemented.The deep-learning architecture is represented in Figure 2.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

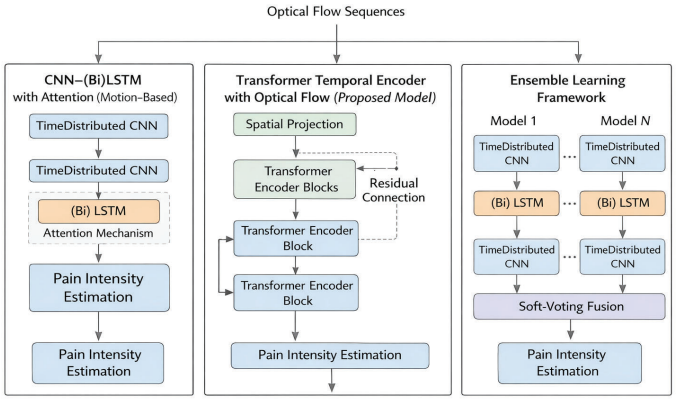


Figure 2: Deep Learning Architecture for Pain Estimation.

4.4.1. CNN-(Bi)LSTM with Attention (Motion-Based)

TimeDistributed CNN layers extract frame-wise spatial motion features from optical flow maps. These features are passed to a Bidirectional LSTM, enabling learning of temporal dependencies in both forward and backward directions. An attention mechanism assigns higher importance to frames exhibiting peak pain expressions, producing a weighted temporal representation for classification.

4.4.2. Transformer Temporal Encoder with Optical Flow (Proposed Model)

The core contribution of this work is a Transformer-based temporal encoder that replaces recurrent units with multi-head self-attention. After spatial projec-tion, the optical flow sequence embeddings are processed by Transformer encoder blocks consisting of Multi-head self-attention, Feed-forward networks, Residual connections and layer normalization. This design allows the model to capture long-range temporal dependencies and global motion interactions, which are cru-cial for modeling gradual and non-linear pain progression patterns.

4.4.3. Ensemble Learning Framework

To enhance robustness, multiple independently trained CNN–(Bi)LSTM mod-els are combined using probability-level (soft-voting) fusion. Final predictions are obtained by averaging class probabilities across models, reducing variance and improving class-wise stability.

4.5. Mental Health Interpretation

In this study, pain intensity levels are treated as an observable quantized indi-cation of latent mental stress. Under controlled experimental conditions, increas-ing pain induces monotonic escalation in affective and cognitive stress responses, which are reflected in facial dynamics. All implemented models learn to discrim-inate pain classes based on facial dynamics, after which a monotonic mapping is applied to infer discrete stress levels. This formulation enables mental health oriented interpretation while avoiding direct clinical stress diagnosis. In this ap-proach, we introduced Latent Variable Modeling (LVM), a statistical approach that uses unobserved (latent) variables to explain patterns and correlations among a set of directly measured (observed) variables, helping quantify mental-health state from measurable pain indicators. In this approach, we define stress as a latent variable as it cannot be observed directly, while pain as an observable or-dinal variable. So mathematically, we can denote stress $S \in \mathbb{R}^{+}$ and pain level $P \in \{0, 1, 2, 3, 4\}$, where 0 = BL1 (no pain), 4 = PA4 (maximum pain). We assume:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

$$P = Q(S+N)+\varepsilon \tag{1}$$

where $Q()$ is a quantization function, ε denotes measurement noise, and N captures input-level perturbations and measurement uncertainty.
Under controlled stimulus conditions (BioVid applies fixed thermal stimuli), $N = 0$ so, (1) can be reduced to:

$$P = Q(S)+\varepsilon, \tag{2}$$

This makes pain a quantized observation of latent stress and corresponds to the baseline (no nociceptive stimulation) condition. To establish a principled relationship between pain intensity and stress level, we define a monotonic mapping function $\psi : P \to S_d$, where P denotes the discrete pain intensity and S_d represents the corresponding discrete stress intensity. The mapping is formally expressed as

$$S_d = \psi(P) = \alpha P + \beta, \tag{3}$$

where $\alpha > 0$ is a scaling factor controlling the sensitivity of stress variation with respect to pain intensity, and $\beta \geq 0$ represents a baseline stress offset. The monotonicity constraint ensures that higher pain levels induce proportionally higher stress responses, satisfying the condition

$$P_i > P_j \;\Rightarrow\; \psi(P_i) > \psi(P_j). \tag{4}$$

This formulation is consistent with psychophysiological findings that pain perception and stress activation exhibit a positive correlated relationship, thereby providing a mathematically sound and interpretable basis for modeling stress as a latent mental-health state inferred from pain intensity.

4.6. Functional Composition in the Experiments

The proposed experiments implement the following composed functional mapping:

$$V \xrightarrow{\phi} Z \xrightarrow{g_\theta} \hat{P} \xrightarrow{\psi} \hat{S}, \tag{5}$$

where V denotes the input facial video sequence, Z represents the encoded motion features, $\hat{P}$ is the predicted pain intensity, and $\hat{S}$ denotes the inferred stress intensity.

4.6.1. Motion Encoding

The motion encoding function $\phi(\cdot)$ extracts stress-induced facial dynamics using optical flow:

$$\phi(V) = \{\text{OpticalFlow}(I_t, I_{t+1})\}_{t=1}^{13}, \tag{6}$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

where I_t and I_{t+1} denote consecutive video frames. This representation captures subtle facial motor activity associated with stress responses.

4.6.2. Learned Pain Classifier

The encoded motion features are mapped to pain intensity through a learned classifier:

$$\hat{P} = g_\theta\big(\phi(V)\big), \tag{7}$$

where g_θ denotes a parameterized deep learning model, such as a CNN–LSTM or transformer architecture with optical flow. The model learns the functional relationship

$$P \approx f(\text{facial stress dynamics}), \tag{8}$$

motivated by the fact that facial muscle tension and micro-movements serve as reliable biomarkers of stress.

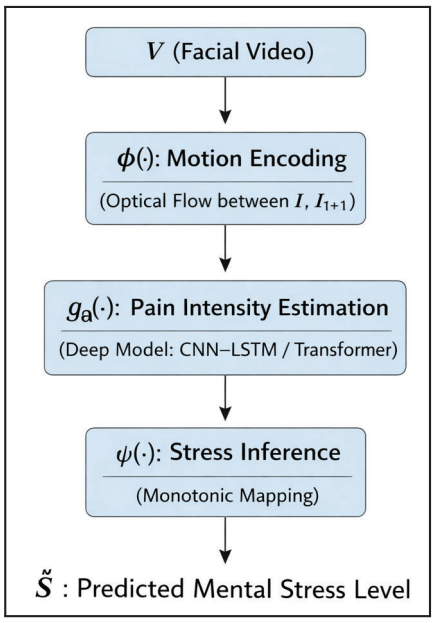


Figure 3: Functional Components for Stress Mapping.

4.6.3. Stress Inference Function

Finally, the predicted pain intensity is mapped to latent mental stress intensity via a monotonic inference function:

$$\hat{S} = \psi(\hat{P}), \tag{9}$$

which converts observable behavioral stress indicator into an estimate of internal mental stress intensity.

Thus, the overall mental stress prediction is formulated as a composition of three functional modules, given by

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

$$\hat{S} = \psi\big(g_\theta(\phi(V))\big)$$

where $\phi()$ extracts stress-induced motion features, $g_\theta()$ estimates pain intensity, and $\psi()$ maps pain intensity to the corresponding mental stress-level. Figure 3 illustrates the model encoding functions.

5. Experiment

Experiments are conducted on the BioVid Heat Pain dataset, which comprises facial video recordings collected under controlled thermal pain stimulation condi-tions. The dataset includes five discrete pain levels, namely a baseline condition (BL1) and four progressively increasing pain intensities (PA1–PA4). All exper-iments are performed on a machine equipped with an NVIDIA GPU NVIDIA GPU (8 GB Virtual RAM), an Intel Core i7 CPU, and $\geq$ 16 GB of system RAM. The models are implemented using Python with deep learning frameworks such as PyTorch and TensorFlow, and CUDA-enabled GPU. The video segments contain-ing exactly 14 consecutive frames are retained for experiment to ensure tarining consistency and fair comparison across all models..

5.1. Preprocessing

To perform the experiment, each video sample is represented as a fixed-length sequence of 14 consecutive facial frames. Temporal ordering of frames is strictly maintained to ensure accurate motion estimation. Prior to feature extraction, all frames undergo standardized preprocessing. First, frames are converted to grayscale to reduce computational complexity while preserving essential motion indicators. Then, the frames are cropped and resized to a fixed spatial resolution of either 128×128 or 32×32, depending on the architectural requirements of the model.

5.2. Feature Extraction

Pain-related expressions are often indicated through micro and transient fa-cial movements. To capture these fine-grained temporal variations, the proposed framework uses motion-based feature extraction strategy centered on dense optical flow analysis. After preprocessing the images, Dense optical flow is computed betweeen consecutive frames using the Farnebäck algorithm. For two successive frames I_t and I_{t+1}, the optical flow algorithm estimates a dense displacement field that encodes pixel-wise motion in both horizontal and vertical directions. For each 14-frame sequence, this results in 13 optical flow maps, corresponding to motion between adjacent frame pairs.

From the estimated optical flow vectors, the horizontal and vertical compo-nents are transformed into motion magnitude maps, which quantify the intensity of movement at each pixel location. This transformation suppresses directional variability while emphasizing motion strength, making it more robust to subject-specific facial structure differences. Two complementary feature representations are derived from these motion magnitude maps:

Statistical Motion Features: In the baseline configuration, the mean optical flow magnitude is computed for each frame pair, resulting in a compact 13 dimen-sional temporal motion vector per sample. This representation summarizes the overall evolution of facial motion intensity across time and serves as a lightweight handcrafted feature descriptor. The average pixel intensity between two consecu-tive frames is represented in Figure 4.

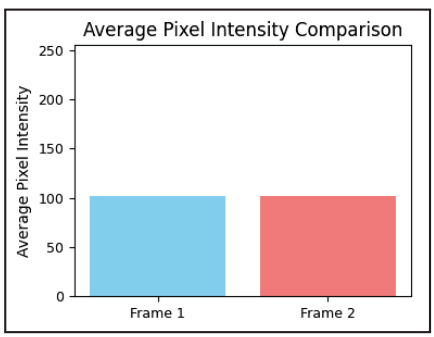


Figure 4: Average Pixel Intensity for BL1 between Frame 1 and Frame 2.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Spatiotemporal Motion Features: For deep spatiotemporal models, the full motion magnitude maps are retained and stacked temporally, forming a high-dimensional 5-D tensor of shape (samples, 13, H, W, 1), where H and W denote the spatial resolution. This representation preserves both spatial motion patterns and temporal progression, enabling deep neural networks to learn discriminative facial motion characteristics automatically. Figure 5 represents the feature extrac-tions strategy based on low, medium and high motion dynamics obtained from the average and maximum optical flow magnitude and its motion significance be-tween consecutive frames for a particular subject.

After this, all extracted features are standardized to zero mean and unit vari-ance to facilitate stable and efficient model training. Corresponding pain-level labels are encoded using label encoding followed by one-hot representation to support multi-class classification.

5.3. Model Training

For model training, the dataset is partitioned into training and testing subsets using 80-20 split. Model training is performed using the Adam optimizer with cat-egorical cross-entropy as the loss function. A small batch size ranging from 8 to 16 is employed to accommodate sequence-based temporal models. The performance

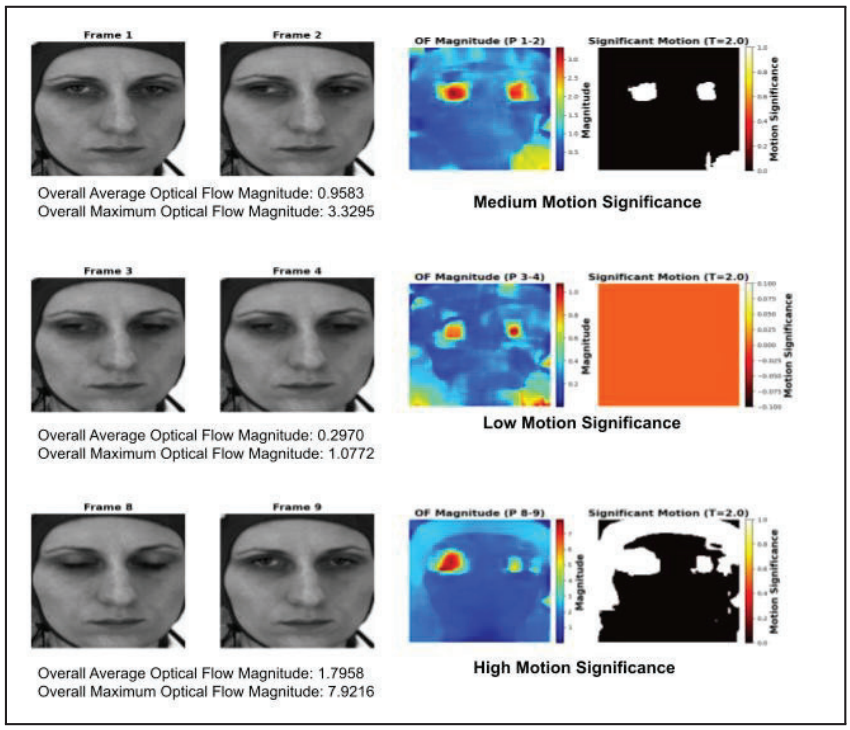


Figure 5: Spatio-Temporal Feature Extraction.

is evaluated using multiple standard classification metrics, including overall accu-racy, class-wise precision, recall, and F1-score. In addition, macro-averaged and weighted-averaged F1-scores were reported to account for class imbalance. Con-fusion matrix analysis is further employed to provide detailed view into class-level prediction behavior. We have evaluated a diverse set of models to assess the effec-tiveness of different spatial, temporal, and motion-based representations. These include: (i) a baseline statistical optical-flow model combined with a deep neural network (DNN); (ii) a CNN–BiLSTM with attention mechanism using raw facial appearance features without optical flow; (iii) a ResNet18–BiLSTM with atten-tion for appearance-based temporal modeling; (iv) CNN–(Bi)LSTM architectures incorporating optical flow features; (v) ensemble CNN–(Bi)LSTM models to im-prove robustness; and (vi) a Transformer-based temporal encoder utilizing optical flow features, which constitutes the proposed approach.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

6. Result and Discussion

This section presents a comparative analysis of all implemented approaches to evaluate the effectiveness of different facial motion representations and temporal modeling strategies for pain-inferred mental health prediction across models.

6.1. Baseline Model: Statistical Optical Flow Features

The handcrafted statistical optical flow and deep neural network baseline es-tablish the lowest reference point among all evaluated methods. By compress-ing each frame-to-frame optical flow map into a single mean magnitude value, this approach captures only coarse motion intensity trends while discarding spa-tial structure and detailed temporal dynamics. A total of 1799 samples with 13-dimensional features were represented. A fully connected deep neural network was employed for multi-class classification of pain intensity levels. The network architecture consists of an input layer corresponding to the 13-dimensional fea-ture vector, followed by four hidden layers with 256, 128, 64, and 32 neurons, respectively. Each hidden layer uses the ReLU activation function to introduce non-linearity. Batch normalization is applied after the first three hidden layers to stabilize training and accelerate convergence, while dropout with a rate of 0.3 is incorporated to mitigate overfitting. The output layer employs a softmax ac-tivation function to predict the probability distribution over the five pain classes (BL1, PA1–PA4). The model is trained using the Adam optimizer with categorical cross-entropy as the loss function. Training is conducted for 50, 100, 150, 200 and 250 epochs with a batch size of 8 and 16, and model performance is observed on validation set. Results exhibit that a huge drop in validation accuracy over train-accuracy which is represented in Figure 6.

The results indicate that such compact statistical descriptors are insufficient for fine-grained discrimination among closely related pain levels. Frequent con-fusion is observed between adjacent pain classes, suggesting that global motion intensity alone cannot adequately represent subtle facial expressions associated with stress and pain perception. This baseline confirms the necessity of richer spatiotemporal motion modeling. Furthermore, the baseline DNN was extended to a hybrid CNN–LSTM architecture to better capture local feature interactions and sequential dependencies within the 13-dimensional facial motion descriptors. Prior to modeling, all features were standardized using z-score normalization and reshaped into a three-dimensional tensor of size (N,13,1),treating the feature di-mension as a temporal sequence. Class labels corresponding to the five pain levels

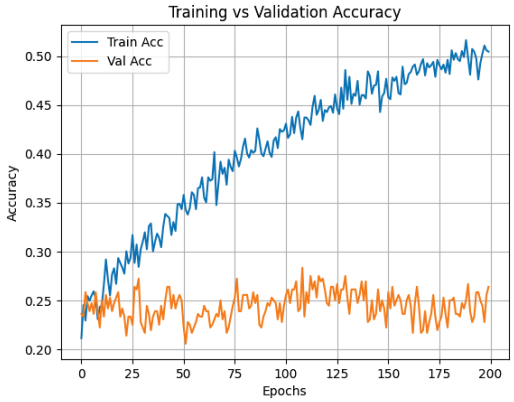


Figure 6: Training and validation accuracy for Baseline.

(BL1–PA4) were encoded using label encoding followed by one-hot representa-tion. The dataset was split into training and testing subsets with an 80:20 ratio. The proposed model begins with a stack of one-dimensional convolutional layers comprising 64, 128, and 256 filters, each with a kernel size of 3 and ReLU ac-tivation, enabling hierarchical feature extraction from the input sequence. Batch normalization and dropout regularization (rate = 0.3) are applied after each con-volutional block to improve training stability and reduce overfitting, while max-pooling is employed after the first convolutional layer to downsample feature maps. The extracted features are then fed into an LSTM layer with 32 hidden units to model temporal correlations across the feature sequence, followed by an additional dropout layer with a rate of 0.4. The final classification is performed using a fully connected softmax layer corresponding to the five output classes. The model is trained using the Adam optimizer with a learning rate of 0.0005 and categorical cross-entropy loss for 50, 100, 150, 200 and 250 epochs with a batch size of 8 and 16, and performance is evaluated using accuracy, class-wise preci-sion–recall metrics. The training and validation accuracy is illustrated in Figure 7.

The accuracy for both the baseline DNN model and extended hybrid CNN-LSTM model remains an area of concern with an empirical value of 26% and 25% respectively. In spite of adding temporal modeling with the baseline DNN model, it fails to achieve better performance. In order to enhance the model perfor-mance, we capture the spatial information as well. The confusion matrix analysis is illusrated in Figure 8.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

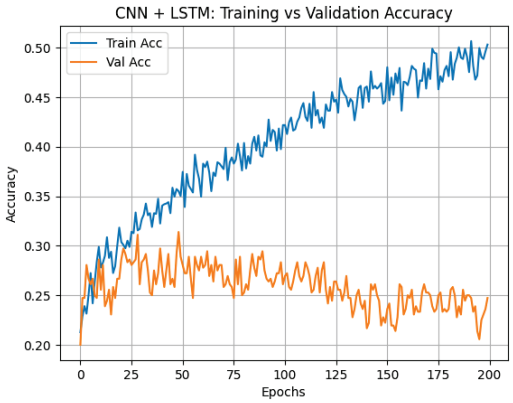


Figure 7: Training and validation accuracy for CNN+LSTM.

True / Pred	BL1	PA1	PA2	PA3	PA4
BL1	**19**	18	12	13	7
PA1	22	**20**	14	13	12
PA2	23	8	**15**	11	8
PA3	16	9	15	**18**	10
PA4	10	7	17	20	**23**

(a) Baseline DNN

True / Pred	BL1	PA1	PA2	PA3	PA4
BL1	**17**	9	17	12	14
PA1	14	**20**	22	14	11
PA2	15	10	**13**	17	10
PA3	9	10	21	**18**	10
PA4	10	14	15	17	**21**

(b) CNN+LSTM

Figure 8: Comparison of confusion matrices for the baseline DNN and CNN+LSTM model.

6.2. Appearance-Based Spatiotemporal Models

A spatiotemporal deep learning model based on a TimeDistributed CNN–LSTM architecture was designed to jointly learn spatial facial representations and their temporal evolution. Each input sample consists of a sequence of 13 grayscale frames with a spatial resolution of 32×32 pixels, represented as a five-dimensional tensor. Spatial feature extraction is performed independently on each frame using a two-dimensional convolutional layer with 16 filters of size 3×3 and ReLU acti-vation, followed by max-pooling for spatial downsampling. The resulting feature maps are flattened within a TimeDistributed framework to preserve the temporal ordering of frame-level features. These sequential embeddings are then processed by an LSTM layer with 64 hidden units to model temporal dependencies across frames. Dropout regularization with a rate of 0.5 is applied to mitigate overfitting, and the final classification is performed using a fully connected softmax layer to predict one of the five pain intensity classes (BL1–PA4). Figure 9 shows that train-accuracy vs validation-accuracy for this experiment achieves the peak value

of 43% for validation with 120 epochs and then it gradually decreases.

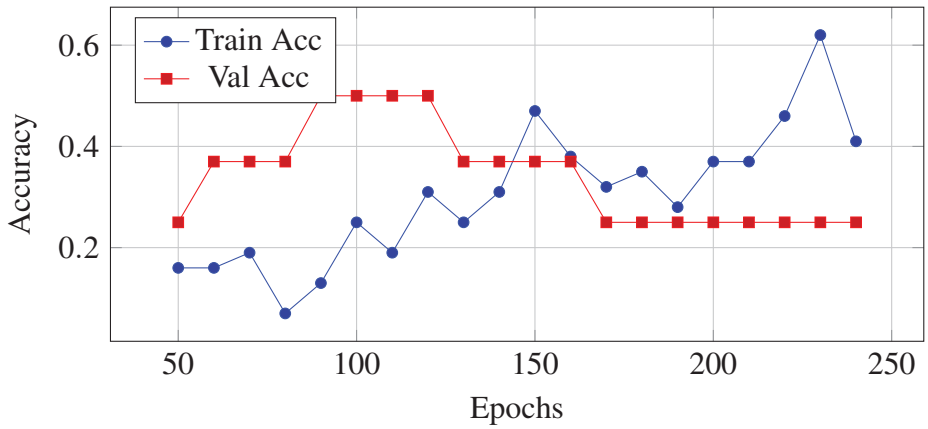


Figure 9: Training and validation accuracy of the TimeDistributed CNN–LSTM model .

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

To observe the importance of optical flow in the above experiment we have modeled a CNN–BiLSTM Attention architecture for spatial facial representations. Given an input tensor of dimensions (B,T,C,H,W), where B denotes the batch size and T the number of frames and $C \times H \times W$ the spatial dimensions of each grayscale frame, spatial feature extraction is first performed using a convolutional neural network. The CNN backbone consists of three convolutional blocks with 32, 64, and 128 filters, respectively, each employing a 3×3 kernel with padding, followed by ReLU activation. Max-pooling layers are applied after the first two convolutional blocks to progressively reduce spatial resolution, while an adaptive average pooling layer produces a fixed spatial output of 4×4, ensuring robust-ness to variations in input frame size. The resulting feature maps are flattened to form compact frame-level embeddings. For temporal dynamics, the sequence of CNN-extracted features is fed into a bidirectional LSTM with 256 hidden units in each direction, enabling the network to exploit both past and future contextual information. The BiLSTM outputs a sequence of 512-dimensional hidden states, which are subsequently processed by an attention mechanism to learn a set of nor-malized weights over the temporal dimension, allowing the model to emphasize salient frames that contribute most strongly to pain-related facial expressions. A weighted temporal context vector is computed as a weighted sum of the BiLSTM outputs and serves as a global representation of the input sequence. Finally, this context vector is passed through a fully connected layer with softmax activation to predict one of the five pain intensity classes (BL1–PA4). This shows a major

drop in the performance of the model with an accuracy of 17%.

We have also employed high-level appearance-based spatio-temporal model using pretrained ResNet18 backbone for spatial feature extraction followed by Bidirectional LSTM (BiLSTM) and an attention mechanism for temporal model-ing to compare with the statistical motion baseline. By jointly modeling spatial facial features and temporal evolution, these approaches are able to capture sus-tained expression patterns over time, thus improving the overall accuracy to 26% . The two models CNN-BiLSTM and RESNET- BiLSTM are evaluated on certain performance metrics, which is represented in Table 3.

Table 3: Class-wise performance comparison of CNN+BiLSTM and ResNet+BiLSTM models

Class	**CNN + BiLSTM**			**ResNet + BiLSTM**		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
BL1	0.17	1.00	0.29	0.22	1.00	0.35
PA1	0.00	0.00	0.00	0.19	0.00	0.24
PA2	0.00	0.00	0.00	0.00	0.00	0.00
PA3	0.00	0.00	0.00	0.00	0.00	0.00
PA4	0.00	0.00	0.00	0.00	0.00	0.00
Accuracy		0.17			0.21	

From these experiments, we observed that the models are unable to classify the PA2, PA3, PA4 classes as the motion-awareness mechanism was missing. To handle this, we used optical flow based motion modeling to capture the frame to frame stimuli. Table 4 demonstrates the comparison study of different appearance-based spatio-temporal model.

Table 4: Comparison of Appearance based spatio-temporal model.

Model	**Motion Modeling**	**Appearance Modeling**	**Role**
CNN + BiLSTM + Attention	No	Moderate	Baseline
ResNet18 + BiLSTM + Attention	No	Strong	Enhanced appearance baseline
Optical flow–based models	Yes	Moderate	Motion-aware

However, despite the use of bidirectional temporal modeling and attention mechanisms, these models exhibit limitations in distinguishing low-intensity and intermediate pain levels. The absence of explicit motion encoding restricts their ability to capture micro-movements and subtle facial muscle activations, which are critical indicators of stress-related mental states. These observations sug-gest that strong appearance modeling alone is not sufficient for robust mental health–oriented pain prediction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

6.3. Motion-Aware Spatiotemporal Modeling with Attention

In our experiments, incorporating dense optical flow representations as input consistently improved performance across all evaluated deep learning architec-tures, underscoring the critical role of explicit motion modeling in pain and stress inference. By applying CNN-based spatial encoders to optical flow magnitude maps, the proposed framework effectively captured localized facial motion pat-terns, while recurrent temporal layers modeled their evolution over time. Our LSTM-based temporal modeling results significantly outperform appearance-only approaches, confirming that dynamic facial information provides complementary and more discriminative cues beyond static visual features. Furthermore, replac-ing unidirectional LSTM with bidirectional LSTM (BiLSTM) yielded additional performance gains, indicating that using contextual information from both past and future frames enhances the interpretation of precise facial motion dynamics associated with pain-induced stress. By assigning higher weights to temporally salient frames, such as those corresponding to peak facial responses to pain stim-uli enables the network to focus on the most informative segments while sup-pressing less relevant or redundant temporal information. Qualitative analysis suggests that attention particularly benefits the recognition of pain levels charac-terized by brief or localized facial reactions, thereby improving robustness under inter-subject variability. Overall, these findings demonstrate that motion-aware spatiotemporal learning, augmented with adaptive attention, is critical for captur-ing the nuanced facial responses underlying pain-related stress and mental health inference.

6.4. Transformer-Based Temporal Modeling

The proposed two-stream Transformer-based framework was evaluated on the BioVid Heat Pain dataset using synchronized facial frame sequences and dense optical flow representations. Each video clip was uniformly sampled to 14 grayscale frames of size 128×128, and dense optical flow between consecutive frames was computed using Farnebäck's algorithm, yielding horizontal and vertical motion components that were normalized on a per-clip basis. Appearance and motion streams were independently encoded using ResNet-18, projected into a shared embedding space, and temporally modeled using Transformer encoder layers with positional encoding to capture long-range facial dynamics. The resulting clip-level representations were fused and classified into five pain levels. Training was performed using weighted cross-entropy loss to address class imbalance, opti-mized with AdamW and cosine annealing learning-rate scheduling. Performance was evaluated using accuracy and macro-averaged F1-score on validation and test

splits, with early stopping applied based on validation F1-score. Quantitative results demonstrating the effectiveness of motion-aware temporal modeling and two-stream fusion are reported in Figure 10.

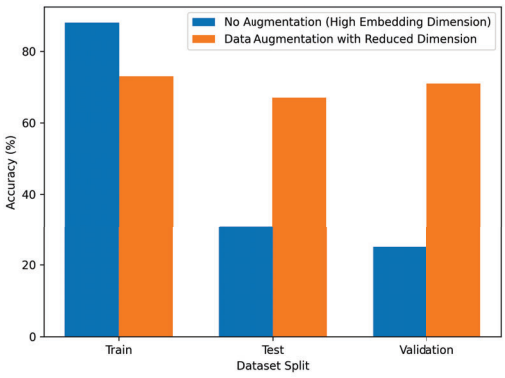


Figure 10: Training and validation accuracy for Augmented and Un-augmented Data.

The Transformer Temporal Encoder with Optical Flow exhibits the most con-sistent and discriminative behavior among all evaluated approaches. By using self-attention, the Transformer is able to model global temporal dependencies across the entire motion sequence, overcoming the limitations of recurrent memory-based models. Performance on the test set before any parameter tuning is rep-resented on Table 5. The test accuracy that we have obatined was 88% while validation accuracy indicates a major drop to 25%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Table 5: Classification performance on the test set

Class	Precision	Recall	F1-score	Support
BL1	0.35	0.28	0.31	60
PA1	0.10	0.03	0.05	60
PA2	0.21	0.20	0.21	60
PA3	0.16	0.10	0.12	60
PA4	0.27	0.62	0.37	60
Accuracy		0.25		300
Macro Avg	0.22	0.25	0.21	300
Weighted Avg	0.22	0.25	0.21	300

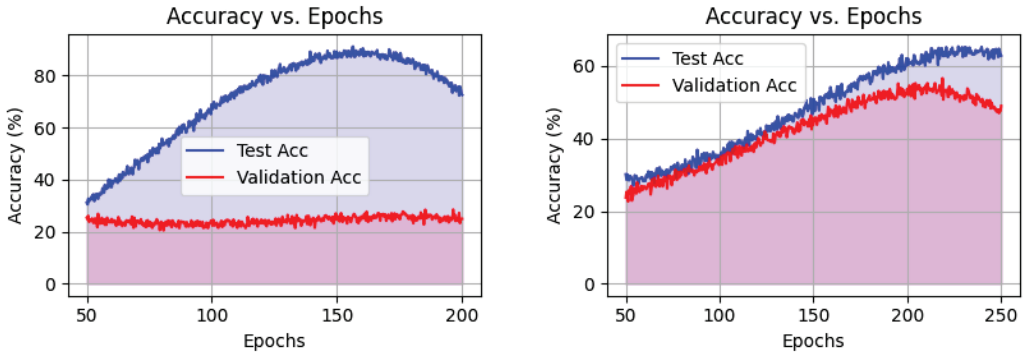
The performance observed a huge validation loss. To address the issue of high validation loss, two new classes were introduced. The first, TwoStream-BioVid V2, incorporates stronger data augmentation strategies, including random

resized cropping and horizontal flipping applied to the training samples. The second, TwoStreamTransformer V2, adopts a more regularized architecture by reducing model complexity through a smaller embedding dimension, fewer at-tention heads and transformer layers, a reduced feedforward dimension, and in-creased dropout rates to mitigate overfitting by replacing original training, valida-tion, and test dataset instances with the new train,test and validation datasets. The performance is illustrated in Table 6.

Table 6: Performance after Data augmentation and horizonatal flipping

Class	Precision	Recall	F1-score	Support
BL1	0.47	0.38	0.45	60
PA1	0.52	0.43	0.45	60
PA2	0.39	0.32	0.38	60
PA3	0.56	0.60	0.52	60
PA4	0.57	0.62	0.37	60
Accuracy		0.47		300

Unlike LSTM-based approaches, the Transformer processes all temporal steps in parallel and dynamically attends to the most informative motion patterns. The results indicate improved separation between pain levels, particularly for cases involving subtle or temporally distributed facial motion. Figure 11 illustrates the performance comparison between the two tansformer-based encoder approach that we have integrated in our experiment with respect to different epochs



(a) Accuracy vs. Epochs without data augmentation.

(b) Accuracy vs. Epochs with data augmentation.

Figure 11: Comparison of test and validation accuracy trends across training epochs under differ-ent training strategies.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

The two-stream formulation using dense optical flow further reinforces this in-terpretation. Optical flow explicitly captures involuntary facial muscle activations, which are known behavioral correlates of both pain and stress. The empirical improvement observed when motion-aware representations are introduced sup-ports the hypothesis that the learned model is responding to stress-induced facial dynamics, rather than superficial visual cues. Consequently, the predicted pain intensity $\hat{P}$ can be interpreted as an intermediate representation of facial stress dy-namics, as formalized by $P \approx f$ (facial stress dynamics). Furthermore, the mono-tonic mapping from predicted pain to latent inferred mental stress is experimen-tally justified by the ordinal structure of the classification task. This monotonicity constraint ensures physiological plausibility while avoiding the need for direct self-reported stress annotations.

Under this formulation, the model does not merely perform pain classifica-tion; instead, it provides an indirect yet objective estimation of mental stress de-rived from observable behavioral stress indicators encoded in facial dynamics. This hierarchical mapping from facial behavior to pain, and from pain to mental stress enables the proposed framework to serve as a non-invasive stress assess-ment model, particularly relevant for mental health monitoring in scenarios where direct stress measurement is impractical or unreliable.

7. Conclusion

Although the model is trained using pain labels, the learned representations en-code facial stress dynamics that provide an indirect yet objective estimate of latent mental stress. The hierarchical inference formulation mapping facial behavior to pain intensity and subsequently to mental stress via a monotonic function offers a principled bridge between observable behavioral stress indicators and internal mental states, without relying on subjective self-reports. This makes the proposed framework particularly suitable for mental health monitoring in controlled or clin-ical environments where direct stress measurement is challenging. Moreover, the absence of explicit mental stress ground truth restricts the evaluation to proxy-based validation. Future work will focus on addressing these limitations by incor-porating multimodal physiological signals (e.g., ECG, EDA, or thermal imaging) to strengthen stress inference, and by exploring ordinal or regression-based for-mulations that better respect the continuous nature of stress intensity. Domain adaptation and subject-independent learning strategies will also be investigated to improve generalization across individuals.

Funding

This research received no funding.

Conflict of Interest

The authors declare no conflict of interest.

Authors' contributions

Conceptualization, S.D., and S.U.; methodology, S.D., and S.U.; software, S.U.; validation, S.D., S.U.; formal analysis, S.D., and S.U.; investigation, S.U.; resources, S.D.; data curation, S.D.; writing—original draft preparation, S.D., and S.U.; writing—review and editing, S.D., and S.U.; visualization, S.D., and S.U.; supervision, S.U.; project administration, S.U.; All authors (Soumalya De (S.D.), and Saiyed Umer (S.U.) have read and agreed to the published this version of the manuscript.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

References

[1] World Health Organization. *Mental Health Atlas 2023*. World Health Orga-nization, Geneva, Switzerland, 2023.

[2] Thomas R. Insel. Digital phenotyping: Technology for a new science of behavior. *JAMA*, 318(13):1215–1216, 2017.

[3] Katja Wiech and Irene Tracey. The influence of negative emotions on pain: Behavioral effects and neural mechanisms. *Pain*, 140(3):447–455, 2009.

[4] A. Vania Apkarian, J. A. Hashmi, and Marwan N. Baliki. Pain and the brain: Specificity and plasticity of the brain in clinical chronic pain. *Neuron*, 87(1):15–28, 2011.

[5] M. Catherine Bushnell, MartaČeko, and Lindsay A. Low. Cognitive and emotional control of pain and its disruption in chronic pain. *Nature Reviews Neuroscience*, 14(7):502–511, 2013.

[6] M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000.

[7] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. *Arti-ficial Intelligence*, 17(1–3):185–203, 1981.

[8] Gunnar Farnebäck. Two-frame motion estimation based on polynomial ex-pansion. In *Proceedings of the 13th Scandinavian Conference on Image Analysis (SCIA)*, pages 363–370. Springer, 2003.

[9] Steffen Walter, Sascha Gruss, Heiko Ehleiter, Junwen Tan, Harald C. Traue, Philipp Werner, Ayoub Al-Hamadi, and Steffen Crawcour. The biovid heat pain database: Data for the advancement and systematic validation of an automated pain recognition system. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 128–131. IEEE, 2013.

[10] Markus Kächele, Patrick Thiam, Mohammad Amirian, Philipp Werner, Stef-fen Walter, Günther Palm, and Friedhelm Schwenker. Multimodal data fu-sion for person-independent, continuous estimation of pain intensity. *IEEE Transactions on Affective Computing*, 7(3):253–264, 2016.

[11] Philipp Werner, Ayoub Al-Hamadi, Reinhard Niese, Steffen Walter, Sascha Gruss, and Harald C. Traue. Automatic pain recognition from video and biomedical signals. *IEEE Transactions on Affective Computing*, 4(1):1–13, 2014.

[12] Y. Zhang, X. Liu, and H. Wang. A multimodal dataset for stress detection using facial and physiological signals. *Scientific Data*, 12:112, 2025.

[13] H. Alshamsi, S. Al-Maadeed, and A. Bouridane. Emotion recognition tech-nologies and their applications in mental health: A systematic review. *BMC Psychology*, 12(1):98, 2024.

[14] Markus Kächele, Philipp Werner, and Steffen Walter. Deep learning–based pain intensity estimation from facial videos using the biovid heat pain dataset. *Scientific Reports*, 15:21987, 2025.

[15] Patrick Thiam, Mohammad Amirian, and Friedhelm Schwenker. Multi-modal temporal fusion for continuous pain intensity estimation. *IEEE Trans-actions on Affective Computing*, 14(2):923–935, 2023.

[16] J. Li, Y. Chen, and L. Zhao. Mental health state prediction using optical flow–based facial motion analysis. *Scientific Reports*, 15:29461, 2025.

[17] M. Rahman, S. Islam, and M. Hossain. Visual encoding of physiological signals for stress and emotion classification. Scientific Reports, 15:1228, 2025.

[18] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C. Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O. Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In 2013 IEEE International Conference on Cybernetics (CYBCO), pages 128–131, 2013.