Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 3, July 2026, pp: 1163-1183 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Deep Learning Based Pain Recognition via Facial Expression Using Feature Fusion of Visual and Thermal Images

Raihan Islamadina[1,2], Fitri Arnia[1,3], Taufik F. Abidin[4], Rusdha Muharar[1,3], Aulia Syarif Aziz[2], and Khairun Saddami[3]

[1] School of Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia
[2] Universitas Islam Negeri Ar-Raniry, Banda Aceh, Indonesia
[3] Department of Electrical and Computer Engineering, Universitas Syiah Kuala, Banda Aceh, Indonesia
[4] Department of Informatics, Universitas Syiah Kuala, Banda Aceh, Indonesia

Corresponding author: Fitri Arnia (e-mail: f.arnia@usk.ac.id), **Author(s) Email**: Raihan Islamadina (e-mail: raihanislamadina@ar-raniry.ac.id), Taufik F. Abidin (e-mail: taufik.abidin@usk.ac.id), Rusdha Muharar (e-mail: r.muharar@usk.ac.id), Aulia Syarif Aziz (e-mail: aulia.aziz@ar-raniry.ac.id), Khairun Saddami (e-mail: khairun.saddami@usk.ac.id)

Abstract **Objective pain assessment in non-verbal patients remains a significant clinical challenge. While automated facial expression analysis offers a promising solution, existing methods often rely on simple cross-modality concatenation or shallow attention mechanisms, which typically fail to fully capture complex, interdependent cross-modal dynamics. To address this limitation, this study proposes a deep learning-based multimodal feature fusion framework that combines visual and thermal images to significantly enhance pain detection accuracy. The proposed framework integrates pretrained convolutional neural network backbones, specifically VGGFace, ResNet50, and DEYOLO, with two novel attention modules: Dual Semantic Enhancing Channel Weight Assignment (DECA), and Dual Spatial Enhancing Pixel Weight Assignment (DEPA) to adaptively optimize joint feature representations. For comparison, a hybrid baseline model that processes visual and thermal images separately was also developed, allowing a comparative analysis between fusion-based and non-fusion-based approaches. The model performance was systematically evaluated on the MIntPain dataset, which comprises 20 healthy subjects experiencing five distinct levels of pain intensity. To ensure data independence and prevent identity leakage across the training and testing phases, a strict subject-wise data split protocol was implemented with 15 subjects allocated for training and 5 subjects for testing. Experimental results demonstrate that the proposed multimodal fusion framework achieves superior performance, attaining a peak F1 score of 0.938 using the VGGFace backbone. Furthermore, external validation on the UNBC McMaster Shoulder Pain dataset yields a classification accuracy of 0.872, confirming the strong generalization capability and stability of the framework across unseen subjects. These findings highlight the effectiveness of visual-thermal synergy and the efficacy of the proposed DECA and DEPA modules, showcasing high potential for robust, non-invasive clinical pain-monitoring and assessment applications.**

Keywords **Facial pain classification; Visual imaging; Thermal imaging; DECA and DEPA modules; Multimodal learning.**

I. Introduction

Objective pain assessment remains a critical clinical challenge, as conventional measurement heavily relies on subjective self-reports [1]. This reliance risks suboptimal pain management, particularly for vulnerable populations unable to communicate verbally, such as infants, elderly individuals with dementia, or unconscious patients [2]. To bridge this gap, automated pain recognition systems utilizing non-verbal cues have emerged as an increasingly vital solution in modern clinical practice. Among these cues, facial expressions are widely recognized as a primary, universal indicator of pain, with the Facial Action Coding System (FACS) providing a structured foundation through specific action units [3], [4], [5], [6]. Given the multidimensional complexity of pain, recent advances have shifted from unimodal analysis to multimodal frameworks that leverage hierarchical fusion to extract richer, more comprehensive features [7], [8]. Existing literature has successfully implemented this paradigm by fusing handcrafted descriptors [9], such as facial landmarks [10], and

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

Histogram of Oriented Gradients (HOG) [11], with deep representations extracted from robust pretrained networks like VGG16 [12] and ResNet50 [13].

Previously, integrating visual (RGB), depth, and thermal modalities has demonstrated substantial improvements in pain recognition accuracy, as evidenced by the multimodal MIntPAIN dataset [14]. However, because acquiring depth data in real-world clinical environments introduce high computational complexity and hardware costs, this study deliberately focuses on the optimal combination of visual and thermal channels [15]. Thermal imaging offers notable operational advantages; it is highly insensitive to ambient lighting variations and uniquely capable of capturing subtle physiological metabolic changes invisible to the naked eye [16]. Consequently, this dual modality coupling provides an excellent complement to visual imaging in challenging, dynamic clinical environments.

To maximize the efficacy of this visual-thermal coupling, various deep learning architectures have been explored. Prior works have integrated the HSV color space and VGGFace with Temporal Convolutional Networks (TCN) [17], or evaluated standard backbones like ResNet [13], MobileNetV2 [18], and EfficientNet [19] directly on thermal images [20], [21]. Beyond baseline feature extraction, systematic reviews and recent medical imaging trends emphasize that feature-level fusion must be reinforced by adaptive attention mechanisms to guarantee cross-modality consistency, semantic alignment, and temporal exploitation [22], [23]. While combining visual and thermal data inherently provides robustness under illumination variations, a critical limitation persists: most existing frameworks rely on superficial feature-level aggregation (e.g., simple concatenation or averaging). They fail to leverage advanced, dual-stage attention mechanisms capable of dynamically guiding the network to isolate 'what' semantic channels and 'where' micro spatial deformations are most relevant from each heterogeneous modality.

Expanding on these limitations, conventional attention based fusion methods generally rely on static feature concatenation that ignores the highly complementary semantic and spatial relationships between visual and thermal domains [24]. Such approaches are inherently insufficient for pain recognition, where pain-induced facial cues are typically subtle, localized, and highly variable across individuals [25]. Furthermore, existing attention mechanisms in multimodal pain analysis focus primarily on global feature weighting, lacking the adaptive refinement necessary to simultaneously emphasize informative regions and discriminative semantic representations [26]. Consequently, the complex, cross-modal interactions between physiological thermal patterns and visual facial expression features are not optimally captured.

To systematically overcome these challenges, this study introduces a dual attention multimodal fusion framework powered by Dual Semantic Enhancing Channel Weight Assignment (DECA) and Dual Spatial Enhancing Pixel Weight Assignment (DEPA) [27]. Unlike conventional static fusion, the proposed framework adaptively refines both semantic channel dependencies and spatially localized facial target regions to fully unlock the complementary synergy of visual thermal representations. Although DECA and DEPA were originally conceptualized within the DEYOLO object detection framework to resolve illumination discrepancies, their transferability to multimodal pain recognition using pretrained, heavy CNN-based feature extractors has not been previously explored [27].

To clarify our boundaries, the core novelty lies not in the baseline invention of DECA and DEPA, but in their strategic domain adaptation and lightweight architectural integration tailored for automated pain classification. Specifically, the proposed framework re-engineers these modules to mitigate modality conflicts between highly textured RGB expressions and low texture thermal maps through a two-fold mechanism. First, DECA dynamically isolates global semantic channels heavily corresponding to pain while filtering cross-modality noise and subject-specific identity artifacts, ensuring the features generalise across diverse individuals. Second, DEPA establishes a robust spatial positional awareness, forcing the model to target precise facial Action Units (e.g., brow lowering or eye squeezing) where muscle deformations and autonomic thermal variations synchronize. Integrating these decoupled refinement mechanisms into pretrained backbones (VGGFace and ResNet50) yields highly discriminative representations while minimizing background interference.

Driven by these advancements, this study aims to deliver an illumination-insensitive, highly robust automated pain recognition system. To fulfill this objective and comprehensively validate our approach, we present two primary architectural contributions:

1. We develop an advanced feature refinement fusion framework that integrates pretrained CNN backbones with DECA and DEPA modules to adaptively enhance cross-modal semantic and spatial interactions.
2. We implement an independent hybrid modeling architecture as a comparative baseline, where visual and thermal modalities are processed separately from input to classification.

To rigorously evaluate these systems, we conduct extensive cross-subject evaluations alongside

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

statistical comparisons. Together, these frameworks provide an in-depth analysis of unimodal, fusion-based, and hybrid strategies, establishing a flexible and deployment-ready benchmark for objective clinical pain assessment.

II. Method

A. The Proposed Framework

This study proposes a classification framework for pain intensity based on facial expressions by integrating two types of images: visual and thermal. The approach aims to overcome the limitations of unimodal models in accurately recognizing facial expressions, particularly under low-light conditions or when expressions are visually subtle. To overcome these challenges, the framework introduces a novel joint feature refinement mechanism powered by the Dual Enhanced Attention (DEA) module, which couples two distinct attention components, DECA and DEPA [28]. Operating in a decoupled yet complementary manner, the DECA module focuses on enhancing the semantic representation of features through a channel-wise weighting mechanism applied to the feature maps. It is designed to emphasize channels that contain important information while suppressing redundant or noisy channels. In contrast, the DEPA module is intended to improve spatial representation by applying pixel level weighting. This module highlights critical location and texture patterns essential to understanding the structural characteristics of objects in the image. The synergy between DECA, which determines what information is globally important, and DEPA, which emphasizes where this information is spatially located in the fused representation, results in significantly richer and more robust feature representations. These two modules serve as key components for improving joint feature representation across both modalities. This framework introduces a novel synergy between semantic and spatial attention mechanisms in multimodal fusion a domain that has not been thoroughly investigated in earlier studies on pain expression classification [7], [20], [21]. Fig. 1 illustrates the overall architecture of the proposed pain intensity classification framework.

The operational pipeline begins with automated facial segmentation using the Multi-task Cascaded Convolutional Networks (MTCNN) algorithm [29] to precisely isolate and crop the facial regions of interest from the raw images. Following segmentation, these cropped facial regions are fed into pretrained Convolutional Neural Networks (CNNs) acting as deep feature extractors. Based on preliminary empirical evaluations comparing various network capacities, VGGFace [12], [27] and ResNet50 [13] were selected as the primary structural backbones due to their superior ability to extract discriminative facial topology and thermal metabolic maps. To preserve these low to mid-level spatial representations (such as edge configurations and localized textures) learned from large-scale datasets, both networks were modified by removing their final classification layers. This architectural re-engineering yields robust 1D feature vectors of 4096 dimensions for VGGFace and 2048 dimensions for ResNet50, successfully mitigating overfitting risks inherent to training on relatively small clinical pain datasets. While the DEYOLO architecture [28] is also evaluated in this study, its utilization is strictly confined to serving as a lightweight comparative baseline to gauge the transferability of the attention modules, rather than as a modified primary feature extractor. To activate the spatial semantic refinement, the extracted 1D feature vectors from each modality are mathematically transformed. These vectors are reshaped into 4D spatial tensors, denoted as $F \in \mathbb{R}^{b\times c\times h\times \omega}$ (where $b \times c \times h \times \omega$ represent batch size,

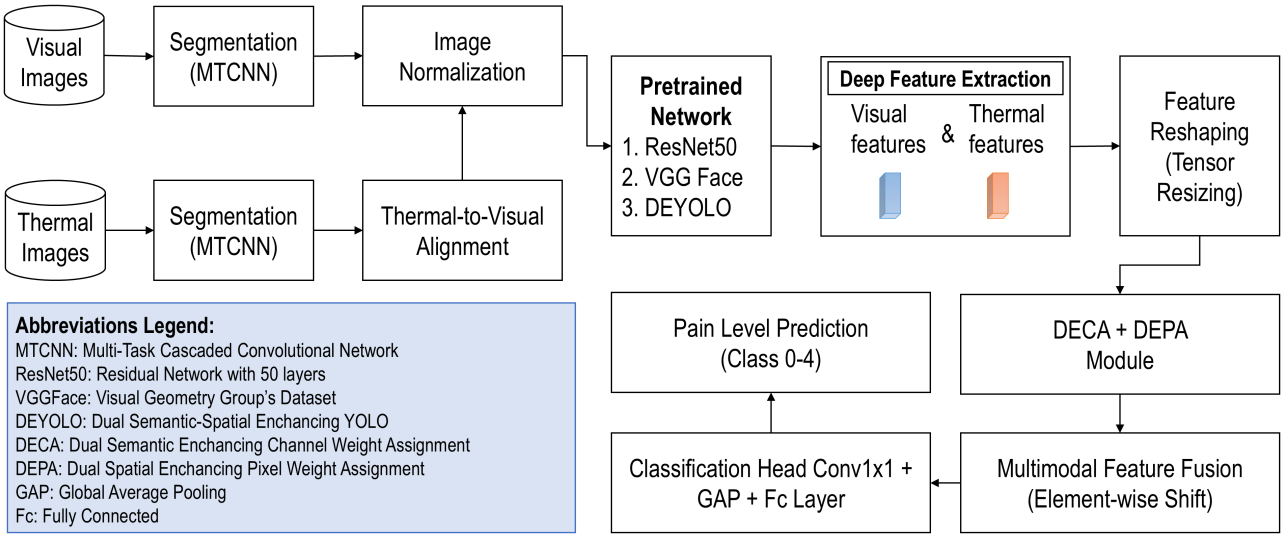


Fig. 1. Fusion of visual image and thermal image features.

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

channels, height, and width, respectively), before being sequentially processed by the DEA module. The mathematical formalization of this entire pipeline is structured as follows:

1. Input Normalization and Computational Feature Extraction

Prior to deep feature extraction, the segmented facial regions undergo input data normalization to ensure numerical stability and accelerate gradient convergence during training. The segmented visual facial image I_V and thermal facial image I_{th} are transformed into a uniform scale space using min-max scaling. The normalized visual image $\bar{I}_V$ is computed via Eq. (1) [30]:

$$\bar{I}_V = \frac{I_V - \min(I_V)}{\max(I_V) - \min(I_V)} \quad (1)$$

where I_V represents the original input visual matrix $\min(.)$ and $\max(.)$ denote the mathematical operators that extract the lowest and highest pixel intensity values within the spatial domain of the image, respectively, and $\bar{I}_V$ represents the resulting scaled visual image matrix bounded strictly within the interval $[0, 1]$. Following the identical standardization procedure, the normalized infrared thermal representation $\bar{I}_{th}$ is obtained using Eq. (2) [30]:

$$\bar{I}_{th} = \frac{I_{th} - \min(I_{th})}{\max(I_{th}) - \min(I_{th})} \quad (2)$$

where I_{th} signifies the raw input infrared thermal matrix aligned via affine transformations, and $\bar{I}_{th}$ is the corresponding normalized thermal tensor generated within the range of $[0, 1]$.

These two normalized image matrices are subsequently fed into their respective deep convolutional neural network (deep CNN) backbones, which serve as feature extractors based on principles of facial expression. The initial deep visual feature map F_{V_0} is formulated via Eq. (3) [30]:

$$\mathrm{F}_{V_0} = \mathcal{H}_v(\bar{I}_V) \in \mathbb{R}^{b \times c \times h \times \omega} \quad (3)$$

where $\mathcal{H}_v(.)$ represents the non-linear feature extraction mapping function of the modified backbone, b denotes the operational training batch size, c signifies the channel depth dimension, and $h \times \omega$ epresents the structural spatial resolution grid consisting of the feature map height h and width ω. Parallelly, the initial deep thermal feature map F_{th_0} is generated via Eq. (4):

$$\mathrm{F}_{th_0} = \mathcal{H}_{th}(\bar{I}_{th}) \in \mathbb{R}^{b \times c \times h \times \omega} \quad (4)$$

where $\mathcal{H}_{th}(.)$ defines the composite feature extraction layers of the backbone neural network architecture, producing an initial thermal tensor space that matches the exact spatial dimensions and channel depth of the visual stream.

2. Mixed Feature Initialization

Prior to entering the attention calibration module stages, the two baseline modality feature maps are concatenated and aligned to construct a single initial mixed feature map $\mathrm{F}_{Mix_0} \in \mathbb{R}^{b \times c \times h \times \omega}$ via Eq. (5) [28]:

$$\mathrm{F}_{Mix_0} = conv(concat(\mathrm{F}_{V_0}, \mathrm{F}_{th_0})) \quad (5)$$

where $concat(.)$ denotes the concatenation or tensor joining operator along the channel dimension axis, and $conv(.)$ represents the dimensional alignment convolution operation used to compress the combined $2c$ channels back to a single channel size c while simultaneously filtering out initial redundant information.

3. Dual Semantic Enhancing Channel Weight Assignment (DECA) Module

To resolve inter-channel information conflicts and highlight discriminative semantic channels, the mixed feature map F_{Mix_0} is passed into the DECA module. The first step involves global cross-modality weight encoding via the Cross Modality Weight Extraction ($CMWE$) operation. This operation progressively compresses the spatial components of F_{Mix_0} until it reaches a pure channel resolution via Eq. (6) [28]:

$$W_{Mix_0} = CMWE(\mathrm{F}_{Mix_0}) \in \mathbb{R}^{b \times c \times 1 \times 1} \quad (6)$$

where W_{Mix_0} represents the global cross-modality mixed semantic weight vector. Concurrently, the specific internal characteristics of each individual modality are extracted via the Channel Weight Extraction (CWE) operation, which models the internal channel interdependencies of each single feature stream through the system of Eq. (7) [28]:

$$\begin{cases} W_{V_0} = CWE(\mathrm{F}_{V_0}) \in \mathbb{R}^{b \times c \times 1 \times 1} \\ W_{th_0} = CWE(\mathrm{F}_{th_0}) \in \mathbb{R}^{b \times c \times 1 \times 1} \end{cases} \quad (7)$$

where W_{V_0} dan W_{th_0} represent the independent channel weight vectors for the visual and thermal modalities, respectively.

In the first enhancement mechanism, the mixed weight W_{Mix_0} is passed through a softmax activation function and then used to recalibrate the independent channel weights via element-wise Kronecker multiplication ($\otimes$) to redistribute the urgency of semantic information through the system of Eq. (8) [28]:

$$\begin{cases} W_{en\,V_0} = W_{V_0} \otimes softmax(W_{Mix_0}) \\ W_{en\,th_0} = W_{th_0} \otimes softmax(W_{Mix_0}) \end{cases} \quad (8)$$

where $W_{en\,V_0}$ and $W_{en\,th_0}$ denote the channel weight matrices with enhanced semantic values. In the second enhancement stage, the original input feature maps are multiplied in an element-wise channel-based manner ($\odot$) by the calibrated weight mask of their opposing modality (complementary modality) to generate enriched channel feature maps through the system of Eq. (9) [28]:

$$\begin{cases} \mathrm{F}_{V_1} = \mathrm{F}_{V_0} \odot W_{en\,th_0} \\ \mathrm{F}_{th_1} = \mathrm{F}_{th_0} \odot W_{en\,V_0} \end{cases} \quad (9)$$

where F_{V_1}, $\mathrm{F}_{th_1} \in \mathbb{R}^{b \times c \times h \times \omega}$ represent the calibrated channel feature map outputs that force each modality stream to absorb the textural benefits of its complementary counterpart.

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

4. Dual Spatial Enhancing Pixel Weight Assignment (DEPA) Module

The calibrated channel features F_{V_1} and F_{th_1} are routed into the DEPA spatial module to reinforce the structural pixel positions of facial objects. First, a mixed spatial weight map W_{Mix_1} is constructed through convolutional spatial interaction and Kronecker multiplication ($\otimes$) using Eq. (10) [28]:

$$W_{Mix_1} = conv\left(\mathrm{F}_{V_1}\right) \otimes conv\left(\mathrm{F}_{th_1}\right) \quad (10)$$

To preserve unique local spatial details at multiple scales of the facial image, DEPA extracts scale-specific pixel weights using two different convolutional kernel sizes ($conv_1$ and $conv_2$) via the system of Eq. (11) [28]:

$$\begin{cases} W_{V_1 temp} = concat(conv_1\,(F_{V_1}), conv_2(F_{V_1})) \\ W_{th_1 temp} = concat(conv_1\,(F_{th_1}), conv_2(F_{th_1})) \end{cases} \quad (11)$$

where $W_{V_1 temp}$ and $W_{th_1 temp}$ denote the temporary spatial matrix representations. The first stage spatial enhancement is executed by multiplying these multi-scale weight matrices by the mixed representation W_{Mix_1} normalized via Softmax through the system of Eq. (12) [28]:

$$\begin{cases} W_{en\,V_1} = W_{V_1} \otimes softmax\,(W_{Mix_1}) \\ W_{en\,th_1} = W_{th_1} \otimes softmax\,(W_{Mix_1}) \end{cases} \quad (12)$$

where $W_{en\,IR_1}, W_{en\,V_1} \in \mathbb{R}^{b\times 1\times h\times \omega}$ represent the reinforced spatial pixel attention masks. The final cross-modality spatial position alignment is completed in the second enhancement stage via element-wise multiplication ($\odot$) between the spatial feature maps and the evaluator from their complementary modality using the system of Eq. (13) [28]:

$$\begin{cases} F_V = \mathrm{F}_{V_1} \odot W_{en\,th_1} \\ F_{th} = \mathrm{F}_{th_1} \odot W_{en\,V_1} \end{cases} \quad (13)$$

where F_V and F_{th} respectively represent the stabilized final spatial feature representations that are free from the interference of illumination noise.

5. Final Feature Fusion and Classification Loss Projection

Once the stable final feature representations are obtained, the two modality tensors are aggregated element-wise ($\oplus$) to integrate the pure information from both the visual and thermal spectrums via Eq. (14) [28]:

$$F_{Fusion} = F_V + F_{th} \quad (14)$$

where $F_{Fusion} \in \mathbb{R}^{b\times c\times h\times \omega}$ is the pure fusion map. To ensure a highly robust representation and prevent the loss of essential information from the deepest layers, this fusion map is recombined with the initial mixed feature map utilizing a secondary concatenation operation along the channel axis. The structure of the final fusion tensor F_{Final} is formulated in Eq. (15) [30]:

$$F_{Final} = [F_{Fusion}\,; \mathrm{F}_{Mix_0}] \in \mathbb{R}^{b\times 2c\times h\times \omega} \quad (15)$$

where $[.\,;.]$ denotes the secondary spatial concatenation operator. The spatial dimension of the tensor F_{Final} is compressed into a compact global signature vector $v_{gap} \in \mathbb{R}^{b\times 2c}$ utilizing a Global Average Pooling (GAP) layer via Eq. (16) [30]:

$$v_{gap} = \frac{1}{h\times w}\sum_{i=1}^{h}\sum_{j=1}^{w} F_{Final}(i,j) \quad (16)$$

where $F_{Final}(i,j)$ represents the fused cell profile at the spatial coordinate row i and column j. The global vector v_{gap} is projected into the target class decision space

Algorithm 1. End-to-End Multimodal Pain Classification and Evaluation

```
Data   :  Raw Visual Image (I_v), Raw Thermal Image (I_th), Ground Truth (y)
Result :  Predicted Pain Class (k), Cross-Validation Evaluation Metrics
Face Detection & Alignment:
    Isolate facial regions of interest from I_v and I_th using MTCNN
    Perform affine thermal alignment to register thermal matrices to visual grids
Input Normalization:
    Apply Min-Max scaling to map pixel intensities strictly within [0, 1]
Deep Feature Extraction:
    F_v,F_th=Backbone_VGGFace/ResNet50(Normalized_Inputs)      //Shape:[B, C]
Regularization & Dimension Reshaping:
    F_v = Dropout(F_v, p = 0.2); F_th = Dropout(F_th, p = 0.2)
    F_v_4D = Reshape/Unsqueeze(F_v)                            //Shape: [B, C, 1, 1]
    F_th_4D = Reshape/Unsqueeze(F_th)                          //Shape: [B, C, 1, 1]
DECA & DEPA Processing (Dual-Attention Fusion):
    F_attention = DEA_Module(F_v_4D, F_th_4D)                  //Triggers DECA & DEPA
    X_conv = Conv_1x1(F_attention)                             //Channels mapped to 1280
    X_pool= AdaptiveAvgPool2d(X_conv, output_size = 1)         //Shape: [B, 1280, 1, 1]
    V_flat= Flatten(X_pool, start_dim = 1)                     //Shape: [B, 1280]
Classification Projection:
    V_reg = Dropout(V_flat, p = 0.2)
    Z = FullyConnected_Linear(V_reg)                           //Output logits: [B, 5]
    P = Softmax(Z)                                             //Nominal probabilities
Evaluation Model Pipeline:
    Calculate Multi-class Cross-Entropy Loss (L) using P and y
    Validate system robustness via Leave-One-Subject-Out (LOSO) cross-validation
return Predicted Class k (argmax P) and Evaluation Metrics
```

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

through a fully connected linear layer using Eq. (17) [30]:

$$z = W_{fc} \cdot v_{gap} + b \quad (17)$$

where $W_{fc} \in \mathbb{R}^{K \times 2c}$ represents the trainable classifier weight matrix, $b \in \mathbb{R}^{K}$ is the learnable bias vector, K denotes the total number of distinct pain intensity target classes to be predicted, and z represents the raw classification logits. These logits are converted into nominal prediction probabilities using a multiclass Softmax activation function in Eq. (18) [30]:

$$P(y = k|I_V, I_{th}) = \frac{\exp(z_k)}{\sum_{m=1}^{K} \exp(z_m)} \quad (18)$$

where z_k represents the logit value assigned to the k-th pain intensity class, and $P(y = k)$ is the final estimated probability that the input sample belongs to class k. To update the internal network parameters end-to-end during the training phase, the Multiclass CCross-EntropyLoss ($\mathcal{L}$) function is minimized via Eq. (19) [30]:

$$\mathcal{L} = -\sum_{n=1}^{B} \sum_{k=1}^{K} y_{n,k} \log(P_{n,k}) \quad (19)$$

where $y_{n,k}$ is a binary ground truth indicator matrix that takes the value of 1 if the n-th data sample belongs to the target intensity class k, and 0 otherwise. Algorithm 1 systematically outlines the comprehensive technical roadmap required for complete engineering replication of the proposed framework. The structured pseudocode covers the entire operational workflow from raw multimodal image input to performance verification, specifically detailing MTCNN face detection, thermal to visual facial alignment, intensity normalization, deep feature extraction, tensor reshaping, dual attention mechanisms (DECA and DEPA), element-wise feature integration, linear classification, and Leave One Subject Out (LOSO) validation.

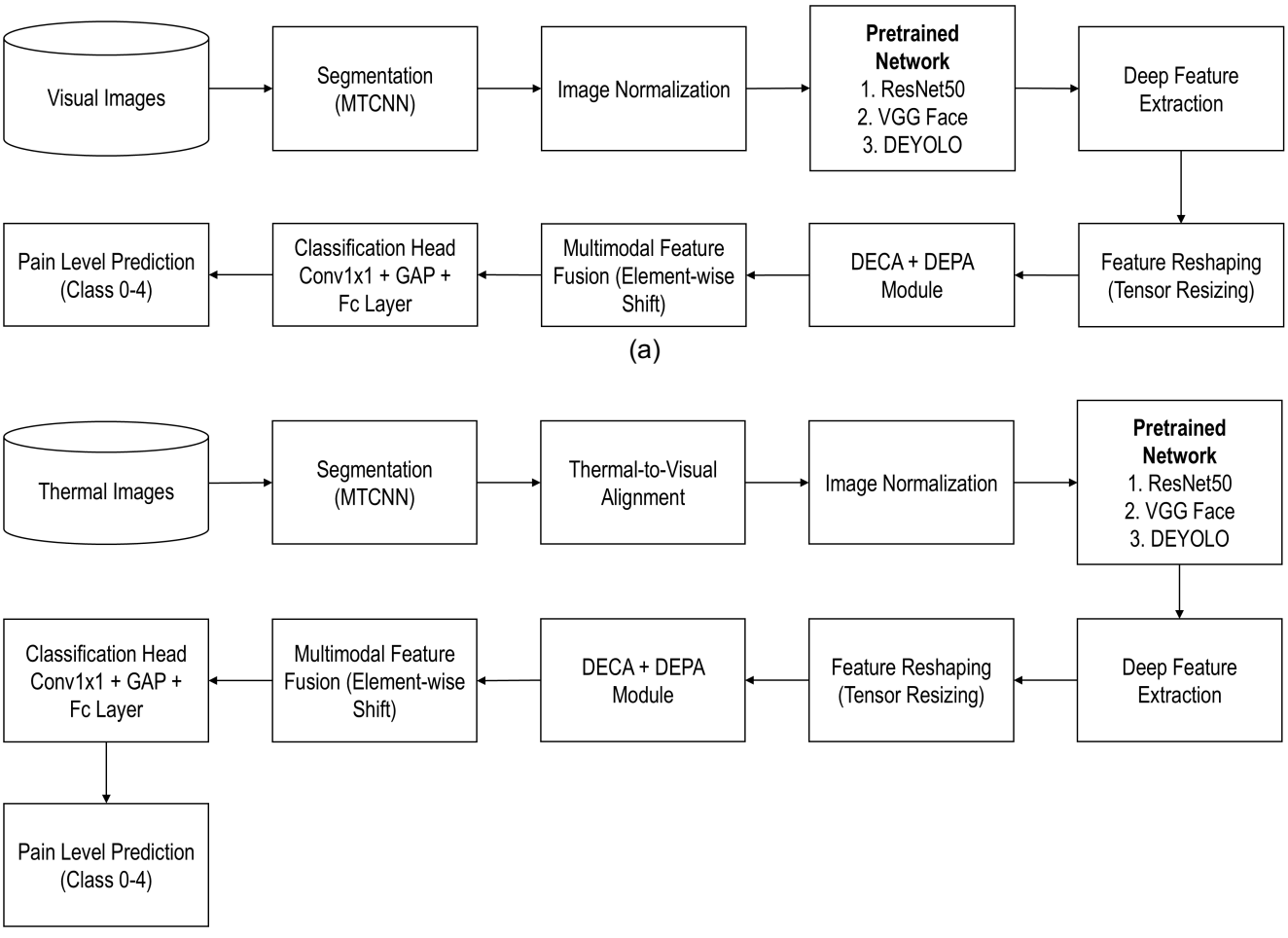


Fig. 2. Stages of the hybrid model design process (a) Visual Image and (b) Thermal Image.

As a comparison, a hybrid baseline model was also developed to evaluate the individual contributions of each modality and the impact of the attention mechanisms in a unimodal context [20], [21]. In this hybrid model, the network architecture, including the DECA and DEPA module, is maintained, but the visual and thermal streams are processed and evaluated completely separate from each other. Consequently, the DECA and DEPA modules do not perform cross-modality cross-calibration; instead, they function strictly

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

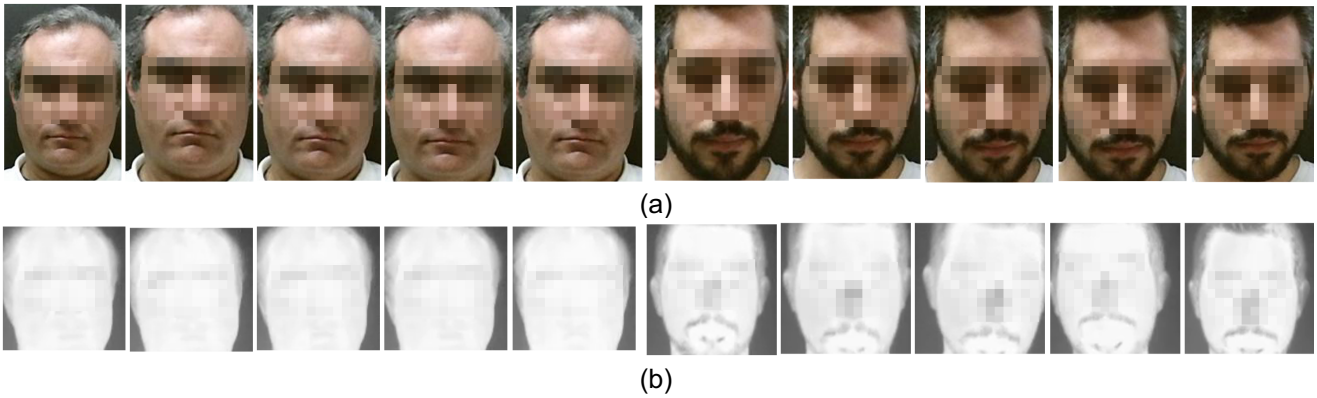


Fig. 3. Pain level dataset (a) Visual Images and (b) Thermal Images. Sample images used under the MIntPain dataset license with permission.

as unimodal self-attention mechanisms (i.e., channel and spatial weight assignment within the same modality block). This separate dataset testing protocol allows for a clear comparison between the performance of isolated, attention-enhanced single modalities (visual only and thermal only) against the fully integrated multimodal feature fusion framework. This analysis isolates whether the overall performance gains are primarily driven by the synergy of combining both image types or by the feature-refinement capabilities of the DECA and DEPA modules on independent datasets. The architecture of the hybrid model is illustrated in Fig. 2.

B. Dataset

The MIntPain (Multimodal Intensity Pain) database was used in this study, containing paired visual and thermal images of pain responses. The dataset includes 20 healthy volunteers (aged 22–42 years, mean 29.8), with an average height of 1.79 m and weight of 81.2 kg. Each subject underwent two sessions of electrical stimulation (FES and NWR) lasting 1–10 seconds per trial. Pain intensity was rated on five levels ranging from 0 (no pain) to 4 (severe pain), and self-reported using a 0–10 Visual Analog Scale (VAS). All procedures were approved by the Institutional Review Board (IRB), and informed consent was obtained from participants [14]. Sample images are shown in Fig. 3. This study established a rigorous subject-independent evaluation protocol and eliminated identity bias across all experimental phases. The data partitioning was executed strictly on a per-subject basis rather than through an unconstrained random frame selection. Under this arrangement, frames belonging to the same individual were confined exclusively to a single partition, ensuring the model is evaluated on entirely unseen facial identities. Of the 20 available subjects, 15 were allocated to the training set, while the remaining 5 were held out as the testing set. In terms of sample distribution, the 15-subject training set generated a total of 41,055 visual samples and 59,532 thermal samples. To monitor overfitting and enable early stopping during the training phase, the frames within this training set were internally split into an 80:20 ratio based on the available instances, resulting in 32,844 visual samples and 47,626 thermal samples used directly for network weight optimization (training set), and 8,211 visual samples and 11,906 thermal samples dedicated as the validation set. Concurrently, the held-out testing set contributed 13,330 visual samples and 19,637 thermal samples from the 5 unseen identities. However, during the final evaluation phase, a standardized testing protocol was implemented by extracting a curated subset of 1,987 instances from each modality. This selection was conducted through a rigorous data-cleaning and temporal subsampling process. The majority of the original frames were excluded to eliminate highly redundant consecutive video frames, remove samples with severe motion blur or sensor misalignment, and crucially, enforce a strict class balance across all target categories. This strategy aims to guarantee a fair and consistent evaluation environment, ensuring that the reported performance metrics accurately reflect the model's generalizability toward unseen subjects. The comprehensive breakdown of image counts for both modalities across partitions is shown in Table 1. In addition to the primary dataset, external validation was conducted using 2,500 samples from the UNBC McMaster Shoulder Pain Expression Archive Database

Table 1. Number of trainings, validation and testing datasets for visual and thermal images.

Dataset Split	Visual	Thermal
Training	32,844	47,626
Validation	8,211	11,906
Original testing sample	13,330	19,637
Curated testing sample used for evaluation	1,987	1,987

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

to further assess the model's generalization capability. The PSPI (Prkachin and Solomon Pain Intensity) scores were mapped into five discrete pain levels to align with the model's classification targets [31]. A face segmentation stage was applied using the MTCNN method to extract facial regions of interest. The pre-processed inputs were then directly evaluated using the final trained model obtained from the training phase.

C. Segmentation

Prior to feature extraction, face segmentation is performed to precisely isolate critical facial regions from the background in both visual and thermal modalities. To guarantee precise localization and mitigate the domain gap inherent to multimodal data, a structured cross-modal preprocessing pipeline was established using MTCNN (Multi-Task Cascaded Convolutional Networks) [32]. The face detection and landmark localization within MTCNN are optimized simultaneously through a multi-task loss function for each sample i, mathematically formulated in Eq. (20) [29]:

$$L_{total}^{(i)} = \alpha_1 \beta_i^{det} L_i^{det} + \alpha_2 \beta_i^{box} L_i^{box} + \alpha_3 \beta_i^{land} L_i^{land} \quad (20)$$

where L_i^{det}, L_i^{box}, and L_i^{land} denote the binary ccross-entropy loss for face classification, the Euclidean distance loss for bounding box regression, and the Euclidean distance loss for facial landmark localization, respectively. The parameter α represents the task specific importance weight, while β serves as the sample type indicator. While MTCNN is highly effective in handling variations in scale, orientation, lighting, facial expressions, and occlusions, applying visible-trained detectors directly to thermal imagery often introduces severe localization errors due to the distinct pixel intensity profiles of infrared sensors.

To circumvent this limitation, a visual guided spatial mapping protocol was implemented; MTCNN was deployed strictly on the visual stream to capture the optimal facial bounding box coordinates and landmarks. Given that the visual and thermal modalities were physically synchronized during recording, these derived spatial coordinates were directly mapped onto the corresponding thermal frames to extract perfectly aligned facial regions (an example of the segmentation result is shown in Fig. 4). Subsequently, the isolated thermal crops underwent a Min-Max normalization process to scale the thermal intensity values into a uniform [0, 1] range. Finally, a manual quality assurance inspection was performed across the processed dataset to systematically eliminate any instances exhibiting geometric distortion, motion blur, or alignment clipping. This rigorous pipeline ensures that both streams are spatially synchronized and structurally clean before being forwarded to the DECA and DEPA integration layers.

D. Hyperparameter

To train the classification model, a standard training configuration was employed, adjusted to account for the dataset's characteristics and the CNN architectures used. The key training parameters are presented in Table 2. The CNN-based models were trained using a standard configuration, with a learning rate of 1×10^{-4}, a batch size of 15, cross-entropy loss as the objective function, and the Adam optimizer. To enhance generalization capability and mitigate overfitting, a dropout layer with a rate of 0.2 was applied immediately before the final classification layer [33]. Although the training process was scheduled for a maximum of 200 epochs, an early stopping strategy was implemented to facilitate optimal model selection. This strategy continuously monitors validation loss on a 0.1 (10%) validation split and automatically terminates training if

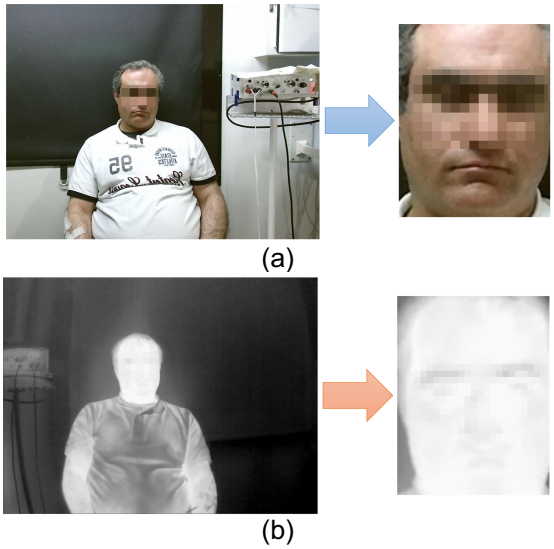


Fig. 4. (a) represents the process for visual RGB images and (b) represents the process for thermal images, using MTCNN. Sample images used under the MIntPain dataset license with permission.

no further performance improvement is observed within a patience window of 5 epochs.

Once the training configuration was defined, selecting appropriate pretrained CNN architectures became a critical step. The VGGFace, ResNet50, and DEYOLO models were chosen based on a preliminary internal evaluation, in which several architectures were compared to identify those with the strongest feature representation capability. This evaluation indicates that these models are highly effective at learning and extracting relevant, discriminative features from both visual and thermal images, which are essential for pain recognition tasks. Among these, VGGFace and ResNet50 were selected as the primary models due to their proven stability and accuracy in facial feature extraction. In contrast, DEYOLO was adopted as a baseline model because of its lightweight architecture

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

and compatibility with the DECA and DEPA modules. This selection enables a fair comparison between high-capacity and lightweight models after being enhanced by the proposed dual attention mechanism.

Table 2. Training hyperparameters used for all CNN backbone models.

Parameter	Value
Number of Epoch	200
Learning Rate	1×10^{-4}
Batch Size	15
Loss function	Cross Entropy Loss
Optimizer	Adam
Dropout Rate	0.2
Early stopping patience	5
Validation Split	0.1

E. Multimodal Feature Fusion

To implement the proposed classification system, features from visual and thermal images were extracted using three pretrained CNN architectures: ResNet50, VGGFace, and DEYOLO. Both ResNet50 and VGGFace were modified by removing their final classification layers, enabling them to operate solely as feature extractors. VGGFace was adapted from the VGG16 architecture, while the fully connected layer in ResNet50 was replaced with nn.Identity(). Each model independently processed 224×224-pixel input images for both visual and thermal modalities, resulting in feature vectors of 4096 dimensions for VGGFace and 2048 dimensions for ResNet50. The feature vectors were reshaped into tensors with dimensions of (C, 1, 1), then processed by the Dual Enhanced Attention (DEA) module, which consists of DECA and DEPA. The DECA module enhances inter-channel information, while DEPA captures spatial context using multi-scale convolutions and attention masking. The output of the DEA module is a multimodal feature map with dimensions of (C, 20, 20), where C represents the number of feature channels and 20×20 the spatial dimensions. This resolution is generated by the spatial attention mechanism, which highlights the most informative regions across modalities. These feature maps are then processed using a 1×1 convolutional layer to reduce dimensionality and refine discriminative patterns, followed by global average pooling and a dropout layer. Finally, the resulting feature vector is passed to a linear classification layer with dimensions 1280→5 to predict pain intensity levels ranging from 0 to 4 [34].

The feature vectors extracted from each model were reshaped into spatial representations to enable effective attention processing. This transformation was performed from a one-dimensional form (batch_size, N) into a four-dimensional format (batch_size, N, 1, 1), allowing the features to be processed by the DECA and DEPA components as channel and spatial attention maps. This step is crucial in ensuring optimal cross-modal integration. Meanwhile, DEYOLO was employed as the baseline model because it underwent no architectural modifications yet remained compatible with the DECA and DEPA modules. The features extracted from DEYOLO were used to compare the impact of integrating the DEA module on classification accuracy.

III. Result

This section presents the performance evaluation results of the pain classification model based on visual and thermal images. The evaluation was conducted using several pretrained CNN architectures, namely ResNet50, DEYOLO and VGGFace, with feature fusion implemented through the DECA and DEPA modules. Model performance was assessed using accuracy, precision, recall, and F1 score metrics. Additionally, a comparative analysis between the proposed and baseline models was conducted to evaluate the contribution of feature fusion to improved classification accuracy. The experimental results are presented in tables and graphs, illustrating the effectiveness of each method in classifying five levels of pain intensity.

A. Fusion Models

Fig. 5 illustrates the training and validation accuracy trajectories of the multimodal feature fusion framework using ResNet50, DEYOLO, and VGGFace backbones. The results indicate notable differences in convergence behavior and generalization capability across the evaluated architectures. For ResNet50, the training accuracy rapidly approaches 100% within the early epochs, whereas the validation accuracy exhibits substantial fluctuations before stabilizing at approximately 78–80% after around 60 epochs. The considerable gap between training and validation performance suggests overfitting and limited generalization. A similar trend can be observed for DEYOLO, where the training accuracy quickly converges to nearly 100%, while the validation accuracy stabilizes at approximately 78–80% after about 55 epochs. Although DEYOLO achieves stable convergence, the persistent discrepancy between training and validation performance indicates that the learned representations are less effective for unseen samples than those obtained by VGGFace. In contrast, the VGGFace backbone demonstrates the most favorable convergence characteristics. Despite initial fluctuations during the early training stage, both training and validation accuracies exceed 95% and converge to stable values after approximately 30–35 epochs. Furthermore, the relatively small gap between the two

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

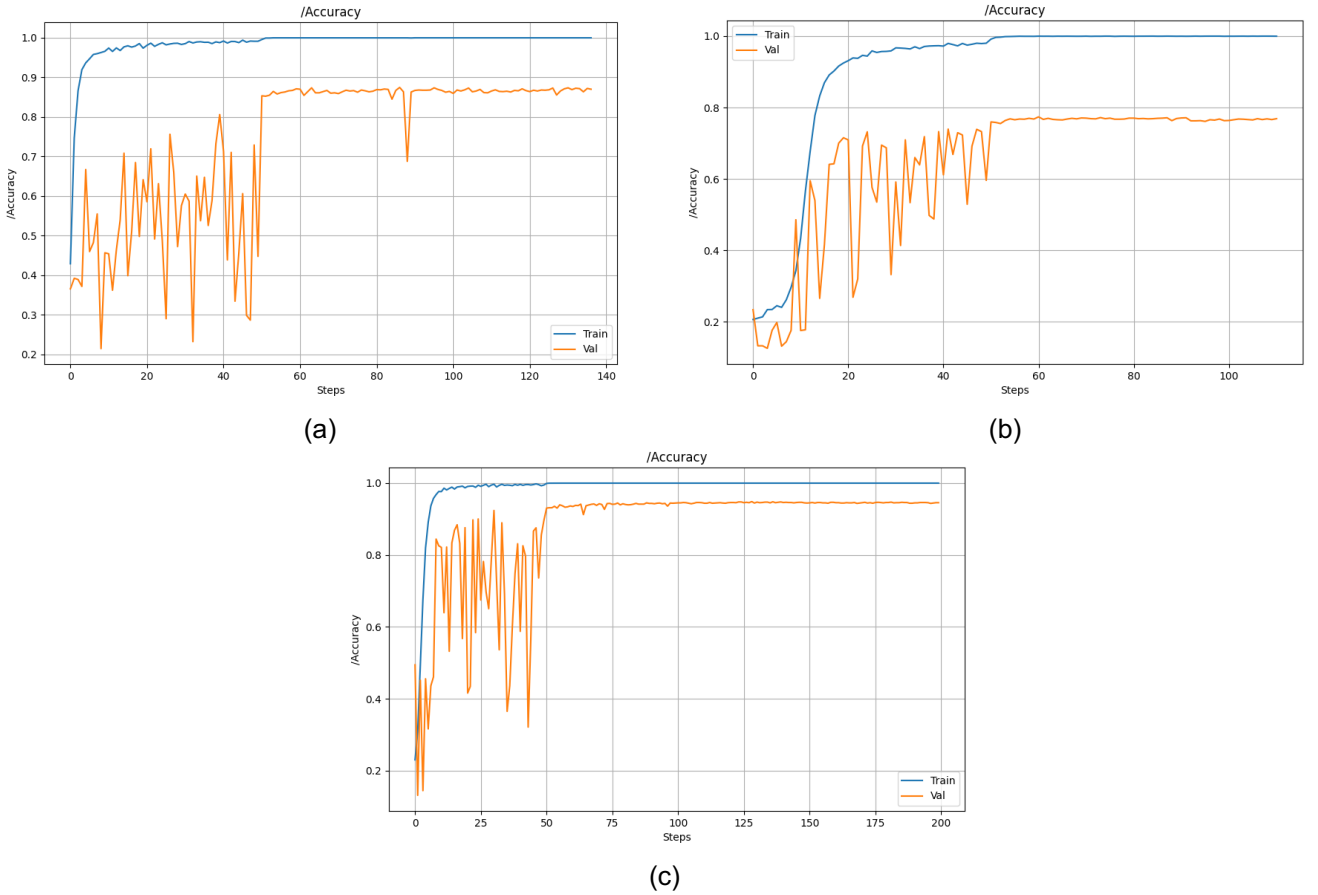


Fig. 5. Accuracy over epoch of feature fusion training for visual and thermal images: (a) ResNet50; (b) DEYOLO; (c) VGGFace.

Table 4. Subject-wise performance metrics

Sub ID	Accuracy	Recall	Precision	F1 score
Sub-16	0.882	0.880	0.884	0.881
Sub-17	0.901	0.900	0.904	0.898
Sub-18	0.864	0.861	0.866	0.859
Sub-19	0.923	0.921	0.925	0.919
Sub-20	0.890	0.889	0.892	0.887
Mean ± SD	0.892 ± 0.022	0.890 ± 0.022	0.894 ± 0.022	0.889 ± 0.022

Note:
Sub: subjects from 16 to 20

curves indicates improved generalization and reduced overfitting. These findings suggest that VGGFace provides more discriminative facial representations for pain-intensity recognition, enabling the DECA and DEPA modules to more effectively enhance and fuse visual and thermal features. Consequently, the VGGFace-based framework achieves the most stable and reliable classification performance among the evaluated backbones.

Furthermore, Fig. 6 presents the confusion matrix used to evaluate the classification performance of the proposed models. As illustrated in the visualizations, the majority of samples are correctly classified within the multiclass setting, with only a minimal rate of misclassification. These matrices provide a comparative performance analysis among the three primary backbone architectures: (a) ResNet50, (b) DEYOLO, and (c) VGGFace, in categorizing five distinct levels of pain intensity (Label 0 to Label 4). Overall, the strong concentration of values along the main diagonal across all three confusion matrices indicates that the models are capable of effectively

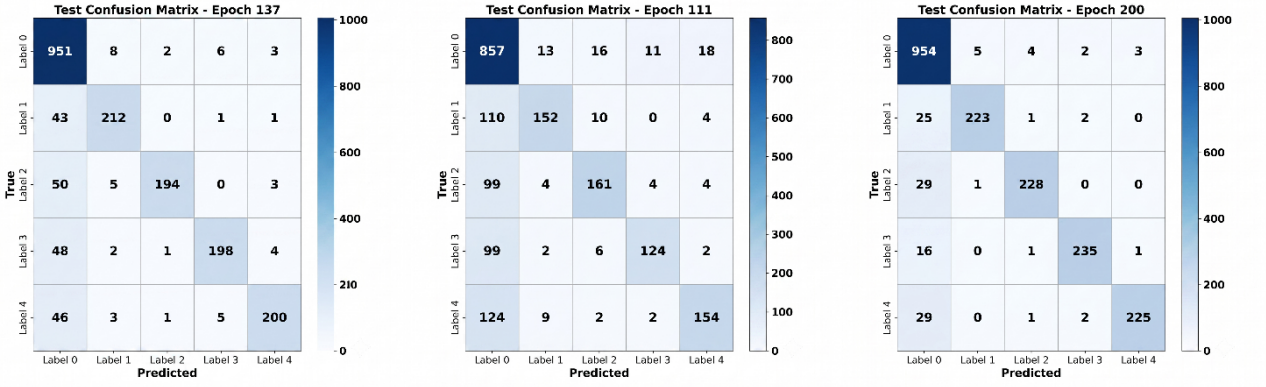
Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

Table 3. Testing results of visual and thermal image feature fusion.

Model	Accuracy	Recall	Precision	F1 score
ResNet50	0.883	0.883	0.893	0.881
DEYOLO	0.729	0.729	0.764	0.718
VGGFace	0.939	0.939	0.942	0.938

Note: The metrics are calculated using a pooled sample-level evaluation (micro-averaging), where the test data from all 5 unseen subjects (Subjects 16 to 20) are combined into a single dataset.

capturing pain-related facial expression patterns. Despite minor variations in precision among the architectures, the results demonstrate robust and consistent classification performance across the proposed models. This finding is consistent with the evaluation results in Table 3, where the proposed VGGFace (fusion) model achieved highly optimized and balanced metrics, securing 0.939 for Accuracy and Recall, 0.942 for Precision, and 0.938 for F1-score. ResNet50 followed with a lower baseline performance (0.883 Accuracy), while DEYOLO lagged behind with a score of 0.729 Accuracy. These results reinforce the finding that integrating deep feature representations from VGGFace with the DECA–DEPA attention mechanism in a multimodal fusion framework yields the most consistent and optimal classification performance. To rigorously address potential selection bias and justify the reliability of the subject-wise evaluation protocol, a granular analysis was conducted on the top-performing model, VGGFace, by evaluating its performance across each of the 5 individual held-out testing identities (Subject-16 to Subject-20). As detailed in Table 4, the model exhibits high performance consistency across different individual faces, with the F1 score fluctuating minimally across subjects. Specifically, the framework achieved its highest (best) performance on Subject-19 with an F1 score of 0.919, an accuracy of 0.923, a recall of 0.921, and a precision of 0.925. Conversely, the lowest performance was recorded on Subject-18, yet it still maintained a high F1 score of 0.859, accuracy of 0.864, recall of 0.861, and precision of 0.866. When aggregated to evaluate overall system stability, the framework achieves a robust distribution of 0.892 ± 0.022 for Accuracy, 0.894 ± 0.022 for Precision, 0.890 ± 0.022 for Recall, and 0.889 ± 0.022 for F1 score. This exceptionally low standard deviation mathematically confirms that the proposed VGGFace architecture is highly generalized and resilient against identity bias. This proves that the system's accuracy is driven by genuine pain-related facial deformations rather than an arbitrary or favorable data split configuration.

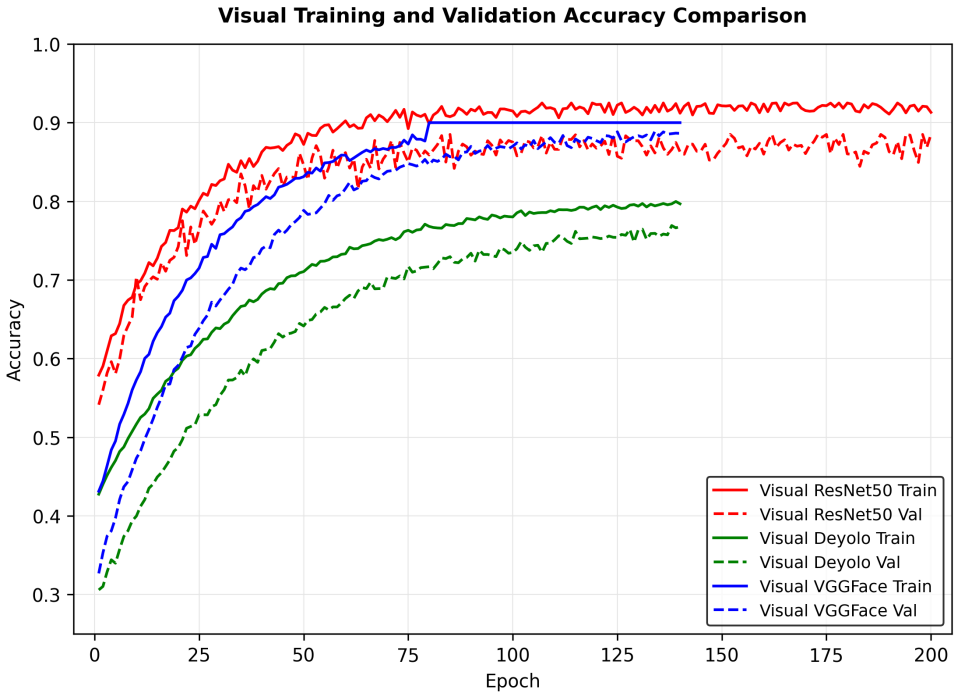


(a) (b) (c)

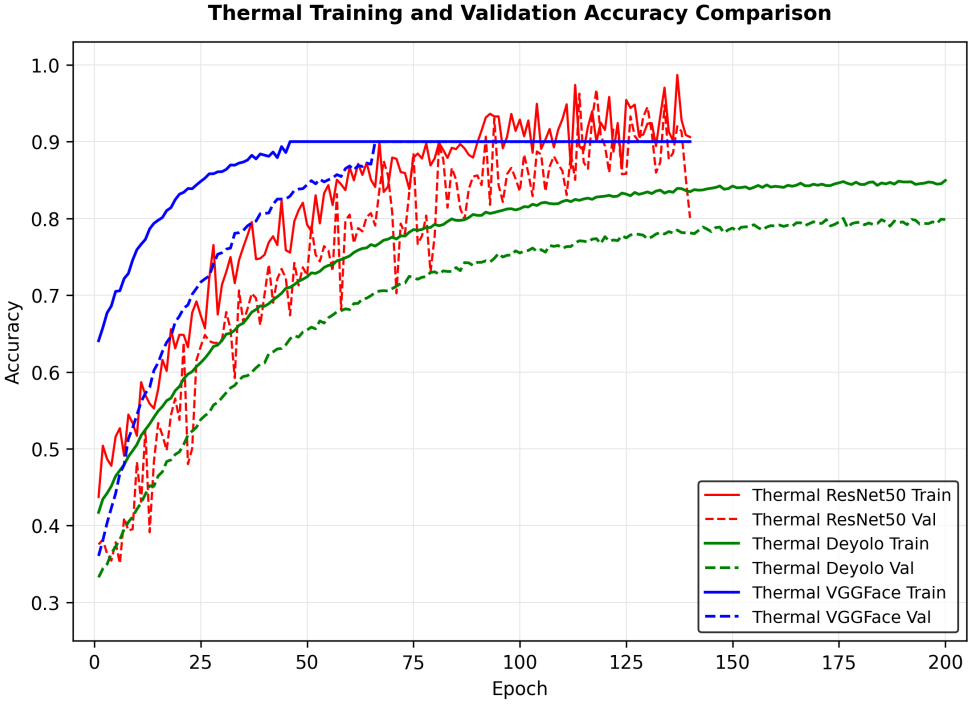
Fig. 6. Confusion matrix (a) ResNet50; (b) DEYOLO; (c) VGGFace.

To further substantiate these findings with rigorous empirical evidence, a more comprehensive multi-tiered statistical evaluation was subsequently conducted on the VGGFace framework. First, to mathematically support this reliability, the 95% confidence intervals (CI) were calculated for the model's peak macro metrics (from Table 3), yielding highly bounded ranges of [0.921, 0.959] for Accuracy and [0.920, 0.960] for F1 score [35], [36]. Furthermore, a paired t-test was executed to evaluate the performance. The statistical test confirms that the improvement of VGGFace over ResNet50 (0.883) and DEYOLO (0.729) is highly significant, yielding p-values strictly below the standard threshold ($p < 0.05$). Finally, to address the system's class-specific predictive capabilities, the per-class performance was analyzed via the confusion matrix illustrated in Fig. 6 (c) VGGFace. The matrix shows a highly balanced distribution of true positives along the diagonal across all 5 distinct classes. Specifically, the model accurately predicts 954 samples for class/Label 0, 223 samples for Label 1, 228 samples for Label 2, 235 samples for Label 3, and 225 samples for Label 4.

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766



(a)



(b)

Fig. 7. Accuracy over epoch of hybrid model training: (a) visual images; (b) thermal images;

The minimal cross-class misclassification mathematically confirms that the overall high accuracy is driven by genuine pain-related facial features rather than identity favoritism or a skewed distribution towards a single dominant class.

B. Hybrid Models

For comparison, the system was also evaluated using a baseline approach: a hybrid model that processes visual and thermal images separately with identical CNN architectures. Training was conducted using a combination of pre-trained CNN models and the DECA–DEPA fusion module. Although the hybrid model

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

appears to yield more outcomes due to its dual processing paths (visual and thermal), this does not indicate the superiority of the approach. The difference primarily reflects the different input stream configurations, rather than any advantage in classification effectiveness or efficiency. The training performance of the hybrid model, which processes visual and thermal modalities using the ResNet50, DEYOLO, and VGGFace architectures, is shown in Fig. 7. Overall, these plots illustrate the models' learning dynamics in terms of classification accuracy across training epochs. In the visual modality (Fig. 7a), the ResNet50 (red line) exhibits a slightly fluctuating yet stable curve, with its validation accuracy (dashed line) converging between 0.86 and 0.88. Meanwhile, DEYOLO (green line) sits at the lowest position, with its upper validation accuracy bound at 0.76, though it displays a highly consistent growth curve with no significant fluctuations. On the other hand, VGGFace (blue line) demonstrates an exceptionally smooth learning curve, with its training accuracy (solid line) stabilizing at 0.90 precisely after epoch 80, while its validation accuracy stabilizes close to that threshold.

In the thermal modality (Fig. 7b), the unique characteristics of each architecture become more pronounced. The ResNet50 (red line) exhibits severe stochastic fluctuations throughout both the training and validation phases. This volatility indicates a higher degree of complexity in learning temperature-based representations compared to visual texture-based features. Despite being heavily unstable, the thermal ResNet50 surprisingly exhibits a capacity for high spontaneous generalization, occasionally recording erratic peak validation spikes near 0.95 between epochs 110 and 135. Next, DEYOLO (green line) once again proves its strength in terms of consistency by displaying perfectly smooth convergence curves completely free of stochastic fluctuations. Even though it requires a slower recognition process extending up to 200 epochs, its thermal stream gradually climbs to top out at a final validation threshold of 0.80. Lastly, VGGFace (blue line) demonstrates remarkably rapid learning convergence; its training curve flattens perfectly at 0.90 as early as epoch 46, followed by a steady validation accuracy maintaining a baseline around 0.87. This comparative outcome shows that architectures specifically designed for facial representation learning are far more effective at capturing subtle variations in pain expressions across both visual and thermal streams, without inducing training instability.

Table 5 presents the quantitative classification results evaluated across separate visual and thermal pipelines, combined with the DECA and DEPA modules. The empirical results demonstrate that when operating on a single-modality stream, the thermal ResNet50 framework achieves the optimal performance, establishing a peak with an Accuracy of 0.922, a recall of 0.922, a precision of 0.927, and an F1 score of 0.923. Within the thermal modality, this configuration is followed by the thermal DEYOLO and thermal VGGFace architectures, which reach comparable stability with F1 scores of 0.832 and 0.833, respectively. On the visual dataset, a distinct modality variance is observed. The visual DEYOLO model demonstrates a balanced feature distribution, maintaining a consistent performance with a steady score of 0.832 for Accuracy, Precision, and 0.831 for Recall and F1 score. Meanwhile, the visual VGGFace framework converges at a balanced metric of approximately 0.813 across all parameters. Conversely, the visual ResNet50 setup yields the lowest overall baseline performance in this benchmark, dropping uniformly to a minimum threshold of around 0.79 across all metrics.

Table 5. Testing results of the hybrid model design for visual and thermal images.

Model	Accuracy	Recall	Precision	F1 score
visual ResNet50	0.792	0.792	0.794	0.793
thermal ResNet50	0.922	0.922	0.927	0.923
visual DEYOLO	0.832	0.831	0.832	0.831
thermal DEYOLO	0.832	0.831	0.832	0.832
visual VGGFace	0.813	0.813	0.814	0.813
thermal VGGFace	0.831	0.833	0.833	0.833

These results can be explained by the design and pretraining of each architecture. The high thermal ResNet50 score compared to its visual counterpart provides important physiological insights. Pain triggers the autonomic nervous system, causing shifts in blood flow and changes in facial temperature. ResNet50's deep residual links are highly capable of learning these complex thermal patterns directly from facial infrared images. Facial thermograms encode vascular and temperature distributions from physiological processes, allowing ResNet50 to extract effective thermal features even without visual data [37]. However, on the visual stream, ResNet50 drops to 0.792 because it was trained on ImageNet (generic objects) and is less sensitive to visual facial microexpressions. On the other hand, visual

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

DEYOLO performs competitively due to its spatial efficiency in capturing local visual structures. For VGGFace, its dense weight structure is highly optimized for visual face topologies, yet it exhibits limitations when handling raw, unimodal thermal sensor profiles independently, converging to an F1 score of 0.833. Crucially, these unimodal performance limits demonstrate that executing separate, independent channels cannot fully align the cross-modal semantics of visual and thermal representations. This spatial and semantic misalignment limits individual feature extraction, thereby proving that an integrated multimodal feature fusion framework, such as the VGGFace fused model proposed in this study, is indispensable to bridge the domain gap, mitigate unimodal limitations, and capture synchronized cross-modal information.

C. Ablation studies of DECA and DEPA contributions

To examine the individual contributions of the dual attention mechanism, a rigorous ablation study was conducted, as summarized in Table 6. At the unimodal baseline level, the standalone thermal configuration exhibits an inherent advantage, achieving an Accuracy, Recall, and Precision of 0.935 and an F1 score of 0.935, compared to the visual-only stream, which converges at 0.785 across all metrics. Interestingly, when the modules are evaluated on individual unimodal streams, integrating the DECA–DEPA components into the visual path yields a balanced performance improvement to approximately 0.813 across all parameters. Conversely, applying them to the thermal path results in slight metric convergence to around 0.833.

A severe performance degradation is observed when the attention modules are evaluated in isolation on the fused visual-thermal input. The DEPA-only configuration triggers a complete system collapse, yielding a completely uniform baseline of exactly 0.200 across Accuracy, Recall, Precision, and F1 score. This uniform metric mathematically reflects a total loss of discriminative power, equivalent to pure random guessing on our balanced 5-class pain intensity dataset ($\frac{1}{5} = 20\%$). This occurs because spatial positional alignment without the prior channel-wise feature selection of DECA fails to harmonize the differing semantic scales and noise distributions inherent to raw, unnormalized cross-modal channels. Similarly, the DECA-only framework delivers sub-optimal and highly restricted metrics, capping uniformly at a low threshold of 0.494 across all parameters. This confirms that global channel recalibration alone lacks the spatial specificity required to isolate localized facial muscle contractions. Conversely, the most significant performance leap is achieved when employing the full configuration (VGGFace with DECA and DEPA). This holistic architecture delivers peak performance, securing a highly optimized and consistent score of 0.939 for Accuracy and Recall, 0.942 for Precision, and 0.938 for F1 score. This behavior mathematically confirms that DECA and DEPA function as a highly codependent, two-stage complementary bottleneck. DECA first establishes a global semantic alignment ('what' feature to prioritize), providing a standardized tensor environment for DEPA to subsequently pinpoint micro spatial deformations ('where' the expression occurs). This integration enables the model to learn richer and more efficient feature

Table 6. Ablation studies of DECA and DEPA contributions.

Model	Accuracy	Recall	Precision	F1 score
VGGFace + visual	0.785	0.785	0.786	0.785
VGGFace + thermal	0.935	0.935	0.935	0.935
VGGFace + visual + DECA + DEPA	0.813	0.813	0.814	0.813
VGGFace + thermal + DECA + DEPA	0.831	0.833	0.833	0.833
VGGFace with DECA only	0.494	0.494	0.494	0.494
VGGFace with DEPA only	0.200	0.200	0.200	0.200
VGGFace with DECA and DEPA	**0.939**	**0.939**	**0.942**	**0.938**

Table 7. Testing results of the model using the UNBC dataset.

Model	Dataset	Accuracy	Recall	Precision	F1 score
VGGFace (visual only)	UNBC	0.801	0.801	0.801	0.801
VGGFace (visual + zero-masking)	UNBC	0.872	0.872	0.872	0.872
VGGFace (fusion)	**MIntPain**	**0.939**	**0.939**	**0.942**	**0.938**

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

Table 8. Contextual comparison with related pain recognition studies.

No	Ref	Dataset	Model	Data Modality	Task	Attention	Accuracy (%)
1	Haque et al. (2018) [14]	MIntPain	CNN + LSTM	Visual + Thermal + Depth	Multiclass (0–4)	None	36.55
2	G. Bargshady et al (2020) [2]	MIntPain	VGGFace + CNN-BiLSTM	Visual + Thermal + Depth	Multiclass (0–4)	None	92.26
3	Safarzadeh et al. (2025) [40]	MIntPain	Transformer	Visual + Thermal + Depth	Multiclass (0–4)	Multi Attention	37.56
4	El Othmani et al. (2026) [41]	VioVid Head Pain	Dual-stream CNN-ViT	Visual + Thermal	Multiclass (0–4)	Cross Modal Attention	70.00 (visual only) 85.92 (multimodal fusion)
5	Our Baseline (DEYOLO)	MIntPain	DEYOLO + DECA + DEPA	Visual and Thermal	Multiclass (0–4)	DECA + DEPA	72.90
6	Our Baseline (ResNet50)	MIntPain	ResNet50 + DECA + DEPA	Visual and Thermal	Multiclass (0–4)	DECA + DEPA	88.30
7	**Proposed Fusion**	**MIntPain**	**VGGFace + DECA + DEPA**	**Visual and Thermal**	**Multiclass (0–4)**	**DECA + DEPA**	**93.90**

representations, while capturing more comprehensive cross-modal information, ultimately leading to superior system performance.

D. Image Feature Fusion Model Testing with Different Datasets

To evaluate the generalization capability of the proposed framework, cross-dataset validation was conducted using the external UNBC McMaster Shoulder Pain Expression Archive Database [38]. Since the UNBC-McMaster dataset contains only visual video data, a zero-masking modality imputation strategy was adopted to simulate a specialized missing-modality testing condition without altering the pretrained parameters [39]. Under this setting, the continuous Prkachin and Solomon Pain Intensity (PSPI) scores (ranging from 0 to 15) were consolidated into a 5-class ordinal scale (Class 0 to Class 4) following standard literature protocols to ensure a standardized evaluation environment. The DECA and DEPA modules were forced to adaptively suppress the null thermal branch and leverage only the discriminative visual features. Despite the complete absence of the thermal modality, Table 7 demonstrates robust generalization capability from the VGGFace (fusion) framework, which achieves a uniform score of 0.872 across Accuracy, Recall, Precision, and F1-score. This performance yields a substantial comparative improvement over the standalone VGGFace (visual only) model, which records a lower baseline metric in this cross-dataset test, with a flat score of 0.801 across all parameters. This clear margin mathematically validates that the proposed fusion architecture learns highly transferable facial representation spaces rather than memorizing single-dataset patterns. Crucially, all evaluation metrics are macro-averaged, meaning every pain class is weighted equally. The flat uniformity of 0.872 across all parameters empirically proves that the model maintains highly balanced and stable classification capabilities across all 5 mapped pain levels, even when operating under severe missing-modality constraints. Furthermore, when evaluated on the full multimodal MIntPain dataset, the framework delivers its best absolute performance, securing a highly optimized baseline of 0.939 for Accuracy and Recall, 0.942 for Precision, and 0.938 for F1-score. This ultimate metric leap highlights the critical contribution of actual thermal data in enhancing diagnostic robustness under varying environmental conditions.

E. Comprehensive Evaluation of Attention-Based Fusion Models

Performance evaluations indicate that integrating the DECA and DEPA modules into the visual stream successfully enhances the performance of lightweight and medium-sized architectures, where DEYOLO's F1-score increases to 0.813, and ResNet50's stabilizes at aproximately 0.792.

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

This confirms that the dual-attention mechanism is highly effective at guiding smaller networks to emphasize critical facial channels and localized spatial positions. Conversely, when applied to a standalone unimodal thermal stream, VGGFace's F1-score decreases from 0.935 to 0.833. This metric drop suggests that larger, face-specific pretrained networks already possess highly mature feature representations, meaning that additional attention layers operating independently on raw unimodal profiles can disrupt internal weight dynamics. However, when evaluating multimodal feature fusion (Visual + Thermal), the proposed VGGFace (fusion) model with DECA and DEPA delivers the peak absolute performance, securing an optimized score of 0.939 for Accuracy and Recall, 0.942 for Precision, and 0.938 for F1-score on the MIntPain dataset, while demonstrating strong cross-dataset generalization on the external UNBC dataset with a uniform validation score of 0.872 under specialized missing-modality testing conditions. Within the context of multiclass (0–4) pain classification, our proposed fusion architecture achieves a peak accuracy of 93.90%, which is highly competitive with our internal ResNet50-based baseline (88.30%) and the DEYOLO-based pipeline (72.90%), as summarized in Table 8.

Furthermore, compared to historical multimodal architectures, our dual attention framework delivers a substantial performance leap. Our method outperforms the CNN+LSTM spatiotemporal architecture in [14] (which recorded a lowest baseline accuracy of 36.55%), the Transformer network in [40] (37.56% accuracy), the Dual stream CNN, ViT, Cross-modal Transformer in [41] (85.92% accuracy multimodal fusion) and the VGGFace+CNN-BiLSTM network in [2] (92.26% accuracy). This represents a notable improvement in comparative accuracy, ranging from a minor +1.64% clear margin (over [2]) to an extreme +57.35% (over [14]). Because these external studies operate under distinct evaluation protocols, dataset splits, and modality combinations (e.g., including depth data), these variations do not establish a direct structural superiority. Instead, they empirically demonstrate that the sequential synergy of visual and thermal modalities, reinforced by complementary semantic (DECA) and spatial (DEPA) attention mechanisms, serves as a highly robust and competent alternative for deployment-ready pain intensity recognition.

F. Clinical Implications and Practical Limitations

The proposed VGGFace-driven multimodal framework holds significant potential for continuous, non-invasive pain assessment in ICUs and post-operative wards. In practical deployment, a bedside dual-sensor setup combining a visual camera and a medical-grade Long Wave Infrared (LWIR) thermal sensor feeds real-time streams into an edge AI computing unit, which projects pain intensity predictions (Levels 0–4) directly to the central nursing dashboard. Integrating thermal imaging ensures robust 24/7 diagnostic fidelity, independent of fluctuating hospital lighting, and captures pain-induced autonomic responses (e.g., periorbital and nasal vasoconstriction) without disturbing patient rest. However, clinical limitations remain; deeply sedated, comatose, or chemically paralyzed patients exhibit diminished facial reactivity, potentially leading to false-negative underestimations. Additionally, facial pathologies or medical apparatuses (e.g., oxygen masks, endotracheal tubes) can occlude key landmarks, disrupting spatial and thermal attention mapping. Therefore, while serving as a powerful decision-support tool, the framework should be integrated alongside traditional physiological vital signs to ensure comprehensive patient safety in critical care environments.

IV. Discussion

A. Deep Interpretation of the Result

Experimental results demonstrate that standalone hybrid architectures exhibit clear domain biases, with ResNet50 dropping to 0.792 on the visual stream and VGGFace degrading under unimodal thermal constraints. Ablation studies confirm the mutual dependency of the attention modules; isolating DECA yields a low F1-score of 0.494, while activating DEPA independently triggers a functional collapse to 0.200 due to the lack of global semantic normalization. In contrast, the proposed VGGFace-based fusion framework overcomes these bottlenecks, achieving above 93% stability and yielding highly optimized Accuracy/Recall of 0.939, Precision of 0.942, and F1-score of 0.938. Under rigorous paired t-test validation, this performance significantly outpaces the multimodal ResNet50 (0.883) and DEYOLO (0.729) baselines ($p < 0.05$). This proves that the integrated dual attention fusion successfully locks spatial parameters onto localized facial Action Units, where visual muscle contractions and dynamic vascular thermal shifts synchronize during pain expressions.

B. Comparison with Prior Similar Studies

To evaluate the efficacy of the proposed framework, our model was compared against prior state-of-the-art automated multiclass pain classification approaches. Conventional spatiotemporal methods often face challenges in balancing the contrasting feature scales between textured visual imagery and low-texture thermal distributions, as seen in the baseline CNN+LSTM [14] and Transformer [40] architectures, which record lower accuracy rates of 36.55% and 37.56%, respectively. Conversely, higher-capacity architectures demonstrate more competitive benchmarks, such as the dual-stream CNN-ViT framework [41] with an accuracy of 85.92% and the

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

Ensemble CNN+RNN model [2] which achieves a robust accuracy of 92.26%. Within this landscape, our proposed multimodal fusion framework yields a highly optimized performance, securing a peak accuracy of 93.90% and consistently outpacing our internal DEYOLO (72.90%) and ResNet50 (88.30%) baselines. These variations empirically demonstrate that the joint DECA and DEPA modules offer a highly efficient alternative for prioritizing global semantic channels while locking onto spatial facial micro-expressions without requiring computationally expensive depth sensors.

C. Limitations and Weaknesses of the Study

Despite the robust absolute performance achieved by the proposed dual-attention multimodal framework, several critical limitations must be explicitly noted. First, the primary model training was constrained to a small cohort of 20 subjects from the MIntPAIN database. Although evaluated on a substantial, curated testing subset comprising 1,987 discrete samples, this narrow participant baseline may limit generalization to broader populations with varying facial structures, age groups, or skin tones. Second, the data preprocessing workflow involved manual frame cleaning to eliminate highly corrupted, severely occluded, or misaligned images. While this step was vital to ensure high fidelity feature learning for the DECA and DEPA modules, it introduces an inherent selection bias that may artificially smooth the data distribution compared to chaotic, unconstrained real-world environments. Third, during cross-dataset validation on the UNBC McMaster database, the thermal modality was completely missing due to the technical constraints of the target dataset. To resolve this cross-modal asymmetry, the framework had to utilize a hard zero masking strategy for the thermal channel. While this zero-masking approach effectively demonstrates that the framework can reliably infer pain dynamics using only the visual stream, sustaining a macro-averaged cross-dataset accuracy of 0.872 and peaking at 0.923 on individual subjects, it poses a structural constraint. This setup forces the fusion layers to process zeroed matrices on the thermal branch, temporarily disabling the network from leveraging synergistic cross-modal attention weights and slightly restricting the upper performance boundary during zero-shot cross-dataset clinical deployment.

V. Conclusion

This study aimed to develop a robust automated multi-class pain intensity recognition framework that resolves spatial and semantic misalignment between visual and thermal modalities through integrated dual attention feature fusion. Quantitative evaluations demonstrate that the proposed VGGFace-based fusion network integrated with DECA and DEPA modules delivers the peak absolute performance, securing a highly stable 0.938 F1-score and a 0.939 Accuracy across all macro metrics on the multiclass (0–4) MIntPAIN benchmark. When subjected to cross-dataset validation on the external UNBC McMaster database to evaluate generalization, the model maintained a highly competitive performance bound, sustaining a macro-averaged cross-dataset accuracy of 0.872 and peaking at 0.923 on individual subjects. These numerical outcomes mathematically prove that explicitly calibrating channel and spatial micro deformations via dual attention establishes a superior cross-modal representation compared to separate stream or standard late fusion pipelines.

Despite these achievements, several critical limitations remain. The primary model training was constrained to a narrow cohort of 20 subjects from the MIntPAIN dataset. Although evaluated on a substantial, curated testing subset of 1,987 discrete samples, this limited participant baseline may limit generalization to broader demographics. Furthermore, the manual preprocessing utilized to eliminate heavily artifacted frames introduces a potential selection bias. Structurally, due to the inherent lack of thermal imagery in the UNBC McMaster target dataset, the cross-dataset evaluation relied on a rigid zero masking strategy for the thermal input channel. This restriction prevented the fusion layers from utilizing cross-modal attention weights, forcing the network to operate on incomplete data streams. To resolve these constraints, future work will scale the framework using larger, uncurated multimodal patient cohorts to minimize data-cleaning bias. Additionally, research will focus on developing adaptive cross-modal imputation and generative adversarial reconstruction networks. This approach will enable the model to dynamically synthesize missing sensory channels, such as the missing thermal stream, in asymmetric clinical environments, rather than relying on hard zero masking, thereby ensuring continuous and deployment-ready automated pain monitoring.

Acknowledgment

The authors thank the providers of the MIntPain and UNBC-McMaster datasets for making the data available for research purposes. The authors also acknowledge colleagues who provided technical and manuscript-related feedback during the study.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

The MIntPain dataset used in this study is not publicly available and can be accessed only through permission from the dataset owners and completion of the required End User License Agreement. The UNBC-McMaster Shoulder Pain Expression Archive is publicly available for research use through its official access channel. The processed data, trained models, and source code generated in this study are available from the corresponding author upon reasonable request, subject to dataset license restrictions.

Author Contribution

Raihan Islamadina conceptualized and designed the study, implemented the models, conducted data collection and experiments, performed data analysis, and prepared the original manuscript draft. Fitri Arnia, Taufik F Abidin, and Rusdha Muharar contributed to methodology validation, data interpretation, and critical revision of the manuscript. Aulia Syarif Aziz and Khairun Saddami contributed to manuscript review, technical editing, and research evaluation. All authors reviewed and approved the final manuscript and take full responsibility for the accuracy and integrity of the work.

Declarations

Ethical Approval

This study used secondary datasets only and did not involve direct recruitment of human participants or new data collection by the authors. The MIntPain dataset was used after contacting the dataset owners and completing the required End User License Agreement. Ethical approval and informed consent for the original MIntPain data collection were obtained by the original dataset creators. The UNBC-McMaster Shoulder Pain Expression Archive was used according to its permitted research-use conditions. Therefore, additional institutional ethical approval for this secondary analysis was not required.

Consent for Publication Participants

No new identifiable participant data were collected in this study. Consent for use and publication of data was obtained by the original dataset providers according to their ethical approval and dataset-use agreements.

Code Availability

The source code and trained model weights are available from the corresponding author upon reasonable request, subject to institutional and dataset-license restrictions.

Use of AI Tools

The authors declare that no generative AI tools were used in the design, analysis, or interpretation of the study. Language editing tools, if any, were used only to improve grammar and readability, and all content was reviewed and approved by the authors.

Competing Interests

The authors declare no competing interests.

References

[1] N. Ben Aoun, “A Review of Automatic Pain Assessment from Facial Information Using Machine Learning,” Jun. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/technologies12060092.

[2] G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker, and H. Wang, “Ensemble neural network approach detecting pain intensity from facial expressions,” *Artif. Intell. Med.*, vol. 109, p. 101954, Sep. 2020, doi: 10.1016/j.artmed.2020.101954.

[3] S. El Morabit, A. Rivenq, M.-E. Zighem, A. Hadid, A. Ouahabi, and A. Taleb-Ahmed, “Automatic Pain Estimation from Facial Expressions: A Comparative Analysis Using Off-the-Shelf CNN Architectures,” *Electronics (Basel).*, vol. 10, no. 16, p. 1926, Aug. 2021, doi: 10.3390/electronics10161926.

[4] T. Alghamdi and G. Alaghband, “Facial Expressions Based Automatic Pain Assessment System,” *Applied Sciences*, vol. 12, no. 13, p. 6423, Jun. 2022, doi: 10.3390/app12136423.

[5] T. Alghamdi and G. Alaghband, “SAFEPA: An Expandable Multi-Pose Facial Expressions Pain Assessment Method,” *Applied Sciences*, vol. 13, no. 12, p. 7206, Jun. 2023, doi: 10.3390/app13127206.

[6] A. R. Wahab Sait and A. K. Dutta, “Ensemble Learning-Based Pain Intensity Identification Model Using Facial Expressions,” *Journal of Disability Research*, vol. 3, no. 3, Mar. 2024, doi: 10.57197/JDR-2024-0029.

[7] M. S. Salekin, G. Zamzmi, D. Goldgof, R. Kasturi, T. Ho, and Y. Sun, “Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment,” *Comput. Biol. Med.*, vol. 129, p. 104150, Feb. 2021, doi: 10.1016/j.compbiomed.2020.104150.

[8] N. Ben Aoun, “Deep Learning-Based Pain Intensity Estimation from Facial Expressions,” 2024, pp. 484–493. doi: 10.1007/978-3-031-64836-6_47.

[9] S. A. S. Lakshminarayan, S. Hinduja, and S. Canavan, “Three-level Training of Multi-Head Architecture for Pain Detection,” in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, IEEE, Nov. 2020, pp. 839–843. doi: 10.1109/FG47880.2020.00071.

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

[10] A. Semwal and N. D. Londhe, "Computer aided pain detection and intensity estimation using compact CNN based fusion network," *Appl. Soft Comput.*, vol. 112, p. 107780, Nov. 2021, doi: 10.1016/j.asoc.2021.107780.

[11] H. Pedersen, "Learning Appearance Features for Pain Detection Using the UNBC-McMaster Shoulder Pain Expression Archive Database," 2015, pp. 128–136. doi: 10.1007/978-3-319-20904-3_12.

[12] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, doi: https://doi.org/10.48550/arXiv.1409.1556.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[14] M. A. Haque *et al.*, "Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, May 2018, pp. 250–257. doi: 10.1109/FG.2018.00044.

[15] Y. Lyu, I. Schiopu, and A. Munteanu, "Multi-modal neural networks with multi-scale RGB-T fusion for semantic segmentation," *Electron. Lett.*, vol. 56, no. 18, pp. 920–923, Sep. 2020, doi: 10.1049/el.2020.1635.

[16] K. Munadi *et al.*, "A Deep Learning Method for Early Detection of Diabetic Foot Using Decision Fusion and Thermal Images," *Applied Sciences*, vol. 12, no. 15, p. 7524, Jul. 2022, doi: 10.3390/app12157524.

[17] G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker, and H. Wang, "The modeling of human facial pain intensity based on Temporal Convolutional Networks trained with video frames in HSV color space," *Appl. Soft Comput.*, vol. 97, p. 106805, Dec. 2020, doi: 10.1016/j.asoc.2020.106805.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[19] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, Sep. 2020, [Online]. doi: 10.48550/arXiv.1905.11946

[20] R. Islamadina, K. Saddami, M. Oktiana, T. F. Abidin, R. Muharar, and F. Arnia, "Performance of Deep Learning Benchmark Models on Thermal Imagery of Pain through Facial Expressions," in *2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, IEEE, Nov. 2022, pp. 374–379. doi: 10.1109/COMNETSAT56033.2022.9994546.

[21] Raihan Islamadina *et al.*, "Learning Rate Analysis for Pain Recognition Through Viola-Jones and Deep Learning Methods," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 13, no. 2, pp. 77–83, May 2024, doi: 10.22146/jnteti.v13i2.9466.

[22] S. Gkikas and M. Tsiknakis, "Automatic assessment of pain based on deep learning methods: A systematic review," Apr. 01, 2023, *Elsevier Ireland Ltd*. doi: 10.1016/j.cmpb.2023.107365.

[23] K. Gasmi *et al.*, "Enhanced brain tumor diagnosis using combined deep learning models and weight selection technique," *Front. Neuroinform.*, vol. 18, Nov. 2024, doi: 10.3389/fninf.2024.1444650.

[24] M. Hussain, M. O'Nils, J. Lundgren, and S. J. Mousavirad, "A Comprehensive Review on Deep Learning-Based Data Fusion," *IEEE Access*, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3508271.

[25] S. Alphonse, S. Abinaya, and N. Kumar, "Pain assessment from facial expression images utilizing Statistical Frei-Chen Mask (SFCM)-based features and DenseNet," *Journal of Cloud Computing*, vol. 13, no. 1, p. 142, Sep. 2024, doi: 10.1186/s13677-024-00706-9.

[26] Y. Cheng and D. Kong, "CSINet: Channel–Spatial Fusion Networks for Asymmetric Facial Expression Recognition," *Symmetry (Basel).*, vol. 16, no. 4, p. 471, Apr. 2024, doi: 10.3390/sym16040471.

[27] P. Rodriguez *et al.*, "Deep Pain: Exploiting Long Short-Term Memory Networks for Facial Expression Classification," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 3314–3324, May 2022, doi: 10.1109/TCYB.2017.2662199.

[28] Y. Chen *et al.*, "DEYOLO: Dual-Feature-Enhancement YOLO for Cross-Modality Object Detection," 2024, pp. 236–252. doi: 10.1007/978-3-031-78447-7_16.

[29] M. Ferguson, R. Ak, Y.-T. T. Lee, and K. H. Law, "Automatic localization of casting defects with convolutional neural networks," in *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2017, pp. 1726–1735. doi: 10.1109/BigData.2017.8258115.

[30] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, "Explicit Attention-Enhanced Fusion for RGB-Thermal Perception Tasks," *IEEE*

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

Robot. Autom. Lett., vol. 8, no. 7, pp. 4060–4067, Jul. 2023, doi: 10.1109/LRA.2023.3272269.

[31] R. Rossi, M. A. Lazarini, and K. Hirama, "Systematic Literature Review on the Accuracy of Face Recognition Algorithms," *EAI Endorsed Transactions on Internet of Things*, vol. 8, no. 30, 2022, doi: 10.4108/eetiot.v8i30.2346.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[33] J. Brownlee, "Understand the Impact of Learning Rate on Neural Network Performance," Machine Learning Mastery. Accessed: Jan. 24, 2021. [Online]. Available: https://machinelearningmastery.com/understand-the-dynamics-of-learning-rate-on-deep-learning-neural-networks/

[34] G. Menchetti, Z. Chen, Di. J. Wilkie, R. Ansari, Y. Yardimci, and A. E. Cetin, "Pain detection from facial videos using two-stage deep learning," in *GlobalSIP 2019 - 7th IEEE Global Conference on Signal and Information Processing, Proceedings*, 2019. doi: 10.1109/GlobalSIP45357.2019.8969274.

[35] C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, vol. 26, no. 4, 1934, doi: 10.2307/2331986.

[36] A. Kimber, B. Efron, and R. J. Tibshirani, "An Introduction to the Bootstrap.," *The Statistician*, vol. 43, no. 4, 1994, doi: 10.2307/2348146.

[37] N. Zaeri and R. Qasim, "Thermal image identification against pose and expression variations using deep learning," *Journal of Engineering Research*, vol. 12, no. 4, pp. 751–760, Dec. 2024, doi: 10.1016/j.jer.2023.10.043.

[38] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The UNBC-McMaster shoulder pain expression archive database," in *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 2011. doi: 10.1109/FG.2011.5771462.

[39] Z. Li, W. L. Zheng, and B. L. Lu, "Multimodal Emotion Recognition with Missing Modality via a Unified Multi-task Pre-training Framework," in *MM 2025 - Proceedings of the 33rd ACM International Conference on Multimedia, Co-Located with MM 2025*, 2025. doi: 10.1145/3746027.3755459.

[40] X. Du, M. Safarzadeh, M. Zhu, S. Prasad, S. Das, and J. Chung, "An Explainable Transformer Model for Pain Intensity Assessment Using Multi-Modal Facial Sequential Images," 2025. doi: 10.2139/ssrn.5201152.

[41] O. El Othmani and S. Naouali, "Cross-spectral fusion of thermal and RGB imaging for objective pain estimation," *PLOS Digital Health*, vol. 5, no. 5, p. e0001424, May 2026, doi: 10.1371/journal.pdig.0001424.

Author Biography

Raihan Islamadina is a lecturer at the Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Ar-Raniry. He received his Bachelor's degree in Electrical Engineering with a specialization in Telecommunications from Universitas Syiah Kuala in 2012. He later obtained his Master's degree in Electrical Engineering, specializing in Informatics, from the same university in 2015. Currently, he is pursuing a doctoral degree in the Doctoral Program of Engineering Science at Universitas Syiah Kuala as part of the 2019 cohort. His academic background is in Electrical Engineering and Informatics, with research interests in image and multimedia information processing.

Fitri Arnia received a B. Eng degree from Universitas Sumatera Utara (USU) in Medan. She completed her master's and doctoral degrees at the University of New South Wales (UNSW) in Sydney, Australia, and at Tokyo Metropolitan University in Japan. She is a professor of multimedia signal processing in the Department of Electrical Engineering at the Faculty of Engineering, Syiah Kuala University. She was a visiting researcher at Tokyo Metropolitan University in Japan and at Suleyman Demirel University in Isparta, Turkey. She has been awarded the AusAID Scholarship, the Hashiya Scholarship, and the Latvian Government's Latvian Fellowship for Research. Since 2008, she has served as the Editor-in-Chief for Jurnal Rekayasa Elektrika (Accredited by RISTEKDIKTI). Her research interests include signal, image, and multimedia information processing. She is a member of PII, IEEE, ACM, and APSIPA.

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766

Taufik Fuadi Abidin is a Professor in Computer Science at Universitas Syiah Kuala. He completed my Ph.D. in Computer Science from North Dakota State University, USA, in 2006 under the supervision of Prof. Dr. William Perrizo and finished a Master's degree in Computer Science from the Royal Melbourne Institute of Technology, Australia , in 2000. He had been a Senior Software Engineer at Ask.com, one of the U.S. core search engine companies, developing algorithms and implementing efficient production-level programs to improve web search results. He was involved in developing utilities for usage mining and query understanding using Natural Language Processing (NLP). My research interests are in Machine Learning, Big Data Analytics, Web and Text Mining, Information Retrieval, and Information Extraction, including acronyms extraction, web classification, social media mining, and processing and analyzing large sets of digital data. He has developed algorithms to automatically extract tropical disease information reported on Indonesian online news articles and developed a repository of tropical disease incidents. He has also successfully developed algorithms to automatically determine the pairs of acronyms and expansions in the Indonesian language and developed what is called IndoAcro, the Indonesian Acronyms and Expansions Repository. The acronyms and expansions are determined automatically using machine learning algorithms, and to expedite the process, Hadoop big data technology was utilized.

Rusdha Muharar obtained the Sarjana Teknik (Bachelor of Engineering) degree in electrical engineering from Gadjah Mada University, Indonesia, in 1999. He received the M.Sc. degree from Delft University of Technology (TU Delft), The Netherlands, in 2004 and the Ph.D. degree from the University of Melbourne, Australia, in 2012, both in electrical engineering. From November 2012 to November 2013, he was a post-doctoral research fellow in the Department of Electrical and Computer Systems Engineering at Monash University, Australia. In April 2006, he joined the Department of Electrical Engineering at Syiah Kuala University, Indonesia, where he is currently a senior lecturer. His research interests are in communications theory, signal processing for wireless communications, and machine learning.

Aulia Syarif Aziz is a lecturer at the Faculty of Tarbiyah and Teacher Training, Universitas Islam Negeri Ar-Raniry. He received the S.Kom. degree in Informatics from Universitas Syiah Kuala, Indonesia, and the M.Sc. degree in Network Learning Technology from National Central University. His research interests include artificial intelligence, educational technology, and computer networks. In addition to his academic activities, he also serves as a journal manager and contributes to various scientific and institutional development programs.

Khairun Saddami obtained a Bachelor's degree in 2015 from Syiah Kuala University, Indonesia. He received his PhD in Electrical and Computer Engineering from Syiah Kuala University, Indonesia, in 2020, during which he was awarded a scholarship from the Ministry of Research, Technology, and Higher Education, the Republic of Indonesia, under the Scheme of Pendidikan Magister Menuju Doktor Untuk Sarjana Unggul (PMDSU). He has been with the Multimedia and Signal Processing Research Group (MUSIG), Syiah Kuala University since 2020. He also acts as a reviewer in reputable journals such as IEEE Access and ACM Computer Survey. He is a member of IEEE and the International Association for Pattern Recognition (IAPR). His research interests are in document image analysis, deep learning, biometric and biomedical image processing and pattern recognition.

Manuscript received 16 May 2026; Revised 15 June 2026; Accepted 26 June 2026; Available online 16 July 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i3.1766