FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

Handling Class Imbalance in Video-Based Pain Intensity Estimation Using ResNetBiLSTM: A Study on Loss Functions and Optimizers

Lu Zhicui[1], Farizuwana Akma Binti Zulkifle[2], Ahmad Zia Ul-saufie Bin Mohamad Japeri[3], Mohd Razif Bin Shamsuddin[4], and Aisyah Binti Mat Jasin[5*]

[1.5]Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM), Malaysia; Email: [1]2022571867@student.uitm.edu.my, [5]aisyahmj@uitm.edu.my
[2]Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM), Cawangan Negeri Sembilan, Kampus Kuala Pilah, Malaysia; Email: farizuwana@uitm.edu.my
[3]Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia; Email: m.razif@uitm.edu.my
[4]Faculty of Computer and Mathematical Sciences, University Teknologi MARA (UiTM), Shah Alam, Selangor, Malaysia; Email: ahmadzia101@uitm.edu.my

***Correspondence:** *Aisyah Binti Mat Jasin; Email: aisyahmj@uitm.edu.my; Phone: +60 1112280557*

ABSTRACT- Accurate pain intensity recognition is vital for improving clinical care, especially in scenarios where patients cannot self-report. However, although existing datasets are often imbalanced, the main challenge is that current optimization and loss functions lack sensitivity to minority classes, adapt poorly to intra-class variability, and are less robust under imbalance, leading to biased recognition performance. To address these challenges, this study proposes a hybrid deep learning model based on ResNet-50 and BiLSTM to capture both spatial and temporal features from facial expression videos while incorporating strategies to mitigate the imbalance issue. To address the challenge of category imbalance inherent in real-world datasets, the study constructs an imbalanced version of the BioVid Part A database, reflecting realistic pain distribution with limited high-intensity samples. The study systematically compares three loss functions—Cross Entropy Loss, Focal Loss, and Cross Entropy Focal Loss (CEFL)—and evaluates their effectiveness in enhancing minority class recognition. A comprehensive benchmarking of four state-of-the-art optimization algorithms, AdamW, LAMB, AdaBelief, and NovoGrad, is conducted across a range of learning rates to systematically evaluate their convergence dynamics and generalization performance within the proposed framework. Experimental results show that the Focal Loss combined with LAMB or NovoGrad achieves superior performance, with the best accuracy reaching up to 84.89%, significantly outperforming traditional configurations. This research highlights the importance of tailored training strategies for imbalanced facial pain recognition and provides a robust baseline for future work. Future directions include expanding to real-world, in-the-wild pain assessment scenarios and integrating multimodal signals to enhance robustness and accuracy.

Keywords: Video sequence, Pain assessment, Balancing loss, Optimizer.

ARTICLE INFORMATION

Author(s): Lu Zhicui, Farizuwana Akma Binti Zulkifle, Ahmad Zia Ul-saufie Bin Mohamad Japeri, Mohd Razif Bin Shamsuddin, Aisyah Binti Mat Jasin;

Received: 28/09/2025; **Accepted:** 07/12/2025; **Published:** 20/12/2025;
E- ISSN: 2347-470X;
Paper Id: IJEER 0210A01;
Citation: 10.37391/ijeer.130422
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130322.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Pain is a complex and prevalent clinical issue, if the issue is inadequately managed, it can lead to serious physical and psychological harm, underscoring the need for accurate and timely assessment [1]. Automated pain recognition has emerged as a promising solution, aiming to objectively evaluate pain through behavioral and physiological cues. Given that pain is a critical diagnostic indicator and a well-documented barrier to recovery, particularly in postoperative and Intensive Care Unit (ICU) [2] settings, this area has garnered increasing attention in healthcare, driving the development of various deep learning-based pain assessment methodologies.

Medasense has developed a medical device for objective pain monitoring, based on the premise that pain induces changes in physiological signals such as blood pressure, heart rate, respiratory rate, and SpO2 acquired through EMG, ECG, or EEG, which may occur individually or in combination, and often in an upward trend. However, acquiring such physiologic data is

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

typically more complex and time-consuming compared to facial video [3]. In contrast, facial expressions, as natural responses to pain, are easier to capture and provide rich information about pain intensity, making them a practical and effective modality for pain assessment [4]. The Facial Action Coding System (FACS) introduced in [5] provides a standardized framework for analysing facial expression by encoding facial muscle movements into discrete components known as Action Units (AUs). Building on this, the study in [6] proposed the Prkachin and Solomon Pain Intensity (PSPI) metric to quantify pain intensity based on selected AUs. This metric has established the foundation for many studies that utilise video-based facial features to automatically recognize and estimate pain.

In recent years, the integration of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) has demonstrated significant improvements across various video-based recognition tasks, including medical and affective computing domains [4, 7-9]. This hybrid architecture effectively captures both spatial features and temporal dynamics, making it especially suitable for sequential data such as facial expression videos in pain assessment. However, in the context of medical applications, two major challenges persist: the limited availability of annotated clinical data and the inherent class imbalance in pain expression datasets, where samples of high-intensity pain are often scarce.

In current pain intensity estimation studies, available pain datasets generally suffer from category imbalance, which restricts the performance improvement of deep learning models in automated pain assessment. The uneven distribution of pain level labels affects both training and generalization, leading to models that achieve better recognition for the majority class (no pain) but poorer recognition for the minority class (high-intensity pain) [10-14]. These limitations hinder the robustness and applicability of deep learning models in real-world clinical scenarios. While model architectures have advanced significantly, existing loss functions such as cross-entropy and focal loss still face challenges in handling imbalanced datasets: cross-entropy is easily dominated by majority classes, whereas focal loss often struggles to distinguish minority samples with subtle expressions. Consequently, few studies have systematically investigated how loss functions and optimization strategies can effectively address these data-driven constraints. To bridge this gap, this study explores balanced loss functions and training optimizers within a Residual Network and Bidirectional Long Short-Term Memory (ResNet-BiLSTM) framework for recognizing imbalanced facial pain expressions.

In the study presented, a deep learning architecture fusing ResNet-50 and BiLSTM is constructed. The proposed model leverages the advantages of ResNet-50 to model spatial features and the ability of BiLSTM to interpret the temporal dynamics, hence enhancing the recognition of variations in continuous pain expressions. In terms of experimental design, this paper builds a five-level pain intensity recognition task based on the BioVid thermal pain database for training and evaluation, and pays special attention to the performance of the model under unbalanced data conditions. To address the problem of category imbalance, the proposed model compares the performance differences of three different loss functions: Cross-Entropy Loss (CE), Focal Loss (FL), [15] Cross-Entropy Focal Loss (CEFL) [16], and analyzes their roles in minority category recognition. In addition, this paper also combines four well-known optimizers - AdamW [17], LAMB [18], AdaBelief [19], and NovoGrad [20] - to compare the performance of the proposed model and finally determines the optimal combination of loss functions and optimizers[21] as a reference for future research directions.

The main contributions of this study are:

- A hybrid ResNet-50 and BiLSTM is developed to extract spatial and temporal features from facial expressions, improving the accuracy of multi-level pain intensity estimation.
- A systematic evaluation of three class-balancing loss functions is conducted under imbalanced data conditions, providing both theoretical and empirical evidence for enhancing the model's ability to recognise minority pain categories.
- A comparative analysis of three optimization algorithms is performed to assess training stability and generalization performance under imbalanced learning, leading to the identification of the optimal configuration as the benchmark scheme.

2. RELATED WORK

Several recent studies have explored automated pain intensity estimation using deep learning approaches across various benchmarked datasets.A. Neg at al.[22]proposed a hybrid feature fusion framework that combines local–global features with temporal context for robust facial expression recognition. Authors in [23] proposed a hybrid architecture combining VGGFace, PCA, and a CNN-BiLSTM structure (EJH-CNN-BiLSTM) to classify four pain intensity levels on the UNBC-McMaster Shoulder Pain Expression Archive Database. Work in [13] utilized a Deep Convolutional Neural Network (DCNN) model on the same dataset to perform binary pain or no-pain classification, achieving impressive results with an AUC of 97% and ROC values ranging between 0.95 and 0.97. Study in [24] applied a CNN-BiLSTM model for binary pain classification, reporting an F1-score of 64.35 ± 10.40. Authors in [11] employed models such as Random Forest classifier (RFc), Long Short-Term Memory (LSTM), and LSTM-sw on the X-ITE pain database to classify four pain levels. More recently, Researchers in [8] implemented a Fully Convolutional Network (FCN)-based BiLSTM model to extract pain-relevant spatiotemporal features for improved recognition performance. These studies collectively demonstrate the widespread applicability and effectiveness of CNN-RNN-based architectures for automatic pain recognition across various datasets and classification tasks.

Current pain intensity estimation research is hindered by the category imbalance in available datasets, leading to deep learning models that perform well on majority classes, such no pain, but struggle to accurately recognize minority classes, such as high-intensity pain [10-14]. Researchers have proposed various strategies to address the challenge of category imbalance, among which the balanced loss function has shown

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

promising results because it does not require modifying the original data and can be integrated directly into the model training. Authors in [16] propose Focal loss, which modifies the standard CE to reduce the contribution of easily classified samples, thereby directing the model's focus toward harder, minority-class samples. Work in [25] embedded focal loss and CosFace Loss into the training model to overcome the imbalance of data sets hindering model training. Focal loss is designed to highlight hard-to-classify samples, but if misclassified samples are abundant, their contribution to the cumulative loss is reduced. Authors in [16] proposed CEFL and CEFL2 loss, which are reweighted based on the CE function and the focus loss function to improve the classification accuracy of imbalanced datasets [26, 27].

3. METHODOLOGY

In this section, a pain intensity assessment framework is proposed, which contains two steps: pain feature extraction and temporal modeling with intensity prediction. First, unprocessed videos are transformed into normalized facial image sequences. Then, both spatial features and spatiotemporal representations are leveraged to enhance the evaluation of pain intensity from facial expression sequences. Finally, the output layer generates predicted intensity scores (five-class classification: no pain to severe pain). At this stage, the framework incorporates specifically designed loss functions, such as Cross Entropy Loss, Focal Loss, and Cross Entropy Focal Loss. These loss functions not only measure the discrepancy between the predicted intensity distribution and the ground-truth labels but are primarily introduced to address the class imbalance problem in pain intensity datasets, where severe pain samples are far fewer than pain-free or mild cases. By dynamically adjusting the weights of minority classes and hard-to-classify samples, they guide the model to pay greater attention to underrepresented pain levels. Subsequently, gradient backpropagation is performed with advanced optimizers (*e.g.*, AdamW, LAMB, AdaBelief, or NovoGrad), enabling the framework to learn more balanced and robust feature representations.

Thus, the position of the loss functions in the framework is at the prediction layer, where they align the model's outputs with the annotated pain intensity levels, completing the end-to-end optimization pipeline illustrated in *figure 1*.

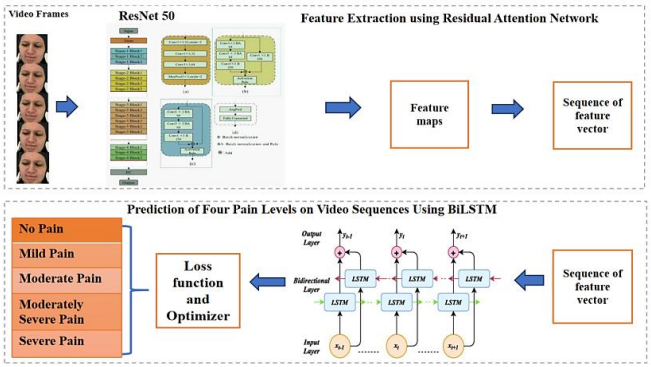

Figure 1. Pain Assessment Framework

3.1. Data Pre-Processing

In this study, the BioVid Heat Pain Database (BVDB) [26] is utilized, which contains recordings of spontaneous facial expressions and physiological signals from laboratory-induced heat pain. The BVDB dataset contains biomedical signals and anterior facial recordings from 87 subsets, which are labeled as pain-free and four levels of pain intensity, such as mild, moderate, moderately severe, and severe pain, as shown in *figure 2*. Each subset contains video sequences recorded under various pain intensity conditions, labeled with the corresponding pain intensity level. Each video has a duration of 5.5 seconds.

Figure 2. Five samples of pain expression in the BioVid dataset

This study establishes a data preprocessing pipeline on the BVDB, as shown in *figure 3*, to enhance the model's ability to recognize pain-related facial features. First, facial regions are detected in each video frame using OpenCV [28]. Subsequently, 68 facial landmarks are extracted using the dlib library [29] to perform face alignment. This step addresses variations in facial position and orientation across frames, thereby improving spatial consistency and enhancing the saliency of facial features relevant to pain expression. Finally, based on the detected landmarks, each image is cropped to a standardized size of 224 × 224 × 3, eliminating background noise and focusing on pain-relevant facial regions. All images are then normalized to ensure consistent pixel distribution. This preprocessing procedure provides the deep learning model with well-structured and semantically focused high-quality input data.

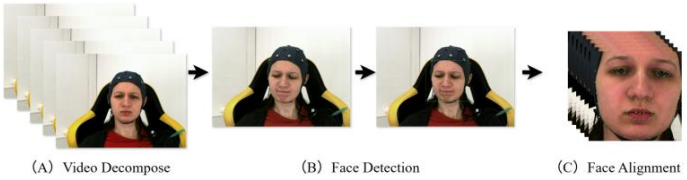

Figure 3. ResultsProcess of pain expression dataset pre-processing

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, and the experimental conclusions that can be drawn.

3.2. Model Construction

Figure 1 presents the overall architecture of our pain intensity estimation framework, which includes two key components. Initially, ResNet-50 is utilized to extract spatial features from each facial frame. Subsequently, a BiLSTM network describes the temporal dynamics across the frame sequence.

3.2.1. Extraction of spatial features

This study employed a deep residual network base on the ResNet-50 architecture as a spatial feature extraction module for

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

facial images. CNN has the ability to extract local to global features layer by layer, but the deep network is susceptible to the gradient disappearance, which makes it difficult to train. Research presented in [9] proposed ResNet, which can effectively mitigate the gradient problem by introducing "shortcut connections" to make the network easier to train and improve the feature extraction in the classification model. ResNet-50 performs well in the expression recognition task with strong generalization and robustness.

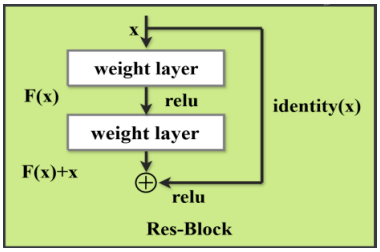

Figure 4. A building block of residual learning

In this study, facial image sequences are fed into a ResNet to extract robust and discriminative features for pain expression recognition. As illustrated in *figure 4*, each residual block is designed to learn the residual mapping rather than the direct transformation. Given an input feature *x*, through the convolution operation, the study obtained the residual function, which is defined as, wherer denotes the ReLU activation function. Then, *via* a shortcut, the input identity mapping is directly added to the output of the convolutional transform, yielding the final response, which serves as the input to the next layer. As shown in *figure 5*, the preprocessed video frames are successively passed through four stages of residual blocks in the backbone network, resulting in a high-level feature representation with 2028 channels in the final feature map. Each residual block achieves the summation of the input and convolution results through constant mapping, which enhances the information flow and stability under the depth of the model, and thus provides high-quality spatial feature support for subsequent time-series modeling.

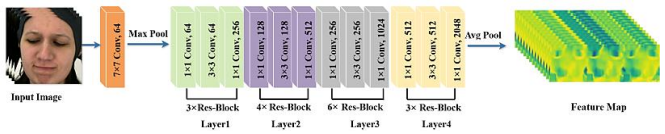

Figure 5. The overflow based on ResNet50 feature extraction

3.2.2. Extraction of Spatiotemporal Features

The LSTM network is a specialized form of RNN. It is designed to address the limitations of traditional RNNs in capturing long-term dependencies, particularly the issues of vanishing and exploding gradients during the training of long sequences. As depicted in *figure 6*, the LSTM architecture includes three gate mechanisms: forget, input, and output gates, which regulate the flow and preservation of important information for long sequences through the network.

This study leverages the ability of the BiLSTM [30] model to effectively capture the temporal dependencies between consecutive frames. Unlike the standard LSTM, which only considers the past context, the BiLSTM processes the input sequence in both forward and backward directions, thereby enabling the model to exploit information from both past and future frames simultaneously. Such bidirectional processing is particularly important for facial expression recognition, where subtle pain-related muscle movements may occur before or after a given frame. *Figure 5* illustrates the internal structure of a single LSTM cell, which is the fundamental unit of the BiLSTM. At each time step $u_t \in R^{1028}$, extracted from the convolutional layers, together with the previous hidden state h_{t-1} and the previous cell state c_{t-1}. The internal operations of the LSTM cell can be described as follows [30];

$$i_t = \sigma(\gamma_i \bullet u_t + W_i \bullet h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(\gamma_i \bullet u_t + W_f \bullet h_{t-1} + b_c) \quad (2)$$

$$c_t = tanh(\gamma_i \bullet u_t + W_c \bullet h_{t-1} + b_c) \quad (3)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes c_t \quad (4)$$

$$g_t = \sigma(\gamma_g \bullet u_t + W_g \bullet h_{t-1} + b_g) \quad (5)$$

$$h_t = g_t \otimes tanh(c_t) \quad (6)$$

Where, i_t, f_t and $o_t$denote the input gate, forget gate, and output gate, respectively, while $\tilde{c}_t$ represents the candidate cell state. The vector $c_t$is the updated cell state that integrates both long-term and short-term memory, and $h_t$is the updated hidden state passed to the next time step. The symbol $\sigma(\cdot)$denotes the sigmoid activation function, $\tanh(\cdot)$represents the hyperbolic tangent function, and $\otimes$ indicates element-wise multiplication. Through this gating mechanism, the LSTM unit is able to selectively retain, update, and output relevant temporal features from the 1028-dimensional convolutional feature sequence. When extended to the bidirectional structure, the BiLSTM significantly enhances the model's ability to capture the temporal dynamics of facial expressions related to different pain intensities.

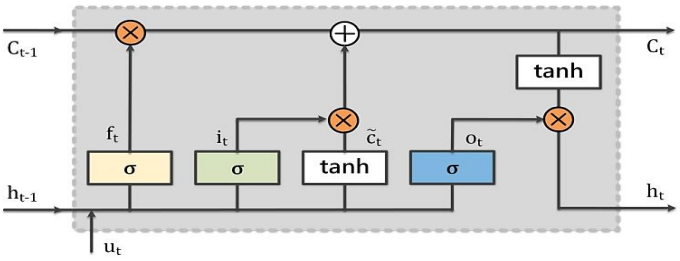

Figure 6. The structure of an LSTM unit [30]

4. EXPERIMENTAL WORK

4.1. Data Preparation

In this study, the BVDB dataset is used to validate the effectiveness of the proposed method in dealing with the category imbalance problem. To ensure fair evaluation and avoid data leakage, all experiments adopt a subject-independent partition, where subjects in the training and testing sets are mutually exclusive. Specifically, all videos from 80% of the subjects were assigned to the training set, and the remaining 20%

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

to the testing set. To simulate realistic pain distributions—where high-intensity pain occurrences are rare in medical and occupational settings—the raw dataset was reconstructed into an imbalanced version through random subject-level down-sampling. The down-sampling process ensures: *(1)* samples are selected uniformly within each class; *(2)* no subject appears in more than one subset; *(3)* reproducible selection via fixed random seeds. The degree of imbalance follows the commonly used imbalance ratio (IR) criterion widely adopted in class-imbalanced benchmarks such as CIFAR-10/100 [16].

The resulting distribution is summarized in *table 1*, where lower pain levels contain more samples while high-intensity pain levels remain scarce. This forms a more challenging benchmark for minority-class recognition. Based on the subject-independent split, 2,992 videos were used for training and 748 for testing, as shown in *table 2*.

Table 1. Pain intensity level distribution

Pain Level	Description	Number of Videos	Percentage
Level 1	No pain	1,340	37%
Level 2	Mild pain	1,050	28%
Level 3	Moderate pain	700	19%
Level 4	Moderate severe pain	400	11%
Level 5	Severe pain	250	6%
Total	—	3740	100%

Table 2. Training and testing picture data

Datasets	Videos	Proportion	Level	Size
Training	2992	80%	0, 1, 2, 3, 4	224×224×3pixel
Testing	748	20%	0, 1, 2, 3, 4	224×224×3pixel

4.2. Experimental Setup

All experiments were conducted on an HP workstation equipped with an NVIDIA Tesla A30 GPU (24GB VRAM) under Ubuntu 20.04, using CUDA 12.4 and PyTorch (≥2.1). The proposed model consists of a ResNet-50 backbone for spatial feature extraction and a two-layer bidirectional LSTM with 256 hidden units and a dropout rate of 0.3 for temporal modeling. Each input frame is resized to 224×224 and processed by ResNet-50 to produce a 2048-dimensional global pooled feature, which is then sequentially aggregated by the BiLSTM. A fully connected layer maps the resulting temporal representation to a five-class softmax output.

During training, Kaiming initialization is applied to non-pretrained layers, and gradient clipping with a max-norm of 5 is used to stabilize optimization. Mixed-precision (AMP) training is enabled to improve computational efficiency. The models are trained for 100 epochs with a mini-batch size of 8. We evaluate four optimizers—AdamW, LAMB, AdaBelief, and NovoGrad—and follow their standard hyperparameter settings (*e.g.*, β_1=0.9, β_2=0.999 for AdamW; ε=*1e-12* and weight decay=*1e-4* for AdaBelief). The base learning rate is selected from [0.01, 0.001, 0.0001, 0.00001] and is decayed using StepLR with a decay factor of 0.1 every 20 epochs. Cross-Entropy Loss, Focal Loss, and CEFL are employed as training objectives depending on the experimental configuration. All feature extraction, data loading, and visualization rely on Torchvision, Scikit-learn, Matplotlib, and Seaborn.

The experimental setup is designed to comprehensively evaluate the model's performance in the facial pain video classification task. A combination of training strategies is employed to conduct comparative experiments. In this experimental work, the three common loss functions, namely CE Loss, FL, and CEFL function will be utilized. Then, for the optimizer, such as AdamW, LAMB, AdaBelief, and NovoGrad are used to explore the effects of different gradient updating methods on the convergence of the model in the learning rate setting. This experiment covers multiple initial learning rates, such as 0.01, 0.001, 0.0001, and 0.00001. The combination of the proposed loss functions and optimizers with the dynamic adjustment of the learning rate scheduler will improve the training efficiency and generalization ability of the model.

To comprehensively assess the performance of the facial pain classification model, a set of well-established evaluation metrics is employed. These metrics include Accuracy, Precision, Recall, and F1-score. In addition, to address the challenges posed by class imbalance—common in pain expression datasets—three specialized metrics are introduced: Probability of Detection (PD), Probability of False Alarms (PF), and Balance (Bal). Bal considers the balance between PD and PF comprehensively, evaluating the model's overall performance in pain recognition versus false alarm avoidance. Higher Bal values indicate better balance between recognition accuracy and false alarm control, with ideal values approaching 1. In multi-class classification, Bal is typically calculated for each category and then weighted or averaged. The definitions of these metrics are as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{7}$$

$$\text{Pr e cision} = \frac{\text{TP}}{\text{TP+FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{9}$$

$$\text{F1} - \text{Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Pr ecision+Re call}} \tag{10}$$

$$\text{PD} = \frac{\text{TP}}{\text{TP+FN}} \tag{11}$$

$$\text{PF} = \frac{\text{FP}}{\text{FP+TN}} \tag{12}$$

$$\text{Bal} = 1 - \frac{\sqrt{(0-\text{PF})^2+(1-\text{PD})^2}}{\sqrt{2}} \tag{13}$$

4.3. Balanced Loss Function

Loss function is used to evaluate the error between the predicted output and the true label. The balanced loss function employed in this study is to improve the class imbalance in facial pain videos for five levels pain intensity assessment. Besides, CE also adopted as the base loss function due to its stable performance in multi-categorization tasks[16]. Formula for the CE loss is as follows:

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

$$L_{CE} = -\log(p_t) \quad (14)$$

However, facing the problem of severe uneven distribution of classes, the standard CE Loss tends to cause the model to be biased towards the majority class[31], which affects the minority class recognition performance. Lin et al.[15]proposed FL function to address the problem of class imbalance. This loss function aims to reduce the contribution of easy-to-classify. samples by down-weighting gradient ratio between classes, Thereby, this effort will be encouraging the model to focus more on hard or minority samples which contribute to less learning and lead to be biased toward the majority samples. Therefore, the FL functions will focus on hard or misclassified samples. The formula is as follows[15];

$$FL = -(1 - p_t)^{\gamma} \log(p_t) \quad (15)$$

However, it has been found that when the predicted probability of the true class (p) is relatively low, such as p≤0.5, Focal Loss tends to reduce the loss value compared to Cross-Entropy Loss[15]. This may hinder the amplification of the loss gap between "easy" and "hard" samples. To address this issue, the authors proposed the CEFL, formulated as follows [16];

$$CEFL = -\frac{1}{N}\sum_{i=1}^{N}[(1-p)\log(p) + p(1-p)^{\gamma}\log(p)] \quad (16)$$

Following Lin et al.[15] , the focusing parameter is set to γ = 2.0, which yielded the most stable convergence in preliminary validation. The class-balancing factor α is computed adaptively according to class frequency, where the weight for class *i* is defined as:

$$a_i = \frac{1/c_i}{\sum_{j=1}^{K} 1/c_j} \quad (17)$$

Where $K = 5$ is the total number of categories. This strategy essentially normalizes based on the inverse of class frequency, giving higher weights to categories with fewer samples, thereby increasing the model's attention to them.

5. DISCUSSION

5.1. Analysis of Test Results

The ResNet-BiLSTM model developed in the study is to train the model on the BVDB unbalanced dataset with 100 rounds of training using CE Loss, FL and CEFL, respectively, and the best accuracies were recorded on the Video1 test set. The experimental results in *table 3* show that the best accuracy of FL is 83.3%, which is the best performance among all the loss functions; the best accuracy of CEFL is 80.0%; and the best accuracy of traditional CE Loss is 78.1%.

Table 3. Comparison of best accuracy for ResNet-BiLSTM model

Function	Accuracy	Precision	Recall	F-score	Bal
CE Loss	78.1%	78.53%	77.32%	78.11%	79.52%
Focal Loss	**83.3%**	**83.76%**	**82.47%**	**83.40%**	**85.88%**
CEFL	80.0%	80.44%	79.20%	80.09%	81.48%

The analysis of the results indicates that FL exhibits greater adaptability when combined with the proposed model and dataset, particularly in addressing class imbalance. This adaptability is reflected in a 5.2% (83.3% vs. 78.1% in *table 4*) performance improvement over the traditional CE Loss, and also in a significantly higher balance performance, with its BAL score reaching 85.88%, compared to CE Loss's 79.52%. The key advantage of FL lies in its modulation factor, which dynamically down-weights the loss contribution from easy-to-classify samples, allowing the model to focus more on hard samples, especially those from minority classes. This leads to enhanced learning of underrepresented categories and improved generalization.

In contrast, although the CE Loss is a classical technique and widely adopted loss function for multi-class classification, this loss function applies uniform weighting to all samples without consideration of their difficulty or class imbalance. This often results in a bias toward majority classes in imbalanced datasets, limiting their effectiveness in capturing minority class characteristics.

Although CEFL attempts to integrate the strengths of both CE and Focal Loss. The performance is not as good as the focal loss; the result obtained for this loss function only reaches 80.0%, and its Bal score is 81.4%, which is notably lower than that of Focal Loss. This could be because its loss structure and weight adjustment mechanism fail to fully coordinate the advantages of the two loss functions, making it difficult for the model to effectively focus on key samples throughout the training process and lowering the overall performance.

Since FL demonstrates superior performance in the comparative results, this loss function was combined with different optimizers (AdamW, LAMB, AdaBelief, and NovoGrad) and learning rates (0.01, 0.001, 0.0001, 0.00001) on the ResNet-BiLSTM model. Experimental results reveal significant disparities in how different optimizers adapt to learning rates, as visualized in the histogram in *figure 7* and detailed in *table 4*.

When the learning rate decreases to 0.0001, almost all optimizers achieve their peak accuracy except AdamW. LAMB achieves the highest accuracy of 0.8489, closely followed by NovoGrad at 0.8459, AdaBelief at 0.8380, and AdamW at 0.8331. At the lowest learning rate of 0.00001, AdamW exhibits remarkable stability, reaching the highest accuracy of 0.8429—an improvement from its 0.8331 at 0.0001, indicating consistent performance enhancement as the learning rate decreases. In contrast, LAMB's performance degrades significantly from 0.8489 to 0.7038, while NovoGrad drops substantially to 0.6431. AdaBelief maintains relatively stable performance, shifting from 0.838 to 0.833, but still falls short of AdamW's optimal performance at this learning rate. In summary, AdamW achieves relatively higher accuracy peaks and optimal stability under changes in learning rates.

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

Figure 7. Evaluation Metrics Across Optimizers and Learning Rates

TABLE 4. Effect of optimizer and learning rate on accuracy with focal loss

Optimizer	0.01	0.001	0.0001	0.00001
AdamW	0.2068	0.2068	0.8331	**0.8429**
LAMB	0.2068	0.4046	**0.8489**	0.7038
AdaBelief	0.2167	0.5974	0.838	0.833
NovoGrad	0.1988	0.2086	0.8459	0.6431

Figure 8 illustrates the training and test accuracy curves of the ResNet-BiLSTM model trained with Focal Loss and the LAMB optimizer on the unbalanced BVDB dataset. The training accuracy steadily increases and eventually saturates, while the test accuracy improves rapidly during early epochs and stabilizes at a high level. The numerical results in *table 5* show that the LAMB optimizer achieves the highest accuracy (84.89%) with a learning rate of 0.0001, highlighting its effectiveness in optimizing model performance under class-imbalanced conditions.

Table 5 summarizes the classification performance of various models and loss function strategies on the facial pain intensity assessment task. Among them, the proposed ResNet-BiLSTM model, integrated with Focal Loss and the adapter optimizer, achieved an accuracy exceeding 82%, substantially outperforming other baseline methods. In contrast, the CNN-LSTM [4] model trained with Regularization Center Loss reached only 37.42%, likely due to its limited ability to jointly capture deep spatial representations and long-range temporal dependencies. The standard ResNet [25] model, equipped with FL and CosFace Loss, improved the accuracy to 61.88%, leveraging a more discriminative feature embedding and enhanced focus on difficult samples. Meanwhile, the Combating Uncertainty and Class Imbalance Network (CUCN) [32] model using a modified CE Loss attained 63.29%, indicating moderate improvement but still falling short in handling imbalanced category distributions effectively.

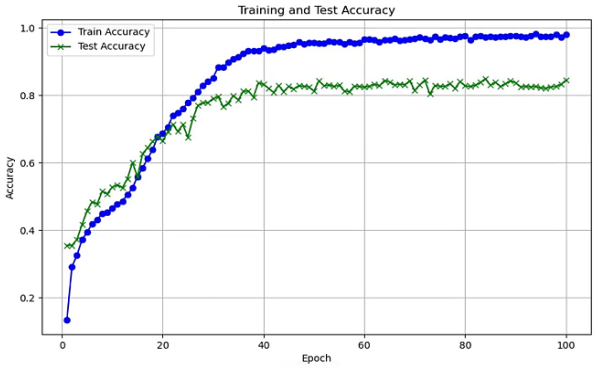

Figure 8. Accuracy variation curve during training of Focal Loss and LAMB optimizer

TABLE 5. Effect of balanced loss functions on model performance under imbalanced data conditions

Classifier	Method	Accuracy
CNN - LSTM	Regularization center loss	37.42%
ResNet	Focal Loss and CosFace Loss	61.88%
CUCN	Modifications Cross Entropy	63.29%
ResNet - BiLSTM	ResNet - BiLSTM + Focal Loss + adapter optimizer	**>82%**

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829 | e-ISSN: 2347-470X

Considering the dataset's class imbalance (Level 1: 37%, Level 5: 6%), we adopted subject-wise 5-fold cross-validation to ensure fair evaluation and prevent subject overlap between the training and testing sets. The model was trained using Focal Loss with the LAMB optimizer (learning rate = 0.0001). Performance was assessed using Accuracy, Precision, Recall, F1-score, and Balance (Bal), reported as the mean ± standard deviation across folds, as shown in *figure 9*. Final results were Accuracy = 85.30% ± 1.45%, Precision = 84.70% ± 1.30%, Recall = 84.90% ± 1.25%, F1-score = 85.36% ± 1.20%, and Balance = 87.40% ± 1.10%, confirming the ResNet - BiLSTM model's generalization under imbalanced conditions.

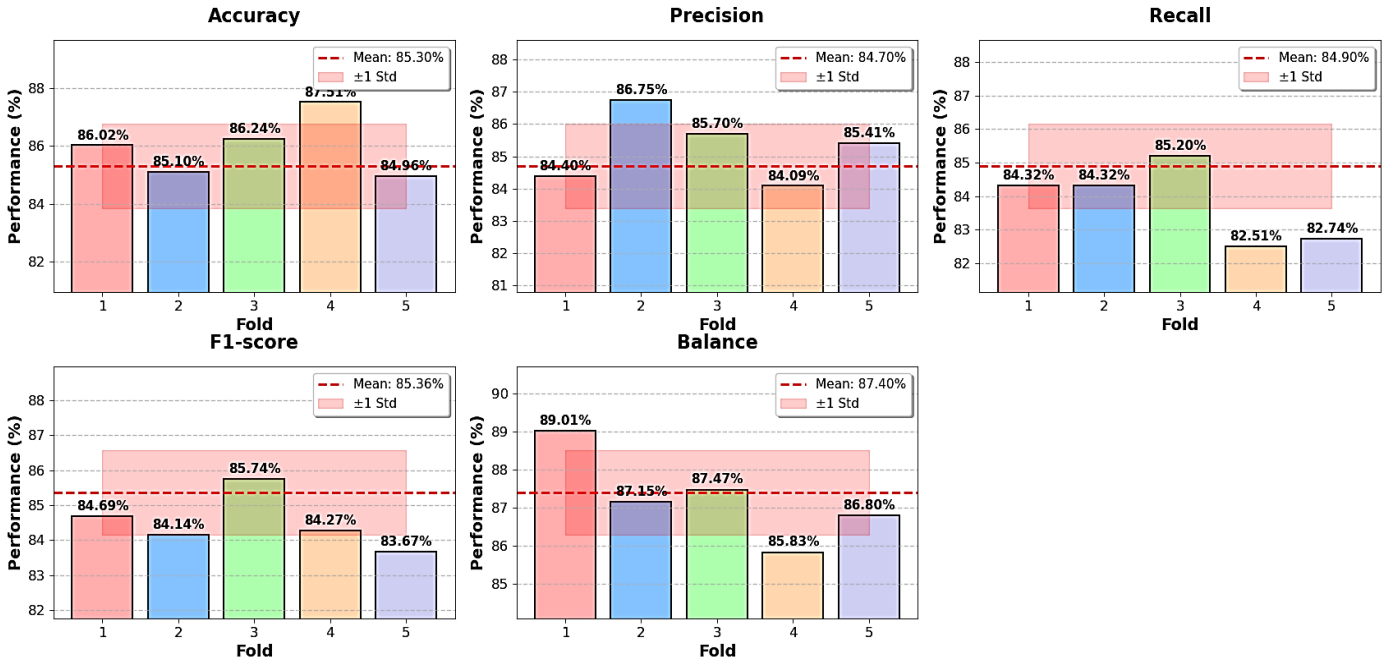

Figure 9. 5-Fold Cross-Validation Performance Comparison

To validate the effectiveness of each component in the proposed model, we conducted a systematic ablation study. Specifically, we examined the performance of three architectural variants: ResNet backbone only, ResNet+LSTM, and ResNet+BiLSTM, in order to quantify the contribution of different temporal modeling strategies to the overall results. The experimental outcomes are summarized in *table 6*. To further illustrate the classification details and inter-class confusion patterns of each model intuitively, *figure 10* presents the confusion matrices of the three variants on the test set. For the ResNet-only model, the confusion matrix shows noticeable misclassification between adjacent pain levels.

This suggests that the single ResNet backbone, which lacks temporal feature modeling, struggles to capture the dynamic characteristics of pain expressions, resulting in limited discrimination between similar pain intensities. The ResNet+BiLSTM model achieves the most concentrated confusion matrix, with the highest diagonal values across all categories—particularly for high pain levels (Level 4 and 5), where 67 out of 80 Level 4 samples and 48 out of 50 Level 5 samples are correctly identified. Misclassifications are minimized to only a small number of adjacent-level cases (*e.g.*, 25 Level 1 samples misclassified as Level 2), benefiting from the bidirectional temporal modeling of BiLSTM that captures both past and future contextual information of video frames.

TABLE 6. Comparison of Model Performance Under Different Sequential Modeling Strategies

Model	Accuracy	Precision	Recall	F-score
ResNet	62.05%	61.55%	63.20%	62.10%
ResNet-LSTM	81.53%	81.30%	81.05%	81.78%
ResNet - BiLSTM	**84.89%**	**84.76%**	**83.47%**	**84.40%**

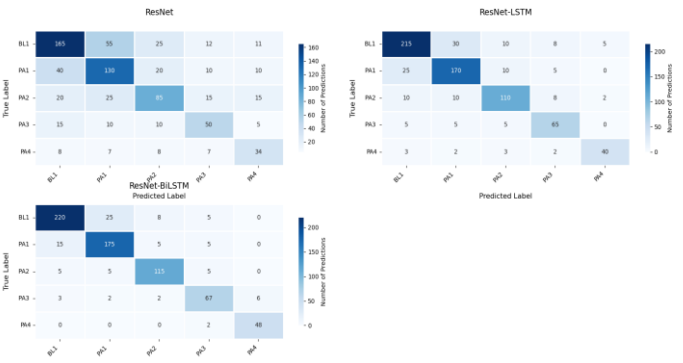

Figure 10. Confusion Matrices for Pain Level Classification Models

FOREX Publication
Open Access | Rapid and quality publishing
International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 820-829|e-ISSN: 2347-470X

5.2. Discussion

This study presents a ResNet-BiLSTM-based deep learning architecture tailored for sequential pain expression recognition, a task characterized by strong temporal dependencies and significant class imbalance. The model's performance is systematically examined by evaluating the impact of loss function selection, optimizer configuration, and learning rate strategies, thereby providing deeper insights into designing effective training for imbalanced learning conditions.

First, the architectural design of the proposed model is inherently tailored to the task characteristics. Unlike static image recognition, pain expression in video sequences is a dynamic phenomenon, requiring the effective capture of temporal contextual information. The ablation study confirms the incremental value of temporal modeling components: the ResNet-only baseline, lacking temporal feature extraction, achieves limited classification performance due to its inability to model frame-wise dependencies; the integration of LSTM mitigates inter-class confusion by leveraging unidirectional temporal sequences; and the adoption of BiLSTM further enhances robustness by capturing bidirectional contextual information, which proves critical for distinguishing subtle differences between adjacent pain levels (*e.g.*, mild vs. moderate pain) and improving the recognition accuracy of high-intensity pain (Level 4–5). This progressive improvement validates that bidirectional temporal modeling is essential for recognizing sequential pain.

Moreover, this study examines the impact of various optimizers and learning rate combinations on the effectiveness of model training. Experiments were conducted using a combination of four optimizers: AdamW, LAMB, AdaBelief, and NovoGrad, with four learning rates: 0.01, 0.001, 0.0001, and 0.00001. The results indicate significant differences in learning rates, sensitivity, and performance among the optimizers.

Comparison with existing methods shows that the traditional CNN-LSTM + regularized central loss method only achieves 37.42% accuracy, while ResNet + Focal Loss + CosFace improves the accuracy to 61.88%, and CUCN + modified cross-entropy loss reaches 63.29%. In contrast, the proposed ResNet-BiLSTM architecture, integrated with Focal Loss and an adapter optimizer strategy, achieves an accuracy exceeding 82%. This represents a substantial improvement in addressing the challenges of recognizing imbalanced and temporally dynamic pain.

6. CONCLUSION

This study aimed to develop a ResNet-BiLSTM model for facial pain expression recognition, addressing the dual challenges of temporal dependencies and category imbalance inherent in the BioVid Part A dataset. To this end, we systematically evaluated the impact of different loss functions, optimizers, and learning rate strategies on model performance.

Experimental results demonstrate that the proposed model, when combined with Focal Loss (FL), achieves the best accuracy of 83.3%, outperforming CEFL (80.0%) and conventional CE Loss (78.1%). The superiority of FL is attributed to its adaptive modulation factor, which enhances the learning of underrepresented pain categories by dynamically down-weighting easy-to-classify samples. Furthermore, optimizer and learning rate experiments reveal that the LAMB optimizer achieves the highest accuracy of 84.89% at a learning rate of 0.0001, while AdamW exhibits the best stability with an accuracy of 84.29% at a learning rate of 0.00001. Overall, the ResNet-BiLSTM integrated with FL and adaptive optimization strategies achieves an accuracy of over 82%, substantially outperforming baseline methods such as CNN-LSTM (37.42%), ResNet with FL and CosFace Loss (61.88%), and CUCN (63.29%).

These findings confirm the effectiveness of the proposed architecture in handling class imbalance and temporal dynamics in pain recognition tasks. Future research can further extend to the pain recognition task in complex environments in the wild, incorporating more robust feature extraction techniques to enhance the generalization ability of the model under uncontrolled conditions; meanwhile, multimodal information (*e.g.*, speech, physiological signals, or gestures) can be introduced to enhance the model's semantic comprehension and judgment ability, laying the foundation for constructing a more realistic and comprehensive automated pain assessment system.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgments: The authors would like to express their sincere gratitude to the creators of the BioVid Heat Pain Database for providing access to the dataset, which has been essential for the completion of this research.

Ethical Approval: This study was approved by the Research Ethics Committee of Universiti Teknologi MARA (UiTM), Malaysia (Reference: 600-TNCPI (5/1/6); REC/10/2024 (PG/MR/526)).

Author Contributions: Conceptualization, Zhicui Lu and Aisyah Binti Mat Jasin; methodology, Zhicui Lu and Aisyah Binti Mat Jasin; Models, Zhicui Lu; validation, Zhicui Lu, Aisyah Binti Mat Jasin, and Ahmad Zia Ul-saufie Bin Mohamad Japeri; formal analysis, Zhicui Lu; investigation, Zhicui Lu; data Apply, Zhicui Lu、Aisyah Binti Mat Jasin; writing—original draft preparation, Zhicui Lu; writing—review and editing, Zhicui Lu, Aisyah Binti Mat Jasin, and Farizuwana Akma Binti Zulkifle; supervision, Aisyah Binti Mat Jasin and Mohd Razif Bin Shamsuddin; All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] J. O. Allen, B. Zebrack, D. Wittman, K. Hammelef, and A. M. Morris, "Expanding the NCCN guidelines for distress management: a model of barriers to the use of coping resources," *The Journal of community and supportive oncology,* vol. 12, no. 8, pp. 271-277, 2014.

[2] S. Taggart, K. Skylas, A. Brannelly, G. Fairbrother, M. Knapp, and J. Gullick, "Using a clinical judgement model to understand the impact of validated pain assessment tools for burn clinicians and adult patients in the ICU: A multi-methods study," *Burns,* vol. 47, no. 1, pp. 110-126, Feb 2021, doi: 10.1016/j.burns.2020.05.032.

[3] M. Kunz, D. Meixner, and S. Lautenbacher, "Facial muscle movements encoding pain-a systematic review," *Pain,* vol. 160, no. 3, pp. 535-549, Mar 2019, doi: 10.1097/j.pain.0000000000001424.

[4] X. Xiang, F. Wang, Y. Tan, and A. L. Yuille, "Imbalanced regression for intensity series of pain expression from videos by regularizing spatio-

temporal face nets," *Pattern Recognition Letters,* vol. 163, pp. 152-158, 2022, doi: 10.1016/j.patrec.2022.09.022.

[5] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of personality and social psychology,* vol. 53, no. 4, p. 712, 1987.

[6] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: evidence from patients with shoulder pain," *Pain,* vol. 139, no. 2, pp. 267-274, Oct 15 2008, doi: 10.1016/j.pain.2008.04.010.

[7] G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker, and H. Wang, "Enhanced deep learning algorithm development to detect pain intensity from facial expression images," *Expert Systems with Applications,* vol. 149, 2020, doi: 10.1016/j.eswa.2020.113305.

[8] A. Semwal and N. D. Londhe, "A multi-stream spatio-temporal network based behavioural multiparametric pain assessment system," *Biomedical Signal Processing and Control,* vol. 90, p. 105820, 2024.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016 2016, pp. 770-778.

[10] G. Bargshady, "A Joint Deep Neural Network Model for Pain Recognition from Face.pdf," 2019, doi: 10.1109/CCOMS.2019.8821779.

[11] E. Othman, P. Werner, F. Saxen, A. Al-Hamadi, S. Gruss, and S. Walter, "Classification networks for continuous automatic pain intensity monitoring in video using facial expression on the X-ITE Pain Database," *Journal of Visual Communication and Image Representation,* vol. 91, 2023, doi: 10.1016/j.jvcir.2022.103743.

[12] A. Semwal and N. D. Londhe, "Automated Facial Expression based Pain Assessment Using Deep Convolutional Neural Network," presented at the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020.

[13] 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.

[14] Y. Huang, L. Qing, S. Xu, L. Wang, and Y. Peng, "HybNet: a hybrid network structure for pain intensity estimation," *The Visual Computer,* vol. 38, no. 3, pp. 871-882, 2022/03/01 2022, doi: 10.1007/s00371-021-02056-y.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2017 2017, pp. 2980-2988.

[16] L. Wang, C. Wang, Z. Sun, S. Cheng, and L. Guo, "Class balanced loss for image classification," *IEEE access,* vol. 8, pp. 81142-81153, 2020.

[17] P. Nabila and E. B. Setiawan, "Adam and AdamW Optimization Algorithm Application on BERT Model for Hate Speech Detection on Twitter," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 10-11 July 2024 2024, pp. 346-351, doi: 10.1109/ICoDSA62899.2024.10651619.

[18] Y. You *et al.*, "Large batch optimization for deep learning: Training bert in 76 minutes," *arXiv preprint arXiv:1904.00962,* 2019.

[19] J. Zhuang *et al.*, "Adabelief optimizer: Adapting stepsizes by the belief in observed gradients," *Advances in neural information processing systems,* vol. 33, pp. 18795-18806, 2020.

[20] B. Ginsburg *et al.*, "Training deep networks with stochastic gradient normalized by layerwise adaptive second moments," 2019.

[21] H. D. Nguyen, P. M. Le, and K. H. Truong, "Searching Optimal Placement and Operations of Energy Storage Systems based on Equilibrium Optimizer," *Engineering, Technology & Applied Science Research,* vol. 15, no. 4, pp. 24174-24180, 08/02 2025, doi: 10.48084/etasr.11238.

[22] A. Negi, M. G. Prasad, R. Kumar, H. K. Gupta, H. N. N. Kumar, and M. Shuaib, "Hybrid Feature Fusion based model: Facial Expression Recognition in the Wild," in *2025 5th International Conference on Intelligent Technologies (CONIT)*, 20-22 June 2025 2025, pp. 1-6, doi: 10.1109/CONIT65521.2025.11166706.

[23] 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.

[24] P. Thiam, H. A. Kestler, and F. Schwenker, "Two-Stream Attention Network for Pain Recognition from Video Sequences," *Sensors*, vol. 20, no. 3, doi: 10.3390/s20030839.

[25] Y. Mao, "Optimization of Facial Expression Recgonition on ResNet-18 using Focal Loss and CosFace Loss," 2022 2022: IEEE, pp. 161-163.

[26] S. Walter *et al.*, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *2013 IEEE International Conference on Cybernetics (CYBCO)*, 13-15 June 2013 2013, pp. 128-131, doi: 10.1109/CYBConf.2013.6617456.

[27] Zhang Shuyu and Chōshō, "AUC Maximization in Deep Neural Network Learning for Imbalanced Classification Problems," 2020.

[28] G. Bradski, "The openCV library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer,* vol. 25, no. 11, pp. 120-123, 2000.

[29] K. A. Majeed, Z. Abbas, M. Bakhtyar, Z. Durrani, J. Baber, and I. Ullah, "Face Detectors Evaluation to Select the Fastest among DLIB, HAAR Cascade, and MTCNN," *Pakistan Journal of Emerging Science and Technologies,* vol. 2, pp. 1-13, 2021.

[30] F. Zhang, C. Hu, Q. Yin, W. Li, H.-C. Li, and W. Hong, "Multi-aspect-aware bidirectional LSTM networks for synthetic aperture radar target recognition," *Ieee Access,* vol. 5, pp. 26880-26891, 2017.

[31] O. M. Alyasiri and Y.-N. Cheah, "Multi-Class Text Classification using Machine Learning Techniques," *Engineering, Technology & Applied Science Research,* vol. 15, no. 3, pp. 22598-22604, 2025.

[32] J. Fan, J. Zhou, X. Deng, H. Wang, L. Tao, and H. K. Kwan, "Combating Uncertainty and Class Imbalance in Facial Expression Recognition," 2022 2022: IEEE, pp. 1-4.