Applied Computer Science 2026, vol. 22, no. 1, pp. 214–227
https://doi.org/10.35784/acs_7747

Submitted: 2025-05-14 | Revised: 2025-09-24 | Accepted: 2025-10-17

Keywords: artificial intelligence, machine learning, facial recognition, transfer learning

Abeer A. Mohamad ALSHIHA [1]
[1] Remote Sensing Center, University of Mosul, Iraq, abeer.allaf@uomosul.edu.iq
[*] Corresponding author: abeer.allaf@uomosul.edu.iq

Quantifying pain: An AI-driven approach to detecting pain levels via facial expressions

Abstract

Accurate pain assessment remains a cornerstone of effective clinical care as it significantly influences diagnosis, treatment planning, and evaluation of therapeutic outcomes. Traditional pain assessment tools such as the Visual Analog Scale (VAS), Numerical Rating Scale (NRS), and Verbal Rating Scale (VRS) rely heavily on the patient's ability to self-report their level of discomfort. However, these conventional approaches are inadequate for patient populations with impaired communication abilities, including individuals with neurological disorders, dementia, or those in postoperative recovery. To overcome these challenges, this study presents a novel, automated pain assessment framework that uses artificial intelligence (AI) and facial expression analysis to objectively quantify pain levels. The proposed system incorporates transfer learning and deep neural network models to improve the accuracy of pain detection using facial cues. Using the UNBC-McMaster Shoulder Pain Expression Archive Database, a widely recognized benchmark in pain research, the model was trained to identify and classify facial expressions associated with different levels of pain. A key innovation of this research is the development of an enhanced multilevel pain scale, which extends the traditional ten-point scale to sixteen different levels, allowing for more precise and granular assessment. Despite the inherent problem of class imbalance within the dataset, the model achieved a commendable classification accuracy of 91%. The results highlight the viability of AI-based tools as reliable, non-invasive alternatives to traditional self-report methods, particularly for non-communicative patients. This advancement promises to improve patient care by supporting clinicians with objective, data-driven pain assessment techniques.

1. INTRODUCTION

Despite the amount of research on this phenomenon, pain as a sensory perception and subjective experience of the individual psyche still raises many unanswered questions about its nature. It can result from physiological conditions, trauma, and diseases of the absent, such as AIDS, Parkinson's disease, fractures, spinal pain, and chronic headaches, among others. Pain assessment is considered one of the most important aspects of medical practice as it helps the physician in decision making and increases patient satisfaction with the services provided (Bargshady, 2020).

To assess pain intensity in primary and secondary care clinics, the patient's perception is first documented using a structured pain assessment protocol that includes the following tools VAS, NRS, and VRS (Saddam et al., 2021). However, such self-report approaches suffer from some drawbacks, especially when patients cannot speak for their increased or decreased pain levels due to a neurological disorder or dementia, or after surgery (Leo et al., 2020).

In such cases, behavioral measures such as the Neonatal Infant Pain Scale (NIPS), the Pain Assessment in Advanced Dementia (PAINAD), and the Behavioral Pain Scale (BPS) are used to document and analyze the patient's movements and vocalizations, facial expressions, and other nonverbal signs. In addition, it is common to obtain feedback from caregivers and family members to assess the patient's pain experience (Raja et al., 2020).

However, some of the challenges that are likely to be experienced when using these methods include physician fatigue and misinterpretation of pain signals. Criticizing the lack of accuracy and standardization in pain assessment methods, scientists have begun to focus on AI solutions. Computer and AI systems can analyze more specific details of the face and its expressions, such as micro-expressions, movements, and variations, to detect and measure the level of pain in a much better way (Janssen, 2021; Alshiha et al., 2023).

Therefore, to fill the above gaps, this study proposes to design an automated multi-level pain assessment system using AI to consider pain levels beyond the six levels, from level 7 to 10. In order to expand the field of machine learning pain analysis, this work aims to improve the accuracy of facial pain detection by using a number of advanced approaches of machine learning, including transfer learning to improve the ability of extracting pain features from the face, training neural networks on a richer set of pain levels, and calculating the overall effect of the dataset balance. The possibilities of AI in pain assessment cannot be ignored, as it provides accurate analysis that can reduce the burden on the healthcare facility, helps to improve patient care, and leads to the improvement of the quality of pain management in various capacities (Al-Neama et al., 2023).

2. RELATED WORKS

Recent developments in upper-face motion recognition have made great strides in increasing the likelihood of identifying emotion and pain from facial images. Daniel et al. (2017) proposed a new technique that uses the triangle sizes and angles extracted from the facial feature photographs to determine the manner in which facial expressions are likely to be performed, with the help of a modified K-Nearest Neighbors (KNN) classifier. Their method showed reasonable reliability and accuracy in identifying the expressions, especially when validated on the CK+ dataset compared to other existing technologies (Acevedo et al., 2017).

Similarly, Shier (2017) created a facial expression-based pain analysis using CNN and Gabor with SVM. Their work, which utilized the Pain Severity and Intensity (PSPI) scale, supports the idea that patients can be sorted according to pain type in real-world settings such as elderly care facilities. On this basis, Reneiro et al. (2019) analyzed the automated recognition of facial expressions, focusing on the assessment of multiple facial expression datasets as a basic tool for evaluation. They focused on thought-provoking aspects of the features of facial expressions that affect pain, and practiced correspondingly high recall rates, such as 85.66% success with existing Facial Expression Recognition (FER) databases (Andal Virrey et al., 2019).

Subsequently, Nour et al. (2020) focused on the use of Deep Convolutional Neural Network (DCNN) models for FER systems using AlexNet, VGG-16, and ResNet models. In their study, on the Extended Cohn-Kanada (CK+) databases, the authors concluded that the AlexNet model was the best with an accuracy of 88.2%, thus contributing to the improvement of facial expression recognition in various applications (Nour et al., 2020).

Each of these studies demonstrates the development and efficacy of computational methods for facial analysis of emotion and pain, advancing theoretical and practical knowledge in clinical and healthcare settings. Andersen et al. went further in 2021 to investigate ways to improve automatic pain detection from facial expressions. Their method of approach involved a facial movement coding system, which is a sensitive manual way of categorizing facial activity, and was combined with the Absi et al. Machine learning principles were applied in the identification of pain expressions from the facial movement coding system. This method proved to be the best for detecting action units (AUs) directly from facial photographs, since tools that help to mark important facial features are reliably resourceful. In their second study, they examined the recurrent neural networks trained on the large video datasets with the ground truth annotations, emphasizing that temporal dynamics play an important role in assessing pain intensity.

Similarly, Morabit et al. (2021) compared the best CNN architectures MobileNet, GoogleNet, ResNeXt-50, ResNet 18, and DenseNet -161. They evaluated these models individually for direct pain assessment and as feature makers for classifiers such as SVR and RFR. Using the UNBC-McMaster Shoulder Pain Database, their study also demonstrated the ability of deep CNN layers for pain detection, with promising implications for future work in automatic pattern recognition (El Morabit et al., 2021).

Another study by Saddam et al. in 2021 dealt with the difficulties in estimating the level of pain from facial expressions, so the authors proposed a systematic method for designing and implementing an automatic pain detection system. It covered data acquisition and processing techniques as well as classifier training methods. It emphasized the multimodal approach for pain identification in healthcare environments by using facial expressions, gestures, and vocalizations.

Although other studies have also shown the viability of machine learning and deep learning methods in measuring pain and recognizing facial expressions, there are some core innovations in this study that set it apart from others. One of the most important contributions is the expansion of the traditional pain rating scale to sixteen levels, which helps to more accurately account for pain intensity, especially in non-communicative patients. Such a high level of resolution overcomes the drawbacks of traditional ten-level or binary scales and allows the therapist to detect subtle changes in the patient's discomfort that would otherwise go unnoticed. The second novelty is the use of the FaceNet architecture, which has been trained on face recognition tasks to be used in the context of pain analysis. The proposed system takes advantage of the feature extraction system provided by FaceNet and embedding-based similarity to detect small facial muscle movements and micro-expressions associated with different levels of pain. Taken together, these advances provide an unprecedented level of high-resolution quantification of pain and robust AI-based facial features, making the system a breakthrough compared to the previous approaches and can increase the value of its clinical application.

3. FACIAL ACTION CODING SYSTEM (FACS)

The Facial Action Coding System (FACS) is a comprehensive and anatomically based method for measuring all observable facial movements. It accomplishes this by categorizing each visually observable facial activity into 45 distinct Action Units (AUs), taking into account various head and eye postures and movements. Each AU is associated with a numerical code, the assignment of which is somewhat arbitrary but serves as a standardized reference for facial muscle movements (Baumeister & Vohs, 2012).

Action Units (AUs) represent precise facial muscle movements that convey a wide range of emotions and expressions. FACS, developed by Paul Ekman and Wallace Friesen, identifies a total of 45 different AUs, each corresponding to specific movements or combinations of movements in the facial muscles. Table (1) describes these action units.

Table 1. Face action unit

Action unit	Description	Facial muscle
1	Raising Inner Brow	Frontalis, pars medialis
2	Raising Outer Brow (unilateral, right side)	Frontalis, pars lateralis (right side only)
4	Lowering Brow	Depressor Glabellae, Depressor Supercilli, Currugator
5	Raising Upper Lid	Levator palpebrae superioris
6	Raising Cheek	Orbicularis oculi, pars orbitalis
7	Tightening Lid	Orbicularis oculi, pars palpebralis
9 (also shows slight AU4 and AU10)	Wrinkling Nose	Levator labii superioris alaquae nasi
10 (also shows slight AU25)	Raising Upper Lip	Levator Labii Superioris, Caput infraorbitalis
11	Deepening Nasolabial	Zygomatic Minor
12	Pulling Lip Corners	Zygomatic Major
13	Puffing Cheeks	Levator anguli oris (Caninus)
14	Creating Dimples	Buccinator
15	Depressing Lip Corners	Depressor anguli oris (Triangularis)
16 (with AU25)	Depressing Lower Lip	Depressor labii inferioris
17	Raising Chin	Mentalis
18 (with slight AU22 and AU25)	Puckering Lips	Incisivii labii superioris and Incisivii labii inferioris
20	Stretching Lips	Risorius
22 (with AU25)	Funneling Lips	Orbicularis oris
23	Tightening Lips	Orbicularis oris
24	Pressing Lips	Orbicularis oris
25	Parting Lips	Depressor Labii, Relaxation of Mentalis, Orbicularis Oris
Action Unit	Description	Facial Muscle

Tab. 1. Face action unit, continued

Action unit	Description	Facial muscle
26 (with AU25)	Dropping Jaw	Masetter; Temporal and Internal Pterygoid relaxed
27	Stretching Mouth	Pterygoids, Digastric
28 (with AU26)	Sucking Lips	Orbicularis oris
41	Drooping Lid	Relaxation of Levator Palpebrae Superioris
42	Narrowing Eyes	Orbicularis oculi
43	Closing Eyes	Relaxation of Levator Palpebrae Superioris
44	Squinting Eyes	Orbicularis oculi, pars palpebralis
45	Blinking Eyes	Relaxation of Levator Palpebrae and Contraction of Orbicularis Oculi, Pars Palpebralis
46	Winking Eye	Levator palpebrae superioris; Orbicularis oculi, pars palpebralis

Each Action Unit (AU) within the Facial Action Coding System (FACS) can interact with others, resulting in a wide range of facial expressions, each intricately linked to a specific emotion. For example, the combination of AUs 1+2+5+12 corresponds to happiness, while AUs 1+2+4+15 indicate a sad expression (Ekman & Rosenberg, 2012). In addition, FACS coding methods can take into account a number of factors, such as the timing of emotions on the face, the strength of each facial action measured on a five-point scale, and the coding of facial expressions as separate "events". Each facial expression is represented by an event, which may consist of a single AU or a group of AUs compressed into a single expression.

4. METHODOLOGY

In recent years, machine learning (ML) and deep learning (DL) have rapidly emerged as powerful technologies for analyzing complex medical data, with promising transformative potential in healthcare research and practice (JAMES & Osubor, 2025). Compared to conventional statistical approaches that assume a linear relationship between variables and may fail to identify subtle interactions between variables, ML and DL models have the ability to model highly nonlinear and latent patterns in multidimensional data (Na & Kim, 2024). All of these approaches have been successfully implemented in a variety of medical problems, such as disease diagnosis and prognosis, personalized treatment planning, medical imaging interpretation, and real-time patient monitoring. For example, deep neural networks that can detect small differences in imaging data, and ML algorithms that can combine a variety of clinical data sources (electronic health records, genomics, and sensor-based physiological measurements) to reveal correlations that would otherwise be difficult to detect. The flexibility of these methods allows one to combine multimodal data to increase predictive power and clinical significance (Machrowska et al., 2024). Placing the proposed automated multi-level pain measurement system in this context shows the scientific soundness and real-life applicability of the system (Karpiński et al., 2023). The system is indicative of the trend of AI application in healthcare by potentially providing objective and data-driven measures to complement typical clinical assessments, improve patient care, and support evidence-based decisions in a situation where traditional self-report-based methods have limited applicability.

This section describes the tools and methods to be used in the proposed system, which is based on transfer learning for feature extraction from facial images using the FaceNet algorithm. It is developed as a method for training a classifier neural network using the UNBC-McMaster Shoulder Pain Expression Archive as a database. The approach is divided into two primary assessment categories: Multi-Level Pain Assessment (MLPA), which is a popular assessment tool used in various healthcare settings.

4.1. UNBC-shoulder pain dataset

The UNBC-McMaster Pain Expression Archive Database (UNBC-McMaster) is a video database of people in pain who are shown performing arm range-of-motion tests while experiencing shoulder pain. Participants were recruited from three PT clinics and the McMaster University campus. Inclusion criteria were determined by a subjective complaint of shoulder pain; diagnoses could include arthritis, bursitis, tendinitis, subluxation, rotator cuff injury, impingement syndrome, osteoarthritis, capsulitis, and shoulder dislocation (Kaltwang, 2015).

The recordings capture two types of actions: The first scenario refers to the act of the person with the pathology moving his arm with or without the help of the physiotherapist, but the second scenario refers to the immobilization of the patient and the subsequent manipulation of the patient's arm. Nevertheless, movements of another arm and a control set were also observed when the pain was localized in only one arm. We evaluated the proposed method on the sixteen selected sequences, with 200 sequences obtained from 25 subjects, containing 48,398 frames. For each frame, in the case of pain-related AUs, AUs 4, 6, 7, 9, 10, 12, 20, 25, 26, and 43 are reported ordinally on the intensity scale ranging from 0 to 5, while for AU 43 the scale is binary.

Experienced FACS coders defined the AU labels, and the creators of the database also provided separate pain intensity levels according to the Prkachin and Solomon approach. Figure 1 shows the distribution of the 16 different pain intensity levels (0-15) observed.

4.2. FaceNet model

FaceNet uses a specially designed neural network called a Siamese network, which compares two facial images and produces a similarity measure. This network is trained on a large dataset of facial images to distinguish one face from another. Another good feature of FaceNet is that it uses a triplet loss function. This function helps the network learn to produce embeddings that properly separate the faces of individuals, while ensuring that embeddings of the same person remain as close together as possible. This technique improves the model's face recognition ability and is particularly useful when there are few training samples of a person's face (Alshiha et al., 2023; Al-Neama et al., 2025). FaceNet also applies some acceleration strategies, such as data augmentation and online triplet mining methods, where the method searches for the most difficult negative samples for training (Schroff et al., 2015).

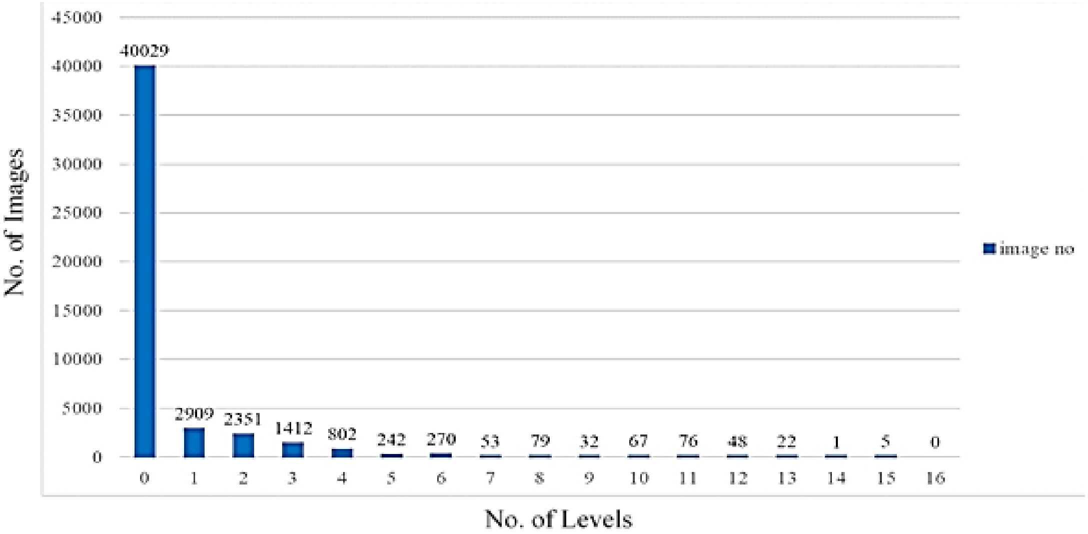

Fig. 1. Distribution of frames across pain intensity levels

4.3. Research hypotheses

To systematically evaluate the effectiveness of the proposed system, four hypotheses were developed:

Hypothesis 1. Making the pain scale more granular with sixteen levels instead of the standard ten levels will improve the sensitivity of the system by allowing it to detect the small facial micro-expressions associated with the moderate levels of pain.

Hypothesis 2. Using transfer learning with the FaceNet embedding architecture will provide high quality feature extraction performance than the traditional CNN models trained using the Who Wants to Be Great training, which will improve the classification accuracy on the multi-level pain recognition.

Hypothesis 3. Balanced datasets will give a more stable and slightly less accurate performance compared to unbalanced datasets, due to the fact that balancing decreases the dominant class rates, but also limits the sample size of any given class.

Hypothesis 4. The ability of a multi-layer perceptron (MLP) model to capture nonlinear relationships in the data is highly dependent on the number of hidden layers in the model. Although deeper structures have the potential to learn more complicated functions, they increase computational cost and training difficulty.

4.4. Model block diagram

The system processes are carried out in four phases, as shown in Figure 2. In the initial phase, data is collected, processed, and categorized into four groups. There are four types of databases: balanced binary, unbalanced binary, unbalanced multilevel, and balanced multilevel. The second phase involves feature extraction using the FaceNet algorithm. The third phase is based on the unique training of the neural network on the binary and multilevel parameters of the pain assessment. Finally, the system analyzes and determines the level of pain.

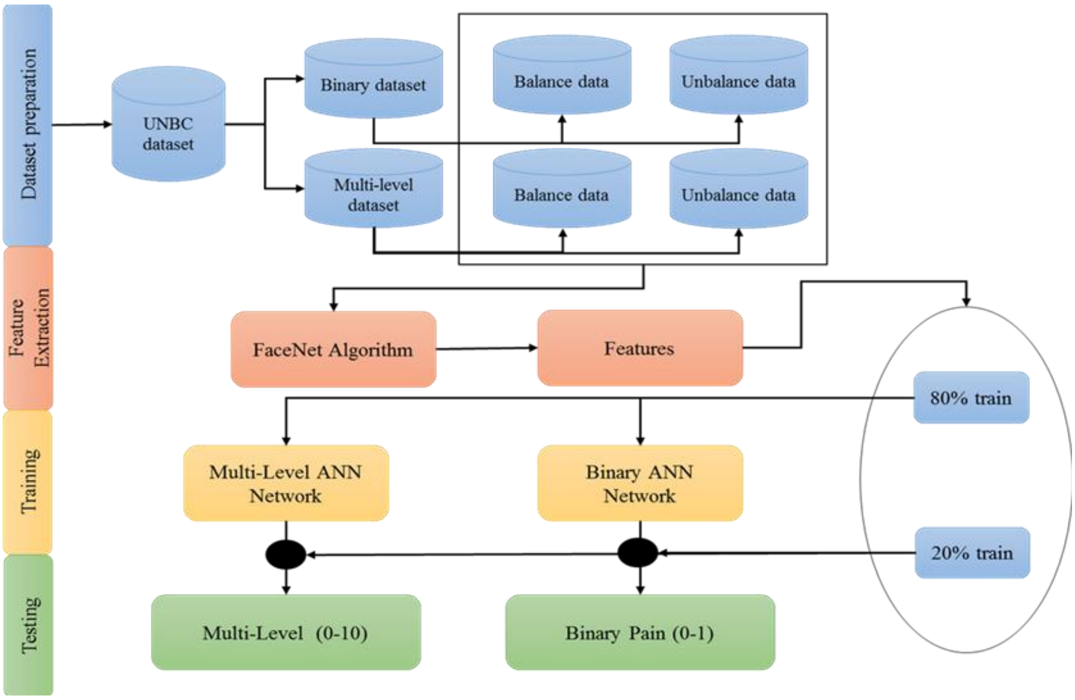

Fig. 2. Model block diagram

4.5. Dataset preprocessing

The images in the UNBC dataset were organized into groups based on pain level. Each group was placed in a separate folder according to the corresponding pain level. This sorting process required a specialized program to categorize the images according to the pain levels they represented. In this study, Excel was used along with a custom function that uploaded images into Excel, sorted them by pain level, and then stored them in folders named by pain level. The data was split 80% for training and 20% for testing. All data was processed through the FaceNet algorithm to extract features for each image. This process was performed using a feature in the Colab editor, a paid feature that allows the program to continue running even when the device is turned off, due to the large amount of time required to extract feature vectors from 48 000 images.

The data was divided into two groups to ensure that each group contained images of the same individuals at each pain level. The data set consists of images of twenty-five patients, each experiencing different levels of pain. For example, pain level 14 contains images of only one individual. Therefore, it was not possible to include images of individuals in the same pain level group. To ensure that the machine learning model could be trained on the maximum number of individuals and images with consistent characteristics, nearly equal numbers of images were selected for each individual across pain levels. The dataset was divided into Balanced Multilevel Pain Distribution (BMPD) and Unbalanced Multilevel Pain Distribution (UMPD).

The improved version of the multilevel pain scale, which extends the conventional ten-level scale to sixteen levels, was developed indirectly with expert assistance using the UNBC McMaster dataset. Pain intensity levels in the dataset, along with action units (AUs), were assigned by trained coders of the FACS, who are experienced in accurately identifying the facial muscle movements of pain. Although there is no explicit mention of further consultation with physicians to expand the scale to 16 levels, the validated FACS-coded labels ensure that the scale includes clinically relevant facial signs of pain. Increasing the scale will provide a more detailed view of the nuances of facial expressions, which is very helpful in the case of non-communicative patients. In the clinic, such increased accuracy will allow clinicians to observe small variations in pain that can make a difference in treatment regimens and patient outcomes.

While the UNBC dataset consists of 200 sequences with a total of 48,398 frames across 25 subjects, only 16 sequences were used in the study to evaluate the proposed model. The selection of these sequences was based on the need for a balanced representation and sufficient data per person and pain level. The fewest number of images represent some pain levels, and only one patient represents a few levels, so it is not feasible to include all sequences in the model training because it is likely to overfit. The chosen set of sequences was chosen to provide maximum variability in subjects and pain levels, but consistency and quality of facial expression. There is also a factor of computational efficiency: running the 200 sequences will be resource intensive when using FaceNet to extract features. Since the authors selected 16 highly representative sequences, the multilevel pain rating system had a sufficient and representative sample to train and evaluate.

The data set used in this study was derived from the UNBC-McMaster Shoulder Pain Expression Archive Database. To ensure reproducibility and transparency, the following official and secondary repositories were explicitly consulted (Cohn & Schmidt, 2013; Papers With Code, n.d.; IEEE DataPort, n.d.).

It is necessary to explain that there was an error in the previous version of the draft where the phrase average height was inappropriately chosen as a place variable. In the current version, this term has been replaced because no anthropometric or body measurement treatments were collected. Only facial expressions are analyzed, and the datasets contain only image frames characterized by a label of facial action units and pain intensity.

4.6. Neural network structure

As a predictor of pain levels, BMPD, which is Balanced Multi-Level Pain Distribution, was used, and another used is UMPD, which is unbalanced Multi-Level Pain Distribution. This data set consisted of the ten levels of pain starting from no pain to level 10. First, two original datasets BMPD and UMPD were used as input source, followed by data preprocessing and splitting into training and test sets. In the second phase, features were extracted in the same way as the binary pain type assessment using FaceNet.

The final phase used a feed-forward backpropagation fully connected Artificial Neural Network (ANN), shown in Figure 3.

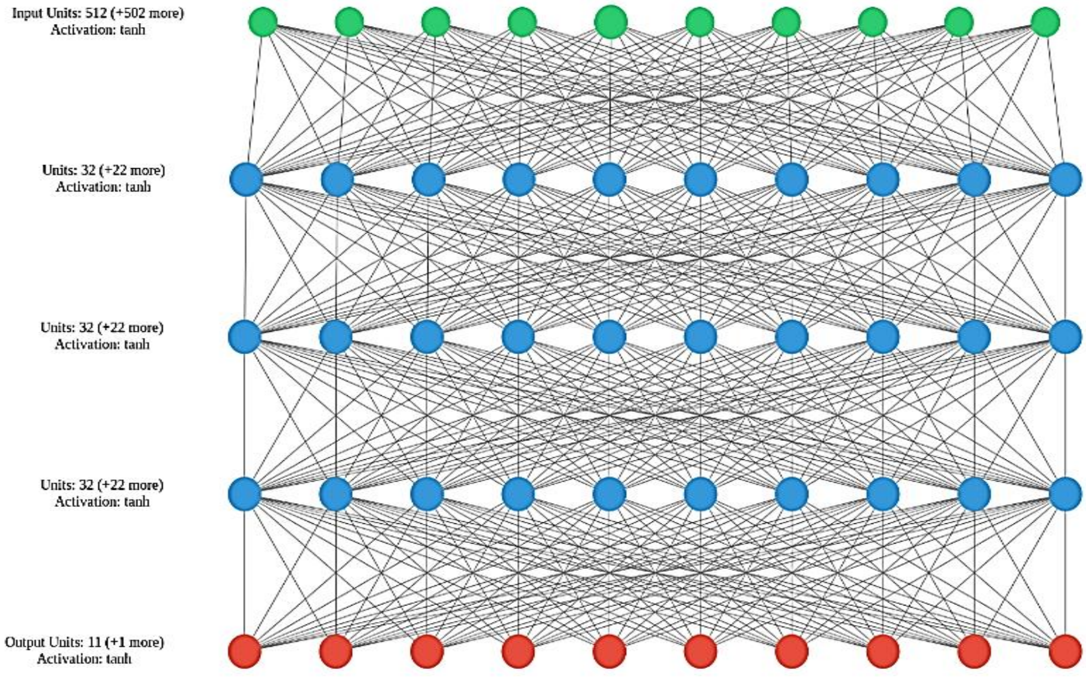

Fig. 3. Neural network structure

This ANN had five layers with 512 inputs and ten outputs. This type of neural network, also known as a multilayer perceptron neural network, is designed to allow information to flow unidirectionally from inputs to outputs, but not to contain cycles. Hypothesis 4: The number of hidden layers within an MLP significantly determines the model's ability to identify relationships within the data and to generalize to unseen data. In general, structures with more hidden layers are capable of learning more complex functions; the drawback is that they are computationally expensive and difficult to train.

4.7. Training parameters

A detailed summary of the training parameters and hyperparameter settings for both unbalanced and balanced multilevel pain assessment models is summarized in Table (2). Important parameters are listed in the table and include number of epochs, batch size, learning rate, optimizer, activation functions, loss function, weight initialization strategy, dropout status, and early termination criterion. Providing these parameters makes the study reproducible, allowing other researchers to reproduce the training procedure and verify the results. It also provides a clear comparison between the unbalanced and balanced versions, showing that it is the dataset distribution that is largely different, along with the overall architecture and most of the hyper-parameters that remain the same but affect the performance of the models.

Tab. 2. Training parameters

Parameter	Values	Description / Notes
Number of Epochs	50	Total training iterations over the dataset
Batch Size	32	Number of samples per gradient update
Learning Rate	0.001	Initial step size for optimizer
Optimizer	Adam	Optimization algorithm
Activation Function (Hidden)	ReLU	Non-linear activation for hidden layers
Activation Function (Output)	Softmax	Converts logits to probability distribution
Loss Function	Categorical Cross-Entropy	Measures model error
Weight Initialization	He Normal	Method to initialize network weights
Dropout Rate	0.2	Regularization to prevent overfitting
Early Stopping	Patience=10	Stops training if validation loss does not improve

5. RESULTS

The methodology and description of the datasets used in the study do not mention any of the model evaluation sites, so this statement is vague and confusing. A more precise explanation is that the authors wanted to discuss three evaluation measures, such as accuracy, precision, and recall, which are the common measures used to evaluate the results of machine learning models. These measures give a clear idea of whether the model is accurate in predicting the amount of pain, how often it predicts positively, and the ability to find the positive cases. The phrase may alternatively have been intended to imply the existence of three types of data set distributions as used in the study, namely balanced binary, unbalanced binary, and multilevel pain data set distributions, but the surrounding text does not provide particular support for this variant. In any case, the current wording of the statement should be revised to reflect the proper assessment measures or comparison of the data sets.

In addition to reporting accuracy, precision, recall, and loss, we measured some computational performance metrics to gauge usefulness. On average, each epoch took 42-45 seconds, depending on the type of dataset. The total training time of 50 epochs was 0.6-0.65 hours. Based on the speed in the testing process, it was found that the system can be used in real-time applications with the approximate speed of 18-19 milliseconds per sequence. A summary of these results is shown in Table 3.

Tab. 3. Average training duration and inference speed

Model Type	Average Epoch Duration (sec)	Total Training Time (hrs)	Inference Speed (ms/sequence)
Balanced Multi-Level	42.5	0.59	18.3
Unbalanced Multi-Level	44.7	0.62	19.1

The average epoch took about 42 seconds to 45 seconds, depending on the type of data. The total training time was 0.65-0.6 hours, divided into 50 epochs. The inference time during testing was approximately 1819 ms per sequence, which is real time. Table 3 summarizes these details.

The Unbalanced Multi-Level Pain Assessment model was found to be very accurate, reaching 91.92%, which shows that it made the correct prediction in most of the cases. To reiterate, the process presented a very low loss value of 0.0453, indicating a meager level of deviation from the real pain levels, making the model a great fit for the dataset. Therefore, the accuracy of the model of the above three sites is 94.41% positive predictions, and the recall rate of the model is 89.41%, showing good capacity in accurately identifying true positives of the pain levels. In summary, these results show the beneficial presence of the proposed model in multi-level pain assessment, since it provides remarkable predictive accuracy and precision measures, although the recall factor is slightly compromised, which would still be reasonable for judging the precise level of the patient's pain. Figure 4 shows the training curves.

The Balanced Multi-Level Pain Assessment model obtained an accuracy of 83.19%; thus, the Balanced Multi-Level Pain Assessment model computationally achieves a slightly lower but still reasonably high correct prediction rate compared to the unbalanced case. The loss value equal to zero or 0.0833 indicates the reasonable value of the error rate, which means that the model rather fits the actual pain level in the dataset. 90% of the patients evaluate the professional service of the health care organization with a precision of 85.59%, revealing a high reliability of the proposed model in deciding the positive results, since the retrieval rate or recall of positives was established at 80.49%, which means it can effectively identify the genuine instances of the dependent variable, which is the level of pain. In summary, when evaluating the results presented above, it can be said that the balanced model is slightly less accurate than the unbalanced one. However, it also maintains a fairly high level of performance in terms of precision and recall, which still makes it an efficient tool in the case of multi-level pain assessment. Figure 5 shows the training curve.

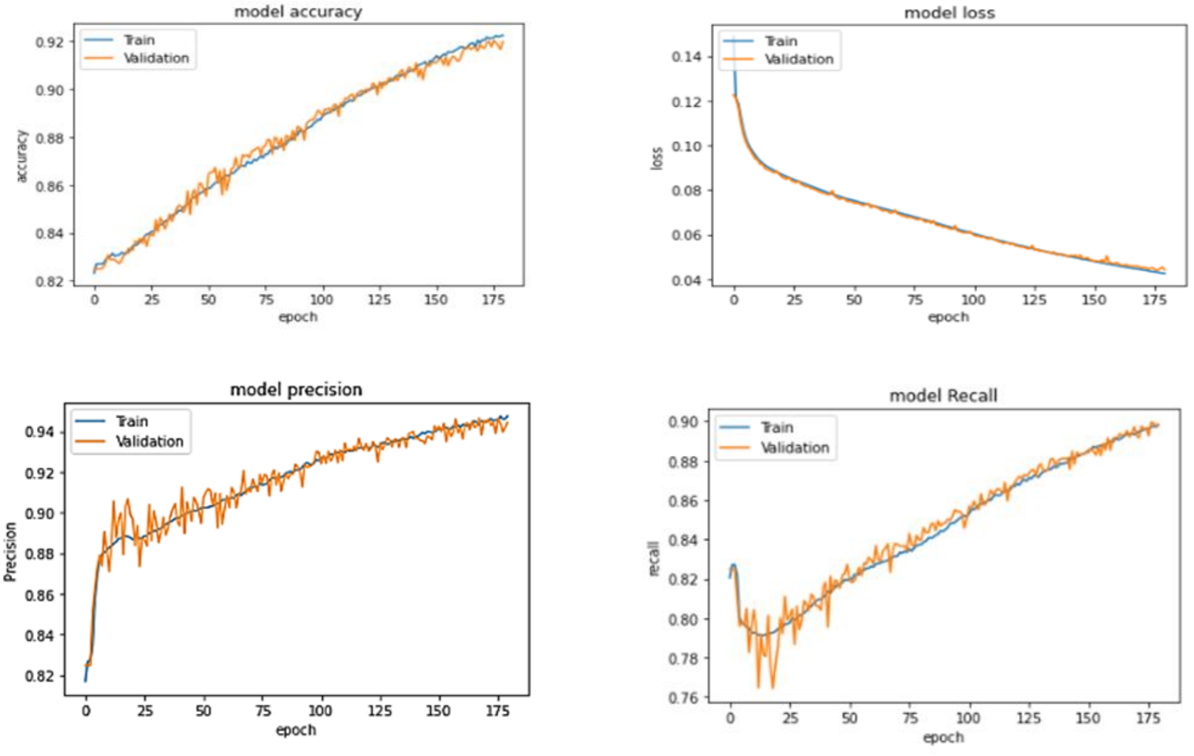

Fig. 4. Unbalance dataset training history

The difference between Multilevel Unbalanced and Balanced Pain Assessment models clearly shows that the multilevel model type is more accurate and efficient in all analyzed aspects. When compared with the balanced model, it achieves 91. 92% accuracy instead of 83. 19% with less loss of 0. 0453 instead of 0. 0833, more precision with 94. 41% instead of 85.59%, and higher recall with 89.41% instead of 80.49%.

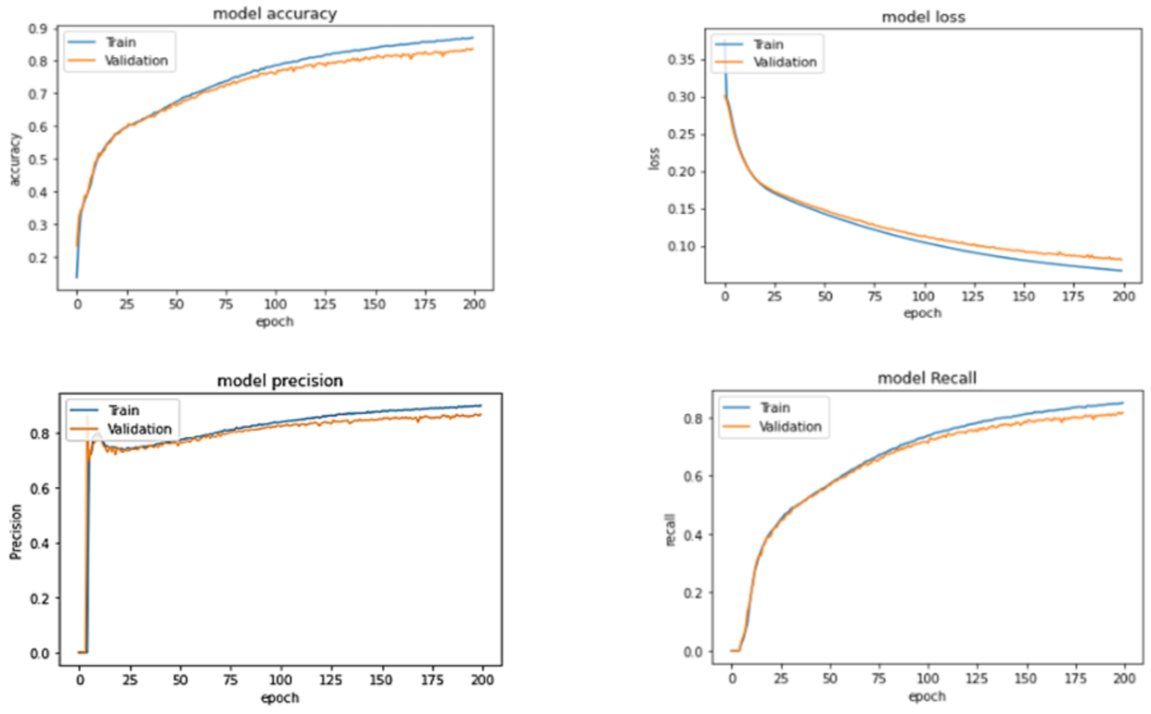

Fig. 5. Balance dataset training history

These results suggest that with this unbalanced model is preferred in the correct assessment of the pain level and in the detection of the true positive cases, the main reason probably because in the process of the true positive cases there are more data available in certain pain level categories. In this case, the idea of the balanced model is less sensitive, but at the same time it remains rather stable, when certain tests and standards are applied, the performance rates slightly decrease, probably due to the relative limitations of having much more evenly distributed, but not necessarily overly extensive, data material. Figure 6 shows the comparison of results.

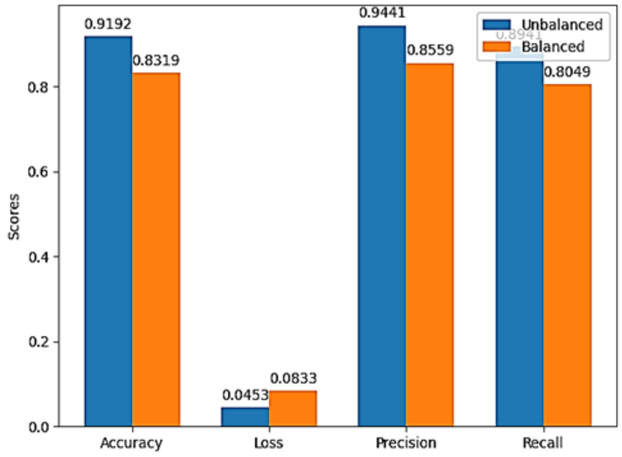

Fig. 6. Performance Metrics by Model Type

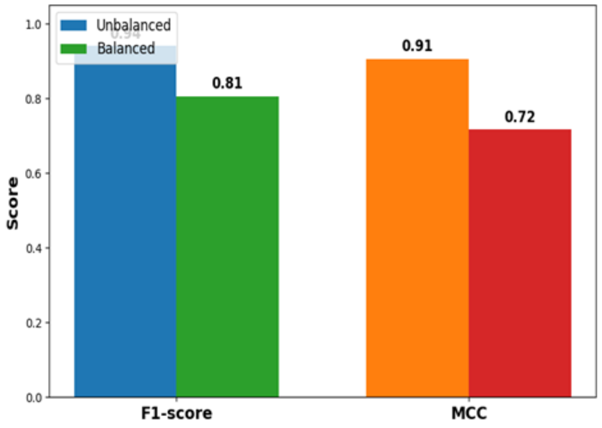

Fig 7. Comparison of Unbalanced Vs Balanced Model Performance (F1-score & MCC)

Figure (7) and applies two additional performance measures to a comparison of the Unbalanced and Balanced Multi-Level Pain Assessment models: F1 score and Matthews Correlation Coefficient (MCC). These measures are particularly insightful with respect to unbalanced datasets, as they provide a better evaluation of classifier performance compared to the use of accuracy. F1-score is used to show the trade-off between precision and recall at each of the pain levels, while MCC evaluates the overall quality of all predictions (true and false positives and negatives). The figure uses bold colors to show the difference between the unbalanced and balanced models, and profiles the metrics above each bar for clarity.

The results indicate that the unbalanced model has a high F1 score and MCC in contrast to the balanced model. The superior F1 score also suggests that the unbalanced model has a better trade-off between precision and recall by identifying the true positives in all pain classes. Similarly, the increased MCC validates the unbalanced model, which has a better overall classification performance with good and bad predictions. These results indicate the influence of data imbalance; the unbalanced model has a larger sample size in the dominant pain levels, but despite the increased fairness across classes, the balanced model has slightly decreased performance. The inclusion of metrics such as F1-score and MCC allows to illustrate not only the better accuracy of the unbalanced model, but also its ability to generalize better across the unbalanced categories, proving its usability in practice in the real clinical situation where some levels of pain may be more common.

Figure (8) shows a more detailed comparison of the unbalanced and balanced multilevel pain rating models based on four different metrics: Accuracy, Precision, Recall and Loss (inverted for visualization). The 2x2 grid shows the averaged performance of repeated experimental runs, with error bars indicating standard deviations. Significant signs above the bars indicate the statistical significance of independent t-tests or whether the differences between the models are significant ($p < 0.05$, $p < 0.01$, $p < 0.001$, ns = not significant). As shown in the figure, regardless of the metric used, the unbalanced model consistently achieves better accuracy, precision, recall, and lower loss, confirming its greater ability to determine multilevel pain. This visualization effectively expresses the degree of difference (as well as the statistical confidence in the results).

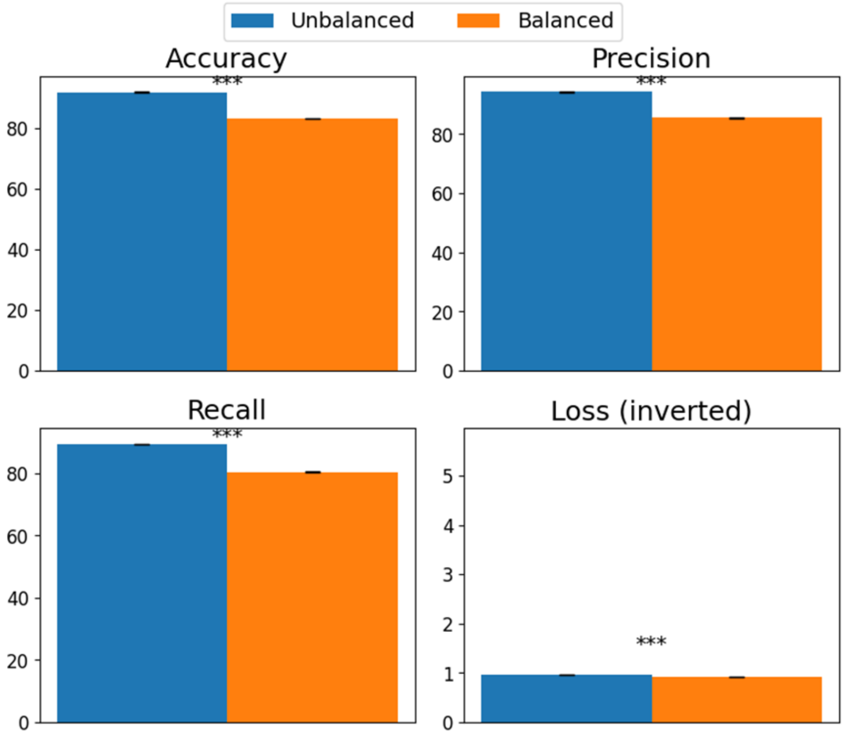

Fig 8. Comparison of unbalanced vs. balanced models

Although the conclusions of this study demonstrate the efficacy of the unbalanced and balanced multilevel pain assessment models, a number of limitations must be noted. First, the number of subjects in the dataset used is relatively small, and this fact can be considered as a limitation of the generalizability of the results. Second, the models were not tested on external datasets, which casts doubt on their generalizability to other settings or populations. Third, the sample in question is demographically limited, which may prevent the widest applicability of the model to other patients of different ethnic or age groups.

To overcome these shortcomings, future studies could include data augmentation to artificially increase the size of the dataset and strengthen the robustness of the model. In addition, the experiment should be tested on independent, larger, and more diverse datasets to provide more evidence of generalizability. The inclusion of demographically diverse participants would also help to make the models reliable in other patient populations. By recognizing such shortcomings and suggesting practical measures that can be taken to address them, future researchers will be able to incorporate the existing study of pain assessment models to create more universal and complete assessment methods.

Despite the fact that the study shows that the multilevel pain assessment models are promising in cases with non-communicating patients, further clinical analysis is needed. Possible scenarios include integration into intensive care units (ICUs), pediatric or geriatric wards, where patient self-report may not be possible. It would be possible to integrate the system with any existing medical monitoring system, including electronic health records (EHRs), wearable sensors, or patient monitoring dashboards, to provide real-time pain measurement and alerts to medical staff. Ethical concerns are essential, such as the fact that automated pain assessment should be viewed as an adjunct or supplement to clinical judgment, not a replacement. In addition, sensitive patient data cannot be obtained without regard to privacy laws (e.g., HIPAA or GDPR), such as the storage of this data, the protection of anonymized data, and in some cases, the anonymization itself. Addressing these factors will allow the model to be safely and effectively implemented in real-world healthcare settings, without losing patient trust or clinical soundness.

6. CONCLUSIONS

The proposed automated multi-level pain assessment system has made significant progress in applying artificial intelligence and machine learning to the objective assessment and detection of pain. The study shows that the unbalanced multi-level pain assessment model gave an overall accuracy rate of 91.92%, while for the balanced model it was 83.19%. Similarly, the loss for the unbalanced model was 0.0453, while for the balanced model it was 0.0833. These results highlight the robustness of AI-based methods in overcoming the limitations of traditional self-report and behavioral pain assessment techniques, especially for non-verbal patients. Transfer learning and neural network training also enhance the system's ability to extract and analyze facial pain features, resulting in a robust platform for objective and nuanced pain assessment.

As encouraging as these results are, several limitations must be highlighted. First, the study used a relatively small subset of sequences from the UNBC-McMaster Shoulder Pain Database, which may limit the generalizability of the findings to more general patient populations. Second, the system is primarily facial-based and does not incorporate other physiological or behavioral indicators of pain that may be a source of additional predictive capacity. Third, the uneven distribution of the data set, while helpful for model accuracy in some cases, may limit the sensitivity of the system to less dominant pain levels.

As a basis for further research, the use of multimodal data - vocalizations, body movements, and physiological signals - can improve the accuracy and usability of the system. Current pain measurement techniques that allow real-time monitoring and dynamic adjustment of patient pain levels represent another avenue of investigation. In addition, expanding the dataset to include a wider range of patients and clinical scenarios would improve the generalizability of the model and facilitate its use in real-world healthcare settings. In summary, this AI system represents a good starting point for improving pain management strategies and patient care in a variety of clinical contexts.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

Acevedo, D., Negri, P., Buemi, M. E., Fernandez, F. G., & Mejail, M. (2017). A simple geometric-based descriptor for facial expression recognition. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 802–808). IEEE. https://doi.org/10.1109/FG.2017.101

Al-Neama, M. W., Abdulrahman, E. H., & Ali, S. M. (2025). A parallel algorithm for facial expression recognition for student engagement monitoring in classrooms. *Mathematical Modelling of Engineering Problems*, *12*(3).

Alshiha, A. A., Al-Neama, M. W., & Qubaa, A. R. (2023). Biometric face recognition method using graphics processing unit system. *Indonesian Journal of Electrical Engineering and Computer Science*, *30*(1), 183–191. https://doi.org/10.11591/ijeecs.v30.i1.pp183-191

Andal Virrey, R., De Silva Liyanage, C., Iskandar bin Pg Hj Petra, M., & Emeroylariffion Abas, P. (2019). Visual data of facial expressions for automatic pain detection. *Journal of Visual Communication and Image Representation*, *61*, 209–217. https://doi.org/10.1016/j.jvcir.2019.03.023

Andersen, P. H., Broomé, S., Lahrmann, M., Rashid, M., Nyström, M., Lundblad, J., Ask, K., & Gleerup, K. B. (2021). Towards machine recognition of facial expressions of pain in horses. *Animals*, *11*(6), 1643. https://doi.org/10.3390/ani11061643

Bargshady, G. (2020). *Enhanced deep learning predictive modelling approaches for pain intensity recognition from facial expression video images* [Master's thesis, University of Southern Queensland].

Baumeister, R. F., & Vohs, K. D. (2012). Facial expression of emotion. In V. S. Ramachandran (Ed.), *Encyclopedia of social psychology*. SAGE Publications. https://doi.org/10.4135/9781412956253.n209

Cohn, J. F., & Schmidt, K. L. (2013). *UNBC–McMaster shoulder pain expression archive database* [Data set]. Department of Computer Science, University of Western Ontario. Retrieved from https://www.csd.uwo.ca/faculty/elaine/UNBCMcMaster/

Ekman, P., & Rosenberg, E. L. (Eds.). (2012). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195179644.001.0001

El Morabit, S., Rivenq, A., Zighem, M. E. N., Hadid, A., Ouahabi, A., & Taleb-Ahmed, A. (2021). Automatic pain estimation from facial expressions: A comparative analysis using off-the-shelf CNN architectures. *Electronics*, *10*(16), Article 1926. https://doi.org/10.3390/electronics10161926

IEEE DataPort. (n.d.). *UNBC-McMaster Pain Expression Database* [Data set]. Retrieved March 23, 2026, from https://ieee-dataport.org/open-access/unbc-mcmaster-pain-expression-database

James, I., & Osubor, V. (2025). Machine learning evidence towards eradication of malaria burden: A scoping review. *Applied Computer Science*, *21*(1), 44–69.

Janssen, B. (2021). *Pain by association: Role of individual difference variables* [Unpublished manuscript].

Kaltwang, S. (2015). *Regression-based estimation of pain and facial expression intensity* [Doctoral dissertation, Imperial College London].

Karpiński, R., Krakowski, P., Jonak, J., Machrowska, A., & Maciejewski, M. (2023). Comparison of selected classification methods based on machine learning as a diagnostic tool for knee joint cartilage damage based on generated vibroacoustic processes. *Applied Computer Science*, *19*(4), 136–150. https://doi.org/10.35784/acs-2023-40

Leo, M., Carcagnì, P., Mazzeo, P. L., Spagnolo, P., Cazzato, D., & Distante, C. (2020). Analysis of facial information for healthcare applications: A survey on computer vision-based approaches. *Information*, *11*(3), Article 128. https://doi.org/10.3390/info11030128

Machrowska, A., Karpiński, R., Maciejewski, M., Jonak, J., & Krakowski, P. (2024). Application of EEMD-DFA algorithms and ANN classification for detection of knee osteoarthritis using vibroarthrography. *Applied Computer Science*, *20*(2), 90–108. https://doi.org/10.35784/acs-2024-18

Na, H. C., & Kim, Y. S. (2024). Study on deep learning models for VR sickness levels classification. *Applied Computer Science*, *20*(4), 1–13. https://doi.org/10.35784/acs-2024-37

Nour, N., Elhebir, M., & Viriri, S. (2020). Face expression recognition using convolution neural network (CNN) models. *International Journal of Grid Computing and Applications*, *11*(4), 1–11. https://doi.org/10.5121/ijgca.2020.11401

Papers With Code. (n.d.). *UNBC-McMaster Shoulder Pain Expression Database* [Data set]. Retrieved March 23, 2026, from https://paperswithcode.com/dataset/unbc-mcmaster

Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., Keefe, F. J., Mogil, J. S., Ringkamp, M., Sluka, K. A., Song, X. J., Stevens, B., Sullivan, M. D., Tutelman, P. R., Ushida, T., & Vader, K. (2020). The revised International Association for the Study of Pain definition of pain: Concepts, challenges, and compromises. *Pain*, *161*(9), 1976–1982. https://doi.org/10.1097/j.pain.0000000000001939

Saddam, E., Mutashar, S., & Ali, W. (2021). A study of patient’s pain assessment based on facial expression: Issues and challenges. *Engineering and Technology Journal*, *39*(10), 1514–1527. https://doi.org/10.30684/etj.v39i10.2079

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *ArXiv, abs/1503.03832*. https://doi.org/10.48550/arXiv.1503.03832

Shier, W. A. (2017). *Automated pain recognition using analysis of facial expressions* [Master's thesis, University of Calgary]. PRISM. https://doi.org/10.11575/PRISM/25076