

Development and internal validation of single-image pain detection from automated facial coding

Corina Espelien

cme2kd@virginia.edu

University of Virginia

Matthew Kongkatong

University of Virginia

Matthew Gurka

University of Virginia

Pavel Chernyavskiy

University of Virginia

Article

Keywords:

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10539662/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Accurate pain assessment remains challenging clinically, especially when patient self-report is unavailable or unreliable. Facial expressions offer a source of information about pain, but prior approaches typically rely on large volumes of data and single-center datasets that may limit clinical applicability. We evaluated whether pain can be detected from a single facial image using automatically encoded facial Action Units (AUs). We retrospectively pooled and harmonized three experimental pain databases, yielding one neutral and one pain image for 212 participants exposed to various stimuli. Model performance was assessed using participant-clustered cross-validation. Across models, pain discrimination was fair-to-good (AUC up to 0.77). Asymmetric AU activation, especially in cheek, brow, and eye regions, emerged as a consistent predictor of pain. These findings demonstrate that clinically meaningful pain-related information can be extracted from single images and support the feasibility of facial expression analysis as a pain-detection clinical decision support tool.

Introduction

Effective pain management is a foundational goal across clinical medicine, but persistent challenges in pain assessment remain. Pain is inherently subjective and highly variable, shaped by its underlying causes as well as individual, social, and contextual factors¹. Clinical practice relies almost exclusively on patient self-reports, most commonly through the Numeric Pain Rating Scale (NPRS) or its visual analog². These self-reports, though informative, are an imperfect measure shaped by a myriad of individual, social, cultural, and systemic factors. In fact, self-reported pain is more likely to be underestimated in women, racial and ethnic minorities, and individuals with chronic pain^{3–5}. Additionally, prior studies illustrate a pain-rating mismatch between clinicians and patients⁶, showing that these biases can contribute to disparities in diagnosis, treatment, and outcomes⁷. The limitations of self-reported pain are further compounded in patients who cannot reliably communicate their pain, including those with cognitive or developmental impairments, children, and individuals who are sedated or intubated⁸. These limitations underscore the need for objective approaches to complement and improve pain assessment for clinicians and patients alike.

Facial expression analysis has long been applied to describe the prevalence and intensity of universal discrete emotions (e.g., happiness, sadness, surprise), with a growing list of applications ranging from mental health assessment to neurological and behavioral disorders^{9–12}. These systems build on the theory that facial expressions are the motor component of affect¹¹ and so the core affective states are expressed through structured, reproducible patterns of facial muscle activity⁹. This approach is formalized in the Facial Action Coding System (FACS), which classifies facial movements patterns via “Action Units” (AUs)¹³. Whereas manual coding is too time-intensive for many applications, advances in facial recognition have enabled automated AU detection, allowing for facial analysis to be used at scale^{14,15}.

Similar to prototypical facial expressions for universal emotions, pain can be signaled by facial expressions characterized by patterns of AU activation^{16–18}. For example, the Prkachin and Solomon Pain Intensity (PSPI) index is a linear combination AUs and remains a widely-used metric in FACS-based pain research. However, most prior work has largely relied on single-database, single-center, and/or single-rater studies^{19,20}, limiting generalizability across populations and settings, including uncertainty in real-world scenarios. Furthermore, many prior predictive models are trained on multiple (usually a high volume) of frames or images per subject, potentially inflating performance estimates while limiting applicability in a single-encounter clinical setting^{21,22}. Recent deep learning approaches have reported performances exceeding 90% accuracy under such settings²², but rely on richer data than what may typically be available in routine clinical workflows. These limitations remain a gap between proof-of-concept algorithms and a deployable clinical decision support tool for evaluating pain.

To address these gaps, the objective of this study is to investigate the feasibility of FACS-based pain detection from a single image, which is easier to obtain than a sequence of images and is thus easier to integrate into clinical practice. We use images from a combination of three databases of pain experiments to avoid overfitting our model to a single cohort or experimental setting. This article lays the foundation for pain detection clinical support tool with realistic constraints, which can complement patient self-report to inform treatment and management practices.

Results

The performance of the models trained on images of participants' neutral and pained faces was evaluated using AUC, sensitivity, specificity, accuracy, F1, and Matthew's Correlation Coefficient (MCC). In addition, we compared models trained on all AU activation vs. a Generalized Additive Model (GAM) with the PSPI index as the sole predictor.

Overall, all models we tested demonstrated fair-to-good out-of-fold predictive performance, with MCC ranging 0.421 – 0.549 and AUC ranging 0.695 – 0.770 (**Table 1**). The Multi-Layer Perceptron (MLP) model emerged as the top overall performer (AUC=0.770 ± 0.074; MCC=0.549 ± 0.130), achieving the highest scores across all metrics except for sensitivity. Given the relatively small number of observations (N=424), the MLP also exhibited the highest cross-fold variability across most of the metrics. The Elastic Net (ENET) was the best likelihood-based model, third-best overall (AUC=0.742 ± 0.047; MCC=0.469 ± 0.087), achieving the highest sensitivity (0.822 ± 0.143) among all models tested. The Random Forest (RF) was the second-best performing model overall, with sensitivity nearly equivalent to that of ENET and slightly higher specificity and accuracy.

The GAM with only the non-linear PSPI index smooth term (full dataset EDF=3.325; $p<2e-16$), out-of-fold EDF=3.255 ± 0.127, yielded the lowest MCC (0.421 ± 0.094) and AUC (0.695 ± 0.062). The poor performance of the GAM model offers evidence that using all available AUs as inputs carries substantially more information about pain than the PSPI alone. However, the Generalized Linear Model (GLM; logistic regression) using all 31 AU predictors did not yield a high-performing model. In contrast,

the ENET kept only 13 AU predictors on average across folds. This suggests that regularization (i.e., exclusion of irrelevant AUs) is vital to achieving high-performance out-of-fold pain detection.

Pairwise probability comparisons showed heterogeneity in how models predicted pain states (**Appendix Figure A1**). ENET and GLM had highly similar prediction structures ($R^2 = 0.93$), while ENET had moderate agreement with SVM and MLP ($R^2 = 0.73$ and 0.78 , respectively). RF exhibited weak agreement with all other models (R^2 between 0.22 and 0.36), suggesting a substantially different prediction structure. The top-performing MLP had only weak-to-moderate agreement with other models (R^2 ranging from 0.36 to 0.73).

Table 1. Model performance metrics (mean $\pm$ standard deviation) across 10-fold cross validation using Youden’s J cutoff for prediction. For each metric, the best performer is highlighted. The GAM includes only the PSPI index as its sole non-linear predictor.

Model	AUC	Sensitivity	Specificity	Accuracy	F1	MCC
GAM	0.695 $\pm$ 0.062	0.816 $\pm$ 0.114	0.585 $\pm$ 0.097	0.700 $\pm$ 0.044	0.729 $\pm$ 0.053	0.421 $\pm$ 0.094
GLM	0.719 $\pm$ 0.057	0.781 $\pm$ 0.177	0.616 $\pm$ 0.150	0.698 $\pm$ 0.046	0.713 $\pm$ 0.073	0.424 $\pm$ 0.103
SVM	0.748 $\pm$ 0.050	0.778 $\pm$ 0.130	0.666 $\pm$ 0.143	0.722 $\pm$ 0.031	0.733 $\pm$ 0.044	0.462 $\pm$ 0.066
ENET	0.742 $\pm$ 0.047	0.822 $\pm$ 0.143	0.617 $\pm$ 0.141	0.719 $\pm$ 0.038	0.742 $\pm$ 0.049	0.469 $\pm$ 0.087
RF	0.747 $\pm$ 0.064	0.821 $\pm$ 0.128	0.637 $\pm$ 0.158	0.729 $\pm$ 0.041	0.749 $\pm$ 0.044	0.483 $\pm$ 0.068
MLP	0.770 $\pm$ 0.074	0.794 $\pm$ 0.111	0.745 $\pm$ 0.121	0.770 $\pm$ 0.064	0.774 $\pm$ 0.064	0.549 $\pm$ 0.130

Model calibration plots (**Appendix Figure B1**) evaluated the agreement between predicted probabilities and observed pain status. We focused on the ENET, the top likelihood-based model, and the MLP, the top ML model. Calibration intercepts were close to zero for both models (ENET: 0.00 95% CI $[-0.20$ to $0.21]$; MLP: 0.04 95% CI $[-0.18$ to $0.25]$), indicating minimal systematic over- or under-prediction. However, the ENET model exhibited miscalibration in the middle of the probability distribution. In contrast, the MLP model demonstrated near-ideal calibration across the entire probability distribution, indicating that predicted probabilities closely corresponded to observed outcomes.

Calibration was further evaluated for males and females separately to assess potential generalizability across different demographic groups (**Appendix Figure B2**). Stratified by participant sex, which was not an input in the model, the MLP still exhibits satisfactory calibration for both sex groups. The calibration for the MLP model was slightly better for female ($n = 192$; C-statistic = 0.80) than male faces ($n = 232$; C-statistic = 0.76). In contrast, the ENET calibration was worse for male faces (C-statistic = 0.74) compared to the overall and female calibration curves.

Asymmetric AU activation was an important predictor of pain status in both the top-performing ML and likelihood-based models. In the MLP model, there were three AU asymmetries in the top 10 most important predictors, asymmetry in: AU6 (asymmetric cheek raise), AU4 (asymmetric brow lowering), and AU2 (asymmetric outer brow raiser), each with greater activation on the right side vs. left side indicating pain status (**Figure 1**). These findings were mirrored in the ENET model, where asymmetric brow lowering (AU4) and eye closure (AU43) were both indicators of pain, retained in all 10 cross-validation folds. These findings suggest that asymmetry captures information that is not fully represented by overall AU activation levels and may provide complementary information for distinguishing pain from non-pain expressions.

Discussion

This study provides evidence that FACS-derived features from a single image can be used to distinguish pain from neutral facial expressions across multiple experimental pain stimuli, supporting the feasibility of facial-expression analysis as a clinical decision support tool for pain assessment. Motivated by the gap between proof-of-concept pain-detection algorithms and clinically deployable systems, we designed our analysis to better reflect the practical constraints that may arise in healthcare settings, including reliance on a single image per encounter and heterogeneity in both participants and pain-induction methods. To do this, we build on prior literature by: 1) working with a larger and more diverse set of participants and painful stimuli (pain induced in three ways, male and female participants aged 17–65 years); 2) using only one image per pain status and participant; 3) documenting the complete analytic pathway from image to prediction; and 4) highlighting the novel predictive potential of facial asymmetry around the eyebrows, eyes, and cheeks.

During our analysis, we found that facial asymmetry (reflected via asymmetric AU activation) is predictive of pain status in an experimental setting. Asymmetric facial expressions have not been previously assessed in the study of pained expressions, though it has been explored in general emotional analysis²³. Interestingly, some of the top asymmetric variables (Fig. 1) highlight the asymmetry of AUs that were originally used in the formulation of PSPI (asymmetry in AUs 4, 6, 7, and 43), which capture brow lowering, cheek raising, and eyes closing. Future work should investigate whether these asymmetry patterns are present across clinical populations, alternative pain stimuli or presentations, and diverse cultural contexts. This highlights a new avenue for understanding and quantifying pain-related facial behavior both mechanistically and for the development of automated pain-detection systems.

Although we still work with experimental data, the pain stimuli in these experiments were designed to induce relatively low-level pain, which is likely lower than an acute clinical presentation. The AU activation associated with severe pain is likely larger and potentially easier to detect than the data we examined. Investigating predictive performance when applied to moderate-to-severe clinical pain remains an important question for future validation studies. The evidence presented in this article has strong potential to inform *in-situ* clinical data collection and analysis, in support of the long term goal of

deploying an automatic pain detection system as a clinical decision support tool in dynamic clinical settings, like an Emergency Department or Perioperative Intensive Care Unit.

In this article, we pooled and harmonized data from three databases: BioVid, D-PASS, and MIntPain. Through the standardization of the inputs from three separate databases, our work mitigates overfitting to a singular pain application method, demographic group, or study setup. Expanding the experimental analysis to include more subjects covering wide range of age, race/ethnicity, and gender would be beneficial. Importantly, only a fraction of the available facial image data was used during model training (one pain image and one neutral image per participant). The motivation for restricting the data in this way was two-fold: 1) to standardize the contribution from each subject from different studies, and 2) to reflect potential constraints of a clinical setting, where a single snapshot or intake photo may be the extent of available data. Additionally, the prediction task in this study was a binary detection of pain vs. no pain (neutral facial expression, unique to each participant). For two of the studies, BioVid and MIntPain, this condensed pain trials from various pain intensities (levels 1 through 4), into a single “pain” classification. This was needed to harmonize with the DPASS data, which was recorded only as binary, and increased the difficulty of the pain prediction task by deliberately inducing heterogeneity in the facial expressions of the pain condition, similar to what may be observed in a clinical setting. Future analyses with labeled stimulus or pain intensity data (including but not limited to BioVid and MIntPain) could be used for quantitative or ordinal pain intensity rather than presence/detection.

How a FACS-based pain detection tool is deployed in a clinical setting will influence model selection, evaluation metric prioritization, and interpretation of results. In practice, predictions may be triggered at discrete clinical time points (e.g., triage intake photographs, postoperative assessments) or continuously via bedside video monitoring in higher-acuity settings such as intensive care units. Threshold selection should reflect the intended clinical role: lower thresholds prioritizing sensitivity may be appropriate for screening applications to minimize missed pain, whereas higher thresholds prioritizing specificity may be more appropriate for triggering interventions such as analgesic administration or escalation of care. These outputs would likely be integrated into clinical decision support systems, with frontline use by nursing staff for monitoring and escalation to physicians when risk thresholds are exceeded or sustained elevations are observed. In this context, the model functions as an assistive tool that augments rather than replaces or obfuscates standard pain assessment practices.

Overfitting was mitigated through cross-validation, regularization in the ENET model, and restriction of model complexity via tuned hyperparameters for ML models. The MLP model’s discrimination (AUC = 0.770, MCC = 0.549), overall and sex-specific calibration, and balanced sensitivity and specificity indicate that it can generate reliable, general purpose estimates for the probability of pain even with this relatively small dataset. In contrast, the ENET model’s high sensitivity could be advantageous for creating a pain screening tool, where providers want to avoid missing clinically significant pain events. Pairwise probability comparisons (**Appendix A**) further suggest that, the top-performing MLP displayed only moderate agreement with competing models despite achieving superior predictive performance. This suggests that different algorithms may be using complementary information. Future analyses can

explore ensemble machine learning to best utilize the strength of each analytic approach and reach better predictive performance than any single model alone. Regardless of predictive performance, model explainability will remain central in clinical adoption, as model transparency and interpretability are key to building trust and buy-in from various clinical groups and across different patient populations.

Beyond performance metrics, calibration analyses provided insight into the potential clinical utility of the models (**Appendix B**). While discrimination metrics quantify how well a model distinguishes between individuals with and without pain, calibration assesses the agreement between predicted probabilities and observed outcomes. In many clinical applications, the quality of predicted probabilities is as important as classification accuracy because a model that systematically over- or underestimates risk may contribute to inappropriate clinical decisions when probability thresholds are used to guide intervention. The MLP demonstrated near-ideal calibration across the probability spectrum, suggesting that its predicted probabilities more accurately reflected observed pain status. This finding strengthens confidence in the model's potential utility in automated pain-assessment settings, where clinicians may rely on prediction magnitudes rather than binary classifications alone. Calibration may be particularly important when extending facial-expression models beyond pain detection to the prediction of pain severity levels (e.g., mild, moderate, severe, or NPRS), where reliable probability estimates are needed to support treatment and monitoring.

Although sex was not included in model development, the MLP maintained satisfactory calibration across male and female participants, whereas calibration degraded for the ENET model among male participants. In a clinical context, ENET's miscalibration at mid-range probabilities and reduced reliability among male participants could translate into less accurate estimation of intermediate pain levels and less robust decision-making. Future studies should routinely assess calibration across demographic and clinical subgroups before deployment to facilitate fairness, robustness, and clinical reliability in algorithmic decision support tools. Collectively, these findings underscore that well-calibrated models are essential not only for predicting the presence of pain but also for supporting clinically meaningful estimates of pain intensity.

Similar to previous research on FACS-based pain detection, this work uses algorithmic facial analysis, introducing potential concerns regarding fairness and bias across populations. Prior literature has demonstrated that facial analysis systems can exhibit systematic performance differences across demographic groups due to imbalances in training data and variation in facial feature detection²⁴. Biases in pain detection may have direct clinical consequences; for example, models that underperform in certain populations could contribute to the under-recognition or undertreatment of pain in already vulnerable groups²⁵. While we evaluated sex-stratified calibration, model performance stratified by other key factors (e.g., race/ethnicity, skin tone, age) fell outside the scope of the current paper. These factors could influence AU detection, as well as expressions of pain based on lived experience. Our analysis was limited by the interpolation of sex and age for one of the datasets, MIntPain, but included one cohort of intentionally diverse participants (DPASS)²⁶. Remediation of potential bias in facial analysis, especially for individuals with darker skin tones, is a dynamic area of research in affective computing^{27–29}. A

worthwhile future direction for our work may be implementing the newly introduced image augmentations (e.g., skin reflectance correction²⁸) and examine if the coded AU activation is substantively different.

Although pooling multiple datasets represents a step toward improving diversity and generalizability, it does not guarantee equitable performance. Data collected both lab and clinical settings should document the robust demographic information for FACS-based pain detection to be included in future modeling efforts. Future work should include intentionally diverse and representative datasets and subgroup assessments to ensure that predictive performance is both accurate and equitable across populations. Transparent reporting of model performance stratified by demographic characteristics will be essential for assessing clinical readiness.

All models were trained and tested on data collected in controlled experimental settings, which impose important limitations and constraints. These conditions include instructions to face the camera and minimize talking. Subjects were aware of on-going data collection, which may influence the expression or masking of pain via social desirability bias³⁰, although we posit this makes for a more difficult pain prediction task. Additionally, the allowable pain exposure for volunteer testing keeps the pain levels within the lower end of the possible pain spectrum. Pain detection in real-world clinical settings introduces variable lighting, camera setup and quality, face obfuscations (e.g., masks, devices, bandages), patient movement, and patient state (e.g., sedated, altered consciousness). Furthermore, clinical data may include unplanned and potentially severe pain exposures found in settings like the emergency department, which go beyond the training datasets in laboratory settings. Emerging work has demonstrated that ML models using facial expressions in postoperative patients can predict severe pain (NPRS ≥ 7) and outperform models based on physiological signals, underscoring the potential of affect-detected pain in clinical settings and the need for larger-scale validation across more variable and realistic conditions³¹.

Future work on the development and validation of a pain-detection clinical support tool should use prospective and external data. One promising approach for prospective data collection would be to incorporate facial-expression analysis into pain intervention studies, where facial images, self-reported pain, clinician assessments, and treatment responses are collected longitudinally. Such designs would permit evaluation of whether model-generated pain estimates are responsive to changes in pain over time, detect treatment effects, and provide clinically meaningful information. In parallel, external validation across independent datasets and healthcare systems will be necessary to determine whether model performance changes across different patient populations, clinical environments, imaging hardware, and institutional workflows. Demonstrating robust performance across care settings will be a critical milestone in transitioning facial-expression pain detection from proof-of-concept to a clinically deployable and actionable decision support tool.

In addition to real-world validation, clinician trust, acceptance, and concerns surrounding the use of automated pain assessment need to be evaluated. Even highly accurate models are unlikely to be

adopted if they are viewed as opaque, unreliable, or inconsistent with clinical judgment. Future implementation studies should assess how different healthcare professionals would interpret and use model predictions in practice. Perspectives may vary by discipline and clinical setting. For example, emergency medicine clinicians may prioritize rapid screening and sensitivity to avoid missing patients experiencing acute pain, whereas palliative care providers may place greater emphasis on longitudinal monitoring, individualized assessment, and alignment with patient-reported experiences. Similar differences may exist among nursing staff, anesthesiologists, intensivists, and rehabilitation providers. Evaluating these stakeholder perspectives will be essential for determining appropriate deployment strategies, establishing relevant probability thresholds, and ensuring that facial-expression models are developed in a manner that supports clinician trust, workflow integration, and patient-centered care.

Automated facial-expression models are promising clinical decision support tools and not intended as replacements for patient self-report, which remains the standard for pain assessment. Although such systems could provide complementary information during routine encounters with communicative adults, their greatest potential may be in settings where self-report is limited, unavailable, or difficult to obtain consistently. Examples include patients with cognitive impairment, altered mental status, language barriers, communication disorders, or those recovering from surgery or sedation. Example deployment scenarios include automatically generated pain scores presented to nurses during vital-sign assessments, incorporated into bedside monitoring systems, or displayed to physicians alongside other clinical information. Probability thresholds could then be linked to management actions, such as prompting repeat pain assessment if deterioration or under-reporting is detected. By providing an objective and complementary source of information, facial-expression pain models could augment existing assessment strategies and help identify patients who may benefit from additional clinical attention.

Materials and Methods

This retrospective analysis used previously collected facial expression data to predict the presence of pain using a two-step method. First, we encoded AU activation from images using Noldus FaceReader10³², a validated automatic facial coding software. For each image, FaceReader recognizes subject faces and codes the observed AU activation into a numeric intensity between 0 and 1. Second, we trained and validated through 10-fold cross-validation a collection of likelihood-based and machine learning (ML) models for pain classification. Model development and reporting were informed by TRIPOD guidelines for the development of prediction models.

Image and Video Databases

Three lab-based experimental studies on facial expressions of pain in healthy adults were used: BioVid Heat Pain database (BioVid), Denver Pain Authenticity Stimulus Set (D-PASS), and Multimodal Intensity Pain database (MIntPain). Images from each of these studies were made available via a user agreement

of the authors with the performing institution. Because the data were collected independently by the original investigators, exact participant recruitment and data collection dates were not consistently available. Participants were healthy adult volunteers meeting inclusion criteria of the original studies.

An overview of the information available for each study is in Table 2. Sample size was determined by the number of participants available across the three databases; no formal sample size calculation was performed.

The BioVid database exposed volunteers to individualized heat pain levels, ranging from pain perception (1) to pain tolerance (4) ^{33,34}. Each pain level was simulated 20 times in a randomized order with variable-length no-stimulus intermissions. Other data available in this database, but not used in this analysis, includes posed and elicited emotion expression and biopotential measures. D-PASS collected a neutral facial expression image, two faked or posed pain expression videos, and an authentic or real pain video ²⁶. The faked expressions occurred with no stimulus applied, while the real expressions occurred with pressure stimulus applied. For this analysis, the faked pain videos were excluded. The pressure stimulus was applied until the subject indicated they were at their tolerance, or a predetermined safety threshold was reached. In the MIntPain database, each subject underwent 40 trials, each of which captured states without pain (level 0) and with pain stimulus (level 1, 2, 3, or 4) ³⁵. Like BioVid, the MIntPain database used an initial calibration for each subject to determine the prescribed stimulus levels with the experimental goal of inducing pain. In all studies, subjects were seated facing the camera with sufficient lighting. In D-PASS and BioVid, the protocol explicitly stated that subjects were asked to minimize talking.

Table 2
Demographics of participants in three databases used in the current article (N = 212 participants; 424 images).

Database	BioVid	D-PASS	MIntPain ^b
N	87	105	20
N _f	43	53	0
N _m	44	52	20
Age range	20–65	17–28	22–42
Number of pain trials per subject	80	1 ^a	40
Pain stimulus	Heat	Pressure	Electric

^a Only “real” (stimulus induced) trials used in analysis.

^b Age and sex were not available on a per-subject basis in the distributed database. Published age ranges have been published previously³⁵ and age and sex were obtained on a per-subject basis using

Pain Coding and Action Unit Data Preprocessing

Image and video preprocessing was used to standardize the data from different studies so that they could be pooled and sampled for the model framework. The goal of preprocessing was to generate numerical action unit (AU) data for every image or frame for each study, subject, and test condition. Subsequently, these data were sampled prior to training a set of statistical models and assessing model performance through cross-validation.

Each image or video from BioVid, D-PASS, and MIntPain was processed in FaceReader10, transforming each 2-D image or video into a row vector of 31 AU numerical values, comprised of 20 AU activation intensities and 11 AU activation asymmetries, computed as left intensity – right intensity. An exemplar image from the BioVid database is in Fig. 2, with FaceReader-coded AU activation labels. The encoding settings were consistent with best practices as outlined in the software manual and the authors' prior work³⁷. FaceReader encoding was automated and performed without reference to outcome labels.

For every row of coded AU data, there were matching pain measures; here, we focused on binary pain classification: pain absent, neutral facial expression (pain = 0) or pain present, facial expression at stimulus application (pain = 1). For BioVid and MIntPain, which had non-stimulus and four pain stimulus trials (up to tolerance), the non-pain stimulus trials were recorded as pain = 0 and the pain stimulus trials (levels 1 through 4) were recorded as pain = 1. For D-PASS, the neutral image was coded as pain = 0 and the real or authentic pain was coded as pain = 1. Outcome labels were determined and assigned by the original experimental protocols and assigned independently of the predictor facial AU values. Pain-status labels were available for all participants, and complete AU data were obtained for all selected images. No observations were excluded because of missing predictor or outcome data.

For each of the 212 total subjects, one pain = 0 and one pain = 1 image was sampled using seed number 80160, resulting in 424 observations with 0.50 pain prevalence. For databases with more than one trial or test condition (BioVid and MIntPain), only one frame from a randomly selected neutral (pain = 0) trial and one frame from a randomly selected pain-inducing (pain = 1) trial were selected.

The encoded AU intensities were transformed from their raw output prior to serving as inputs into the models. Each intensity is continuous and theoretically bounded on the closed interval [0, 1]. Untransformed intensity distributions were all right-skewed, with almost all distributions concentrated at or near 0 (**Appendix Figure C1**). To maximize our chances of detecting differences in AU activation patterns between pain conditions we employed a deterministic monotonic transformation. Let x_{ij} be the raw encoded intensity for the j th AU on the i th participant. We define the transformation function $g : [0, 1] \rightarrow \mathbb{R}$ as follows:

$$g(x_{ij}) = \text{logit} \left[\frac{1}{N} (x_{ij} (N - 1) + 0.5) \right] = \ln \left[\frac{\frac{1}{N} (x_{ij} (N - 1) + 0.5)}{1 - \frac{1}{N} (x_{ij} (N - 1) + 0.5)} \right]$$

The function g is strictly monotonic in x_{ij} for a given N and can be recognized as the logit transformation meant to “nudge” variables on the unit interval away from 0 and 1³⁸. All statistical models used $g(x_{ij})$ as inputs. To serve as input in the comparison model, the PSPI was computed using the raw AU activation values as follows:

$$PSPI_i = AU4_i + \max(AU6_i, AU7_i) + \max(AU9_i, AU10_i) + AU43_i.$$

Model Development

We tested multiple likelihood-based and machine learning (ML) approaches to predict pain status (pain = 1). Likelihood-based parametric models (e.g., logistic regression) tend to generate more interpretable estimates and be more familiar to clinicians, thus presenting fewer barriers to widespread adoption. ML algorithms (e.g., Random Forest) tend to perform better on complex prediction tasks, in part because the functional form of how the inputs correlate with the response need not be linear or even be specified prior to model estimation.

We evaluated three likelihood-based models (Generalized Linear Model/logistic regression (GLM), Elastic Net regression (ENET), and Generalized Additive Model (GAM)) and three ML models (Random Forest (RF), Support Vector Machine (SVM), and the Multilayer Perceptron (MLP)). For the unilateral AUs where both left and right activation were available, the mean activation was taken as the unilateral AU. Except for the GAM, regression models and ML algorithms included the same 31 input variables: the transformed activation of 20 unilateral AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, AU27, AU43) and 11 AU asymmetries (AU1asym, AU2asym, AU4asym, AU5asym, AU6asym, AU7asym, AU12asym, AU14asym, AU15asym, AU20asym, AU43asym), computed as left activation – right activation. No other participant or study variables were included in the models.

Model Evaluation

For each model, predictive performance was evaluated using participant-clustered 10-fold cross-validation, wherein the folds were constructed such that the neutral (pain = 0) image and the pained (pain = 1) image from each participant were allocated to the same fold. With a sample size of $N = 424$, 10-fold cross-validation provides a balance between bias and variance in performance estimation, while maintaining sufficient sample size within each fold for stable model training³⁹. To prevent data leakage across folds, we used participant-level clustering (i.e., Participant 1 neutral and pain image allocated to same fold) and applied preprocessing steps, including AU transformation and scaling, within each fold.

We computed mean and standard deviation performance metrics across validation folds, with the standard Youden’s J used to compute the classification threshold. To evaluate model performance, we used out-of-fold Matthew’s Correlation Coefficient (MCC; ⁴⁰) and the Area Under the Operating Characteristic curve (AUC) as the primary criteria. In addition to these metrics, we also computed out-of-fold sensitivity, specificity, accuracy, and F-1 score. Calibration plots, both overall and stratified by participant sex, were used to evaluate how closely predictions matched observed pain status. Details on the model tuning parameters and model implementation in R are summarized in **Appendix Table D1**.

Declarations

ACKNOWLEDGEMENTS

The authors acknowledge the creators and maintainers of the publicly available pain expression databases used in this work. We thank the following researchers for developing and curating their databases:

- BioVid Heat Pain database: Dr. Sascha Gruss and colleagues at Ulm University;
- Denver Pain Authenticity Stimulus Set: Dr. E. Paige Lloyd and colleagues at University of Denver;
- Multimodal Intensity Pain database: Dr. Andreas Aakerberg and colleagues at Aalborg University.

FUNDING

This work did not receive specific funding for the conduct of the study.

AUTHOR CONTRIBUTIONS

The author contributions are described below with the CRediT model.

Conceptualization: CE, PC

Data curation: CE, PC

Formal analysis: CE, PC

Funding acquisition: PC

Investigation: CE, PC

Methodology: CE, PC

Project administration: CE, PC

Resources: CE, PC

Software: CE, PC

Supervision: CE, PC, MG

Validation: CE, PC

Visualization: CE, PC

Writing—original draft: CE, PC

Writing—review & editing: CE, PC, MK, MG

COMPETING INTERESTS

All other authors declare they have no competing interests.

DATA AVAILABILITY

The datasets used in this study are publicly available but subject to access and data use agreements. Upon request, the BioVid Heat Pain Database is available from Ulm University; the Denver Pain Authenticity Stimulus Set is available through the University of Denver; and the Multimodal Intensity Pain database is available from Aalborg University.

CODE AVAILABILITY

The code used for data preprocessing and model development is available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study used secondary data from previously published experimental databases involving human participants (BioVid, D-PASS, and MIntPain), each of which received prior ethical approval from their respective institutional review boards. No new data involving human participants were collected for this study. Therefore, additional ethical approval was not required.

References

1. Coghill, R. C. Individual Differences in the Subjective Experience of Pain: New Insights Into Mechanisms and Models. *Headache: The Journal of Head and Face Pain* 50, 1531–1535 (2010).
2. Williamson, A. & Hoggart, B. Pain: a review of three commonly used pain rating scales. *Journal of Clinical Nursing* 14, 798–804 (2005).
3. Zhang, M., Zhang, Y., Li, Z., Hu, L. & Kong, Y. Sexism-Related Stigma Affects Pain Perception. *Neural Plasticity* 2021, 6612456 (2021).
4. Meints, S. M., Cortes, A., Morais, C. A. & Edwards, R. R. Racial and ethnic differences in the experience and treatment of noncancer pain. *Pain Manag* 9, 317–334.

5. Perugino, F., De Angelis, V., Pompili, M. & Martelletti, P. Stigma and Chronic Pain. *Pain Ther* 11, 1085–1094 (2022).
6. Seers, T., Derry, S., Seers, K. & Moore, R. A. Professionals underestimate patients' pain: a comprehensive review. *PAIN* 159, 811 (2018).
7. Emerson, A. J. et al. Patient and Provider Attitudes, Beliefs, and Biases That Contribute to a Marginalized Process of Care and Outcomes in Chronic Musculoskeletal Pain: A Systematic Review –Part I: Clinical Care. *Pain Med* 23, 655–668 (2022).
8. Herr, K. et al. Pain Assessment in the Patient Unable to Self- Report: Clinical Practice Recommendations in Support of the ASPMN 2024 Position Statement. *Pain Management Nursing* 25, 551–568 (2024).
9. Ekman, P. & Friesen, W. V. Facial Action Coding System. <https://doi.org/10.1037/t27734-000> (1978) doi:10.1037/t27734-000.
10. Huang, Y. et al. Insomnia and impacts on facial expression recognition accuracy, intensity and speed: A meta-analysis. *Journal of Psychiatric Research* 160, 248–257 (2023).
11. Ekman, P. Facial expression and emotion. *American Psychologist* 48, 384–392 (1993).
12. Flynn, M. et al. Assessing the Effectiveness of Automated Emotion Recognition in Adults and Children for Clinical Investigation. *Front. Hum. Neurosci.* 14, 70 (2020).
13. Ekman, P., Friesen, W. V. & Hager, J. C. Facial Action Coding System. Manual and Investigator's Guide. (2002).
14. Zeng, Z., Pantic, M., Roisman, G. I. & Huang, T. S. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 39–58 (2009).
15. Cohn, J. & De la Torre, F. Automated face analysis for affective computing. *Handbook of affective computing* 131–150 (2014).
16. Prkachin, K. M. The consistency of facial expressions of pain: a comparison across modalities. *Pain* 51, 297–306 (1992).
17. Kunz, M., Scharmann, S., Hemmeter, U., Schepelmann, K. & Lautenbacher, S. The facial expression of pain in patients with dementia. *PAIN®* 133, 221–228 (2007).
18. Lautenbacher, S. et al. Automatic Coding of Facial Expressions of Pain: Are We There Yet? *Pain Res Manag* 2022, 6635496 (2022).
19. Torralba, A. & Efros, A. A. Unbiased look at dataset bias. in *CVPR 2011* 1521–1528 (2011). doi:10.1109/CVPR.2011.5995347.
20. Taati, B. et al. SynPAIN: A Synthetic Dataset of Pain and Non-Pain Facial Expressions. Preprint at <https://doi.org/10.48550/arXiv.2507.19673> (2025).
21. Werner, P. et al. Automatic Recognition Methods Supporting Pain Assessment: A Survey. *IEEE Transactions on Affective Computing* 13, 530–552 (2022).

22. Cascella, M., Esposito, D., Muzio, M. R., Cascella, V. & Cerrone, V. Artificial intelligence for pain assessment via facial expression recognition (2015–2025): a systematic review. *Explor Med.* 6, 1001370 (2025).
23. Murray, E. M. et al. Asymmetry of Facial Expressions of Emotion. in *Understanding Facial Expressions in Communication: Cross-cultural and Multidisciplinary Perspectives* (eds Mandal, M. K. & Awasthi, A.) 73–99 (Springer India, New Delhi, 2015). doi:10.1007/978-81-322-1934-7_5.
24. Grother, P., Ngan, M. & Hanaoka, K. *Face Recognition Vendor Test Part 3: Demographic Effects*. NIST IR 8280 <https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf> (2019) doi:10.6028/NIST.IR.8280.
25. Hoffman, K. M., Trawalter, S., Axt, J. R. & Oliver, M. N. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences* 113, 4296–4301 (2016).
26. Lloyd, E. P. et al. Denver pain authenticity stimulus set (D-PASS). *Behav Res* 56, 2992–3008 (2024).
27. Kara, O., Churamani, N. & Gunes, H. Towards Fair Affective Robotics: Continual Learning for Mitigating Bias in Facial Expression and Action Unit Recognition. Preprint at <https://doi.org/10.48550/arXiv.2103.09233> (2021).
28. Krishnapriya, K. S., Pangelinan, G., King, M. C. & Bowyer, K. W. Analysis of Manual and Automated Skin Tone Assignments. in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)* 429–438 (IEEE, Waikoloa, HI, USA, 2022). doi:10.1109/WACVW54805.2022.00049.
29. Zeng, L. & Kalantari, N. K. Analyzing and Improving the Skin Tone Consistency and Bias in Implicit 3D Relightable Face Generators. Preprint at <https://doi.org/10.48550/arXiv.2411.12002> (2024).
30. Oceno, M. How social desirability bias impacts the expression of emotions. *Political Science Research and Methods* 1–10 (2025) doi:10.1017/psrm.2025.10016.
31. Park, I. et al. Machine learning model of facial expression outperforms models using analgesia nociception index and vital signs to predict postoperative pain intensity: a pilot study. *Korean J Anesthesiol* 77, 195–204 (2024).
32. Noldus Information Technology. Noldus FaceReader 10: Tool for automatic analysis of facial expressions. (2025).
33. Werner, P. et al. Towards Pain Monitoring: Facial Expression, Head Pose, a new Database, an Automatic System and Remaining Challenges. in *Procedings of the British Machine Vision Conference 2013* 118.1-118.11 (British Machine Vision Association, Bristol, 2013). doi:10.5244/C.27.119.
34. Walter, S. et al. *The BioVid Heat Pain Database: Data for the Advancement and Systematic Validation of an Automated Pain Recognition System*. 2013 *IEEE International Conference on Cybernetics, CYBCONF 2013* 131 (2013). doi:10.1109/CYBConf.2013.6617456.
35. Haque, M. A. et al. Deep Multimodal Pain Recognition: A Database and Comparison of Spatio-Temporal Visual Modalities. in *2018 13th IEEE International Conference on Automatic Face & Gesture*

Recognition (FG 2018) 250–257 (2018). doi:10.1109/FG.2018.00044.

36. Terzis, V., Moridis, C. N. & Economides, A. A. Measuring instant emotions during a self-assessment test: the use of FaceReader. in *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research* 1–4 (Association for Computing Machinery, New York, NY, USA, 2010). doi:10.1145/1931344.1931362.
37. Kutaka, T. S., Chernyavskiy, P. & Hofkens, T. Achievement emotions in kindergarten: the association of solution accuracy with discrete joy, sadness, and surprise. *Front. Psychol.* 15, (2025).
38. Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11, 54–71 (2006).
39. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 14, (2001).
40. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6 (2020).

Figures

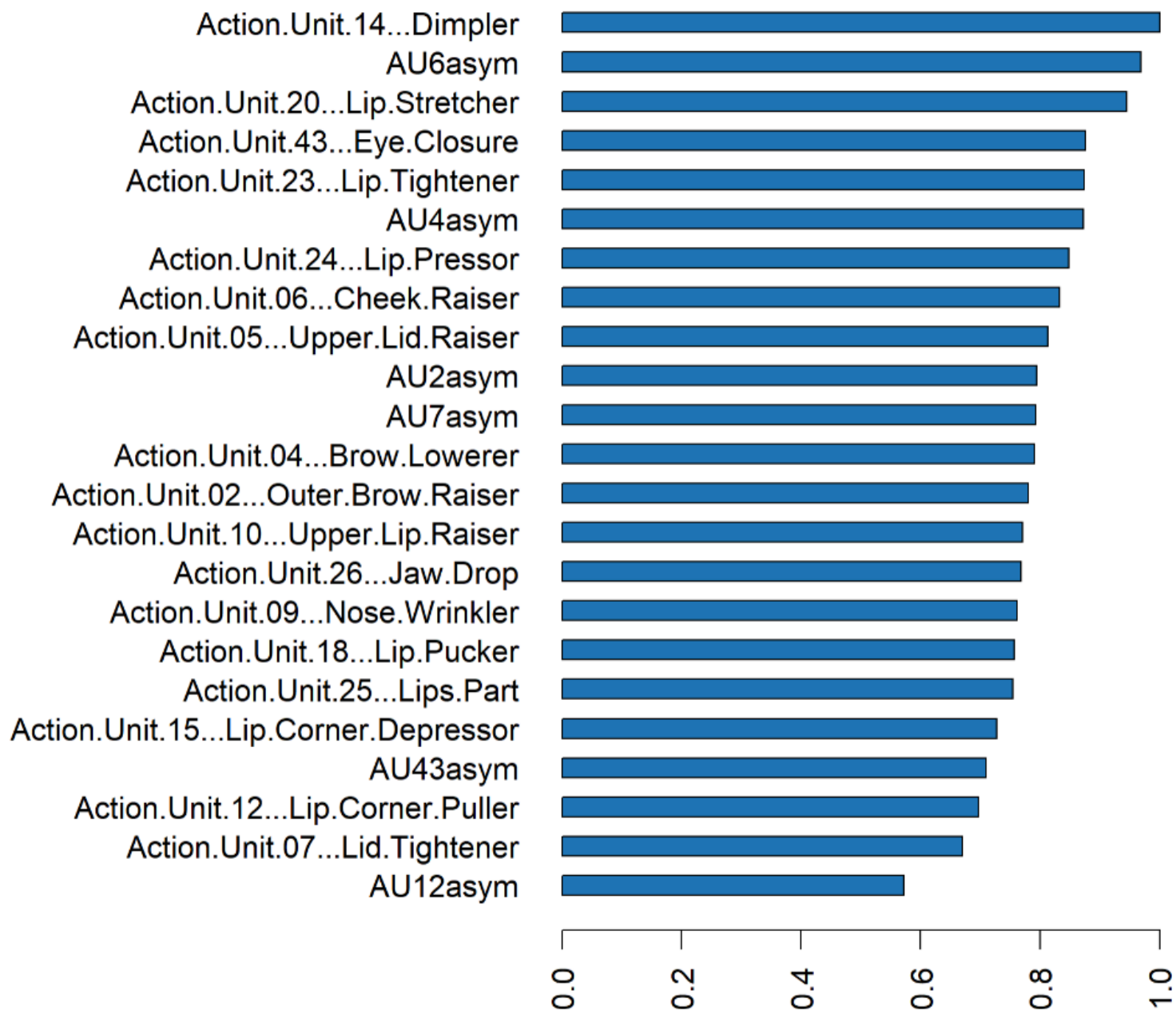


Figure 1

Variable (AU intensity) importance for the MLP model.

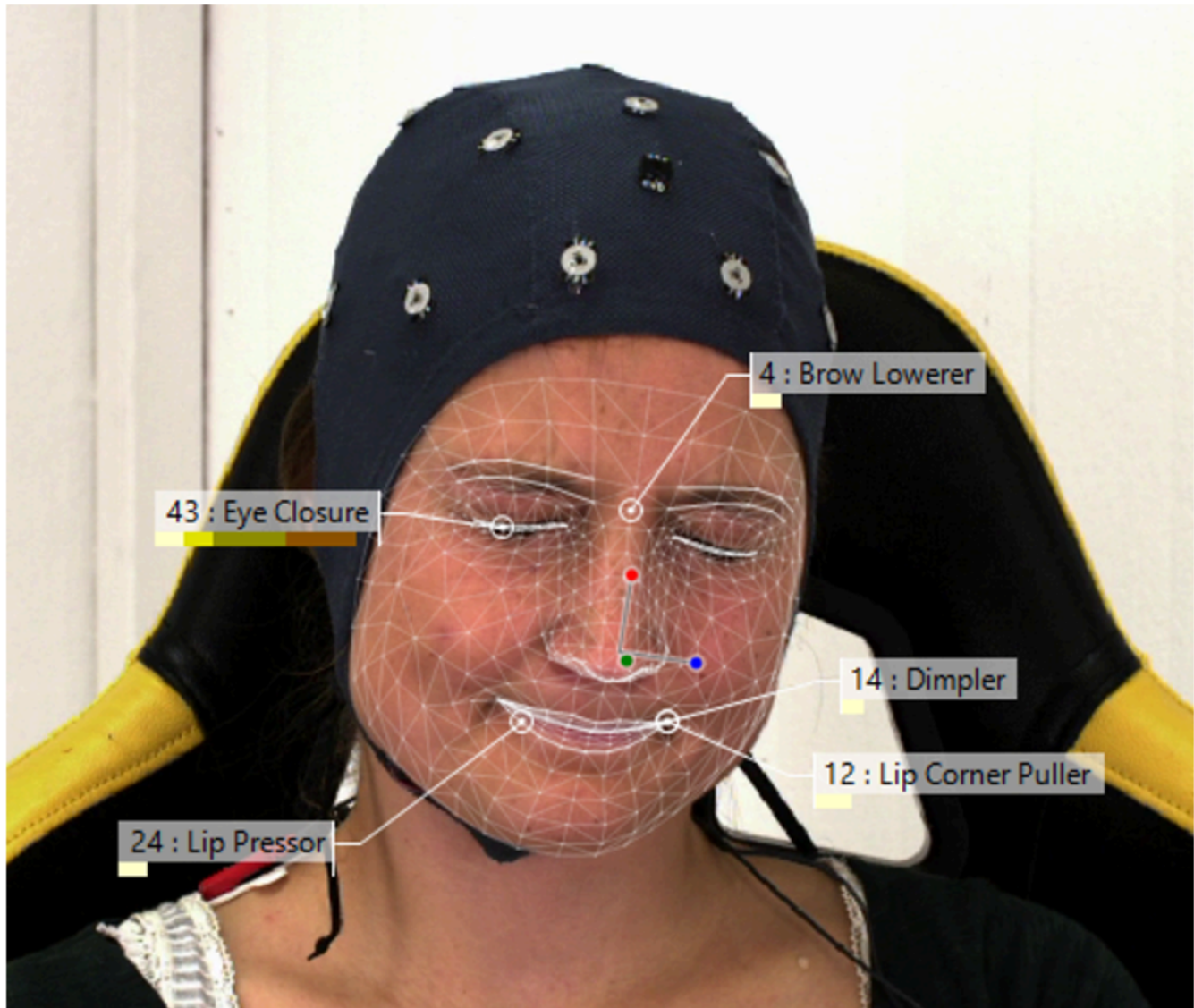


Figure 2

An illustration of FaceReader10 digitizing human facial expressions during experimentally-induced pain (BioVid exemplar).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [npjPainSupp1submit.docx](#)