The effect of human ratings reliability on machine learning model performance: a case study in infant pain assessment

Edoardo Passarotto[1], Daniel Müllensiefen[2], Ivan Tomasi[3], Valentina Brino[1], Matilde Paramento[1,4], Emanuela Formaggio[1], Stefano Masiero[1,5,6], Maria Rubega[1]

[1]Department of Neuroscience, University of Padova, Padova, Italy; [2] Institute of Systematic Musicology, University of Hamburg, Hamburg, Germany; [3]L'Inglesina Baby Spa, Altavilla Vicentina, Vicenza, Italy; [4]Department of Information Engineering, University of Padova, Padova, Italy; [5]Padova Neuroscience Center, University of Padova, Padova, Italy; [6]Ospedale Riabilitativo di Alta Specializzazione di Motta di Livenza, Motta di Livenza, Treviso, Italy

Corresponding author: Edoardo, Passarotto; Email: edoardo.passarotto@unipd.it

Funding statement. This work was supported in part by "Ministero delle Imprese e del Made in Italy" (D.M. MISE 31/12/2021 – Accordi per l'innovazione - Id: 10210 - MISE 127 - DESIGN FOR BABY WELLNESS). Edoardo Passarotto is supported by Fondo di Beneficenza Intesa San Paolo (Grant B/2024/0215, project title: "Un protocollo innovativo per l'indagine e il trattamento della scoliosi idiopatica dell'adolescente"). Matilde Paramento is supported by Fondazione Cassa di Risparmio di Padova e Rovigo (CUP C95F21009660007, project title: "Nuovi modelli di

integrazione tra trattamenti termali e terapia riabilitativa in soggetti con disabilità cronica nell'ambito del progetto di rilancio delle terme").

Competing interests. The author(s) declare none.

Abstract

Human-annotated data is foundational for supervised machine learning (ML). Low inter-rater reliability often introduces noise that degrades model performance. This study investigates how human rating reliability and panel size impact ML efficacy, and introduces a novel debiasing procedure utilizing Random Effects Models (REMs) to mitigate annotator noise. We conducted two complementary experiments to evaluate these dynamics. Experiment 1 analyzed real-world assessments from nine evaluators classifying 355 infant images. Results demonstrated that panel size and specific psychometric reliability indices—namely Cronbach's α, Generalizability, and Dependability coefficients—are strong predictors of ML performance across four algorithms, whereas inter-class correlation coefficients proved less robust. Experiment 2 generated simulated datasets mimicking Experiment 1 - incorporating 40 virtual raters with varying expertise levels and structured pattern noise - to evaluate the robustness of model aggregation across diverse rating scenarios. These simulations confirmed that while annotator noise significantly impairs classification, the proposed REM-based

debiasing procedure effectively recovers ground-truth scores. Notably, ML models trained on REM-debiased data from merely two raters achieved predictive performance comparable to models utilizing mean-aggregated scores from eight raters. Ultimately, this study underscores the critical importance of psychometrically sound data curation, demonstrating that advanced debiasing techniques can substantially enhance ML accuracy and efficiency even with small expert panels.

Keywords: machine learning; reliability; human judgments; ratings; debias.

1. Introduction

Machine learning (ML) models are commonly used to classify and label all sorts of perceptual objects and stimuli automatically. ML models are statistical algorithms that can learn from data (i.e., estimate sets of weights that represent the relationship between stimuli features and their corresponding labels) and generalize to unseen data. Object classification and data labelling are frequently performed by human experts who, similarly to ML models, assign attributes or class labels to the stimuli based on their characteristics. However, human ratings often suffer from poor reliability and agreement [1–3], which might lead to the misclassification of stimuli. The literature provides several examples of low agreement between human evaluators in many domains such as medicine, law, and aesthetic judgments [4–6]. Errors are therefore

expected, and their definition and identification strongly depends on the labelling process. In psychophysical terms, labelling tasks can fall into two main categories: performance and appearance tasks [7]. In the first case, human judgments can be compared to an objective ground truth (GT), as labels refer to objectively measurable characteristics of the stimuli. For instance, raters may be required to identify traffic-signs in real-world images [8,9], and errors can be identified by comparing labels to objective stimulus characteristics or meta-data. In appearance tasks, GT are not available as the labels strongly depend on the evaluator's perception, experience, and subjective understanding of a particular task or domain. This is the case, for example, for emotion classification from facial expressions, which relies on the subjective nature of emotion perception and discrimination. Classification errors cannot be measured from a GT and labelling reliability is usually estimated by collecting multiple judgments and assessing the agreement between multiple evaluators [10–12].

Kahneman et al. [4] suggests that human judgments in appearance and labelling tasks behave similarly to physiological or acoustic signals, where information often is recorded together with background noise. Instead of an objective GT, the reference value or label of an object can be estimated by aggregating ratings from a large number of unbiased evaluators [4]. This is based on the assumption that estimates based on data collected from large samples will be similar to those from other, similarly sized samples, and can therefore be generalized to the population from which the evaluators were drawn [13,14].

Labelling errors represent non-systematic deviations from the reference values of the objects being evaluated and are therefore referred to as *noise*. In Kahneman's framework [4], noise is further developed into two main concepts: level and pattern noise. Level noise represents the average deviation of all ratings given by an evaluator from the assumed reference value. In contrast, pattern noise refers to the deviations of the average rating given by an evaluator to specific subgroups of stimuli. While level noise provides a measure of overall strictness or leniency in rating behaviors relative to other raters, pattern noise identifies systematic bias for specific classes of stimuli. Thus, level and pattern noise assess interindividual differences in rating approaches, which may be affected by subjective perception and experience within a given field. When considering data from a large cohort of raters, level noise is expected to have a zero-centered distribution around the population mean rating, whereas pattern noise is assumed to have a zero-centered distribution around individual level noise intercepts. As with physiological or acoustic signals, a key property of noise in human judgements is its randomness. This suggests a simple and effective solution to deal with noise in human judgments: by collecting and aggregating data from enough raters, it is possible to account for noise and identify reliable ratings which can be generalized to larger populations [4,15].

When scores are aggregated from multiple evaluators, it is important to consider that it is assumed the data originates from the same GT, which was sampled with varying degrees of noise. Therefore, score aggregation

is only appropriate when data has been sampled from a homogeneous and consistent population. Proceeding with aggregation when the homogeneity assumption is violated entails practical risks: if raters are assessing fundamentally different constructs, the aggregated mean may represent an invalid compromise rather than a refined ground truth. Ultimately, reliability indices measure the agreement between evaluators and indicate the extent to which data aggregation will overlook differences in the rating approaches adopted by evaluators. In practice, problematic heterogeneity can be detected when reliability indices fail to improve with larger panel sizes, or by observing multimodal distributions in the raw ratings, which suggests the data should be segmented rather than pooled.

In the context of ML model development, noise that comes with a dataset containing human labels can have negative consequences for learning the correct relationships between stimulus features and labels during the training phase of ML models. Thus, noise coming from human labelling can negatively affect the performance of ML models and future use [16]. Several approaches have been suggested to overcome or mitigate the presence of noisy labels (for a comprehensive summary, see Frenay et al. [17]). These mitigation approaches can be grouped into three main categories: data filtering, rater penalization, and noise-robust algorithms. The first involves filtering out noisy observations that contribute substantially to the model's prediction error [18]. The second weights individual raters based on the agreement with their peers, penalizing those who deviate the most from the crowd [19]. The third approach uses

ML models that are insensitive to outliers and perform well even in the presence of noise [20]. These solutions mostly aim at minimizing the effects of noisy data or labels, ensuring model generalizability. While these approaches proved effective in overcoming noise in performance tasks (i.e., when a GT is available), they may be hardly applicable in appearance tasks as they underestimate the psychometric issues associated with noisy labels. On the one hand, filtering out observations or penalizing divergent raters implicitly assumes that noise can be attributed to single problematic cases which are not representative of the population from which they were drawn. On the other hand, robust ML models may be effective in improving model performance, but they do not necessarily improve the efficiency, validity, reliability, and generalizability of model predictions. These solutions may potentially bypass or suppress noise rather than accounting for it, limiting the extent to which ML models can be generalized and implemented in real-world scenarios. Finally, most solutions are specifically designed for ML models that classify stimuli into categories and may not be applicable to ML regression models, which assign a score to stimuli on a continuous scale. These limitations may be overcome by considering the psychometric properties of noise in human ratings.

A previous work by [5] provided a statistical implementation for Kahneman's framework in the context of aesthetic judgments [4]. By means of random effects regression models (REMs), level and pattern noise were operationalized as random intercepts and their contribution to the overall variability in ratings was assessed in terms of variance

components. Thus, these models were used to aggregate individual ratings while accounting for noise and the complex structure of individual rating schema, demonstrating the advantages of this aggregation method over more common alternatives such as mean and median aggregation. By means of sensitivity analyses on data collected from human raters as well as data simulations, the authors demonstrated a positive association between panel size (i.e., the number of human raters) and the reliability of aggregated ratings, in line with Kahneman's framework [4]. Furthermore, the Generalizability (G) and the Dependability (D) reliability coefficients from the Generalizability Theory [13] were particularly sensitive to noise and effective in measuring the reliability of aggregated ratings.

This study investigates the relationship between noise in ratings and model performance in the context of supervised ML algorithms and appearance tasks involving labels assigned by humans, using an approach similar to that presented in [5]. It aims to a) quantify the effect of panel size on ML model performance and identify reliability indices (RI) of human rating data that effectively predict ML model performance, and b) explore the use of a debiasing procedure based on REMs to improve ML model performance, comparing it to alternative methods. To achieve these objectives, extensive sensitivity analyses were conducted on both real-world (Experiment 1) and simulated data (Experiment 2).

2. Experiment 1

The first part of the study focuses on real-world data and contextualizes the issues within the assessment of pain from facial expressions in newborns and infants. Experiment 1 considered the development of an ML model to assess pain in infants aged between 0 and 12 months, applicable in both clinical and domestic settings. The assessment of pain in infants is challenging due to their limited ability to communicate. However, this challenge is of significant clinical importance, as effective monitoring and timely intervention are crucial for patient management [21–23]. Several studies have explored the use of ML models to automatically classify pain from images of infants and newborns [24,25]. Despite the good results in terms of model performance, the approaches proposed in the literature present some common limits. First, pain is usually considered a dichotomous variable [26,27] (i.e., pain response to heel prick devices compared to resting state), discarding its complex multidimensional nature. Second, the human labeling process is vaguely described, and standardized reliability indices are often omitted [26,28,29]. Third, databases implementing different pain scales are sometimes merged [28], disregarding possible differences in the operational definitions of pain. In Experiment 1, ratings were collected from a panel of nine raters who evaluated images of infants and newborns from two publicly available databases using two standardized multilevel pain scales. Extensive sensitivity analyses were performed on this data, to explore the relationship between sample size, RI and model performance. We expect to find significant associations between N, RI and ML performance. In addition, score aggregation based on REMs

should lead to more reliable score aggregates and better ML model performance compared to alternative aggregation methods, in line with [5]. A greater and more reliable cohort of raters should provide a more complete representation of the population from which it was drawn, thereby reducing sampling error. Thus, the value of the stimuli under evaluation should be estimated with better precision and reliability. When trained on a subset of reliable ratings, it should be more likely that ML models will learn weights that can effectively predict ratings for unseen data partitions, demonstrating superior predictive performance.

2.1. Methods

2.1.1. Dataset

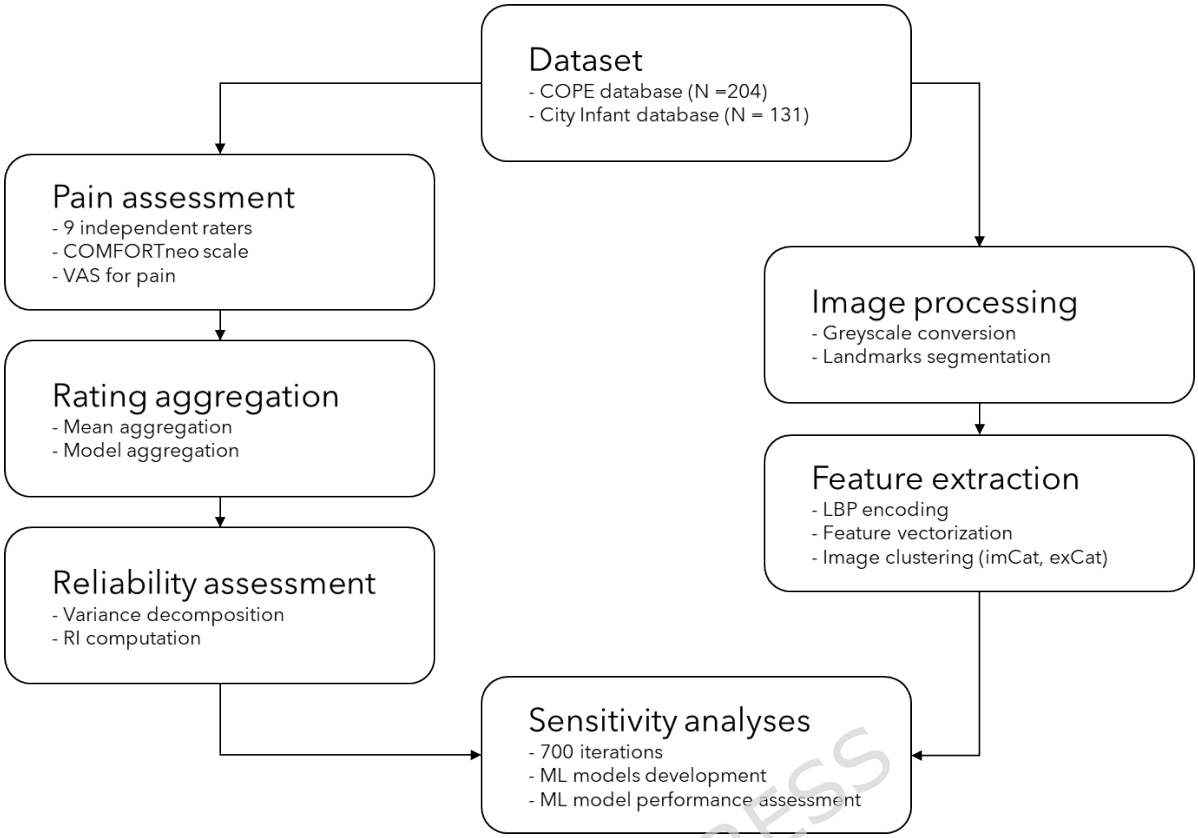
355 images of newborns and infants were used in the study. The dataset was created by merging the Infant COPE and the City Infant databases [26,27]. The first is a collection of 204 images of newborns, taken between 18 and 36 hours from their birth. 50% of the images depicted female infants. The pictures were collected during different experimental conditions, including rest and procedural pain induced by heel prick tests. The second collects 154 images of infants between 0 and 12 months of age. In this database, the mean age of the infants was 6.56 months (SD = 2.73) and 61% were females. The pictures were taken under ecologically valid conditions and present a broad variety of facial expressions. After removing duplicates (i.e., files labelled as copies of others were excluded, then all images were visually inspected to verify their content and ensure uniqueness), 131 images from the City Infant database were retained for the analysis. The use of two datasets was

meant to increase the number of stimuli and their variability, widening the age range considered. Further information regarding the images used in this study can be found in [26, 27].

2.1.2. Pain measures and assessment

Pain was assessed through two measurement scales: the *facial tension* subscale from the COMFORTneo scale [30] and a Visual Analogue Scale (VAS) for pain. The former consists of a single item scale, ranging from 1 to 5 points. Each level of the scale is accompanied by a description of the morphological features typically associated with that level of pain. Higher COMFORTneo scores indicate greater discomfort. VAS consists of a 10 cm line, representing the distance between the absence of pain and the worst pain possible, corresponding to a VAS score of 0 and 10 respectively. Subjects were asked to graphically indicate the level of pain perceived by moving a cursor between the two extremities. VAS scores were calculated by rounding subjects' input to two decimal places.

Figure 1. Experimental procedure and data processing pipeline.



Note: the right side of the figure shows the workflow used to extract features from the images in the database. The left side summarizes the assessment and data aggregation procedure followed to generate pain scores. Feature vectors and aggregated ratings were subsequently used to run sensitivity analyses, train ML models, and compare aggregation methods (i.e., mean-aggregation VS model-aggregation) in terms of ML model performance.

Ratings from a panel of 9 independent raters were collected. The panel included 1 pediatrician, 4 researchers from the local university, and 4 parents. On the one hand, the different profiles included in the panel were intended to provide a high-level representation of pain that could be generalized to different subgroups of users. On the other hand, it could increase variability in ratings, providing an ideal context for the subsequent analysis. The parents did not have a familiar relationship with any of the other raters. All raters voluntarily participated in the study. A sample size of 9 raters with all raters assessing all stimuli is within the range of similar studies focusing on classification of facial

expressions in infants [31–35]. The agreement between raters was good, with inter-class correlation (ICC2) values greater than .80 as well as G and D values greater than .97 for both ratings scales. Each rater evaluated all 355 images in the database using both the COMFORTneo and VAS scales. Ratings were collected in person, through a computerized assessment platform programmed in MATLAB [36] where the order of the images was randomized across participants. The experimental procedure and data processing pipeline are summarized in Figure 1.

2.1.3. Image processing and feature extraction

The image preprocessing and feature extraction pipeline used in this study is the same as in Martínez et al [37]. Five salient landmarks were segmented from the images, including right and left eye, eyebrow, nose, and mouth. Subsequently, image chunks were converted to the grayscale, normalized to reduce differences in illumination between the two databases, resized to matrices of fixed dimensions, and encoded into Local Binary Patterns (LBP). This method has been often used for image classification, including pain detection algorithms [37–40]. Finally, LBP matrices of different landmarks were vectorized and concatenated, resulting in vectors of 1920 features per image. Thus, an n*p features matrix was created, where n represents the number of images in the dataset and p the number of features extracted. For detailed information about the LBP feature extraction procedure, see Martínez et al [37].

2.1.4. Image categorization

Noise was assessed on different categories of images, namely explicit image categories and implicit image categories. Explicit image categories (*exCat*) were manually specified by the researcher. Each image was labeled as either belonging to the Infant COPE or the City Infant database (*exCat1* binary variable). This took into account the differences in the characteristics of the two datasets, such as the age range, image quality and shooting conditions. In addition, the *exCat2* variable represented the original pain and comfort classification proposed by the authors of the COPE and City Infant databases [26,27]. For *exCat2*, pain and discomfort classes from the Cope and City Infant databases were considered equivalent. Implicit image categories, *imCat*, represented data-driven clusters of images identified in the feature matrix by k-means algorithms. Each *imCat* category groups together images with specific combinations of features arising from technical and content aspects (i.e., capturing similarity in faces and in the way the image was shot). A large value for k is plausible to capture the many idiosyncrasies of the images included in the dataset. For this study, k was set to 53, namely the largest value of k that grouped at least two images per cluster.

2.1.5. Rating aggregation methods

Individual ratings were aggregated through two alternative approaches: mean-based score aggregation (*mean-aggregation*) and model-based score aggregation (*model-aggregation*). In the first case, each image was scored as the average of the individual ratings. In the second case, model-aggregation was performed via REMs, which allowed us to estimate and account for noise in ratings. The REMs used individual

human ratings as dependent variable (y), while image identifiers (*stim*), rater identifiers (*rater*), explicit image categories (*exCat1* and *exCat2*), implicit image categories (*imCat*), and the interactions *rater:exCat1, rater:exCat2,* and *rater:imCat* were used as random effects. Note that the model did not contain any fixed effects and can therefore be considered a variance decomposition model. The model is represented by the following probability distribution:

$$y = N(\mu + u_{stim} + u_{exCat1} + u_{exCat2} + u_{imCat} + u_{rater} + u_{rater:exCat1} + u_{rater:exCat2} + u_{rater:imCat}, \sigma^2) \quad (1)$$

Each random effect (u) was estimated as a zero-centered normal distribution:

$$u = N(0, \sigma_u^2) \quad (2)$$

Note that REMs estimate individual coefficients through partial pooling, thus penalizing raters who show inconsistent rating behaviors [41]. Individual coefficients are computed as the weighted sum of no-pooling and complete-pooling estimates, as shown in the following formula:

$$u\hat{}_i \approx \frac{\frac{n_i}{\sigma_i^2} * \underline{u_i} + \frac{1}{\sigma^2} * \underline{u}}{\frac{n_i}{\sigma_i^2} + \frac{1}{\sigma^2}} \quad (3)$$

n_i, $\underline{u_i}$, and σ_i^2 represent the number of observations, mean, and variance of the i-th group (i.e., rater or noise category) while $\underline{u}$ and σ^2 represent the mean and variance across all groups. Partial-pooling operates hierarchically across each predictor, estimating a unique shrinkage factor for every level. By weighing the group-specific average against the global mean based on sample size and variance, the model prevents over-

shrinkage: estimates supported by consistent, abundant data retain their distinct values, while only those based on sparse or highly variable data are pulled strongly toward the population mean. In the model, level noise is represented by *rater* main effect, while the interactions *rater*:*exCat1, rater*:*exCat2,* and *rater*:*imCat* quantify pattern noise. The total amount of noise in the data is estimated as the sum of level and pattern noise variance components. Thus, ratings (y) were aggregated through predictions from the model, by reformulating its structure to account for interindividual variability in ratings (i.e., level and pattern noise) as follows:

$$y = N(\mu + u_{stim} + u_{exCat1} + u_{exCat2} + u_{imCat}, \sigma^2) \quad (4)$$

An example of the difference between score aggregates computed by the two alternative methods (i.e. mean-aggregation and model-aggregation) is reported in Figure 2.

2.1.6. Reliability assessment

Five reliability indices (RI) were computed to quantify the agreement between raters and the reliability of the estimated ratings: Cronbach's α, ICC2, ICC3, G, and D coefficients [13,42].

Cronbach's α is a measure of internal consistency, which assesses how closely related a set of items are as a group and is commonly used to evaluate scale reliability. The formula for α is:

$$\alpha = \frac{N * \underline{c}}{\underline{v} + (N-1) * \underline{c}} \quad (5)$$

where N is the number of items, $\underline{c}$ is the average covariance between item pairs, and $\underline{v}$ is the average variance of each item. To compute the remaining RI, variance components, residual and overall model variance were estimated through REMs. Individual variance components were computed by raising each random effect coefficient in formula 1 to the power of two. Subsequently, ICC2 and ICC3 estimate raters' agreement and consistency respectively by considering the ratio between the variance explained by the object evaluated (i.e., the images, σ^2_{stim}), the variance due to the raters (σ^2_{rater}), and the residual variance ($\sigma^2_{residual}$).

$$ICC2 = \frac{\sigma^2_{stim}}{\sigma^2_{stim} + \sigma^2_{rater} + \sigma^2_{residual}} \quad (6)$$

$$ICC3 = \frac{\sigma^2_{stim}}{\sigma^2_{stim} + \sigma^2_{residual}} \quad (7)$$

G and D [13] assess the reliability of the relative and absolute score values, respectively. Compared to ICC coefficients, G and D measure reliability of complex variance structures and penalize each noise category by the panel size, as shown in the following formulae:

$$G = \frac{\sigma^2_{stim} + \sigma^2_{exCat1} + \sigma^2_{exCat2} + \sigma^2_{imCat}}{\sigma^2_{stim} + \sigma^2_{exCat1} + \sigma^2_{exCat2} + \sigma^2_{imCat} + \frac{\sigma^2_{rater:exCat1}}{n_{rater}} + \frac{\sigma^2_{rater:exCat2}}{n_{rater}} + \frac{\sigma^2_{rater:imCat}}{n_{rater}} + \frac{\sigma^2_{re}}{n_r}} \quad (8)$$

$$D = \frac{\sigma^2_{stim} + \sigma^2_{exCat1} + \sigma^2_{exCat2} + \sigma^2_{imCat}}{\sigma^2_{stim} + \sigma^2_{exCat1} + \sigma^2_{exCat2} + \sigma^2_{imCat} + \frac{\sigma^2_{rater}}{n_{rater}} + \frac{\sigma^2_{rater:exCat1}}{n_{rater}} + \frac{\sigma^2_{rater:exCat2}}{n_{rater}} + \frac{\sigma^2_{rater:im}}{n_{rater}}} \quad (9)$$

The values of all five RI range from 0 to 1, with higher values indicating greater ratings' reliability.

2.1.7. Sensitivity analysis: the effect of panel size and aggregation method

To assess the effects of panel size (N) and data aggregation method on ML model performance, a sensitivity analysis was run on real data. At each iteration of the analysis, ratings from a random subset of raters were selected and RI were computed on the selected data. Then, individual scores were aggregated by either mean-aggregation or model-aggregation methods (*debias*). Four ML models were trained on the feature matrix and the aggregated ratings. Their performance was evaluated on 5 non-overlapping data folds, resulting in a 5:1 ratio between train and test data (see the next paragraphs). The number of randomly selected raters ranged between 2 and 8, increasing by one unit at each iteration. The entire procedure was repeated 10 times. Therefore, the sensitivity analysis had a 7 (number of raters) * 5 (folds) * 10 (runs) *2 (aggregation method: mean- vs model-aggregation) structure, resulting in 700 iterations in total.

2.1.8. ML models

Four ML regression models were trained to predict pain ratings based on feature vectors: k-nearest neighbor (KNN), least-squares support vector machine (LSSVM), random forest (RF), support vector machines (SVM) algorithms. In addition, LSSVM represents a noise-robust variant of SVM

models [17]. KNN, LSSVM, and RF models were implemented using the *caret* R package [43]. For these models, hyperparameter-tuning was run on pain scores from the entire panel of raters aggregated via mean-aggregation (see Appendix 1). Note that hyperparameter-tuning led to similar results (i.e. hyperparameters) for the VAS and COMFORT neo scales. LSSVM were run using the *neo-ls-svm* python package [44], which offered adaptive hyperparameters tuning at each sensitivity analysis iteration.

ML model performance was evaluated out-of-sample using a 5:1 ratio between train and test data. Three performance indices, assessing the difference between ML model predictions and the actual data, were computed: R2, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). R2 measures the proportion of the variance in the dependent variable explained by independent variables, with values ranging from 0 to 1. Higher values indicate better model fit. RMSE is the square root of the average of the squared differences between the predicted and actual values, providing a measure of the model's prediction error in the same units as the dependent variable. MAE is the average of the absolute differences between the predicted and actual values, offering a straightforward measure of prediction accuracy. R2, unlike RMSE and MAE, is independent of the units and range of the dependent variable (i.e., the rating scale in use), providing a scale-free measure of model performance. When comparing the effectiveness of mean-aggregation or model-aggregation methods, ML models were trained and tested on the same data partitions.

2.1.9. Statistical analysis

The initial analysis consisted of a variance decomposition of the raw rating data using a REM following Eq. 1. Next, the association between N, RI, and model performance was assessed using linear mixed-effects models (LMM), with performance indices (i.e., R^2, RMSE, and MAE) as dependent variables, N and RIs (i.e., N, α, ICC2, ICC3, G, and D) separately entered as fixed effects. The model intercept was allowed to vary across runs of the sensitivity analysis (see section 2.1.7. Sensitivity Analysis), thereby representing the random effects component of the LMMs. The effect of each predictor (i.e., N and individual RIs) was assessed separately. N was normalized to the 0-1 range. Thus, N represents the proportion of raters in the panel considered at each iteration, where N = 0 corresponds to 2 raters and N = 1 corresponds to 8 raters. All other predictors were standardized across iterations (i.e., M = 0, SD = 1). Thus, model coefficients estimated the change in the dependent variables associated with a 1SD increase in the predictors. Finally, LMMs were used to assess the effect of noise and rating aggregation methods on model performance. In this case, *debias*, N, and *debias*:N interaction predicted performance indices as fixed effects and random intercepts per each sensitivity analysis run were modelled. *Debias* was dummy coded as *mean-aggregation* = 0 and *model-aggregation* = 1 while N was normalized to the 0-1 range. Given their bounded range (i.e., between 0 and 1), R2 values were modelled via beta distributions and the logit link function. As the RMSE and MAE values computed in the analyses were greater than 0, the two variables were

modelled via gamma distributions and the log link function. P-values for the LMM models were estimated via Satterthwaite's degrees of freedom method and no correction for multiple testing was applied. When conducting a sensitivity analysis, it is important to consider that increasing the number of iterations would expand the set of observations analyzed in the regression models and likely decrease the p-values. This is because sensitivity analysis is an inherently exploratory approach, focusing primarily on identifying consistent patterns across conditions rather than conducting individual significance tests. Therefore, when interpreting the regression models, the main focus should be on effect sizes, which remain unaffected by the number of iterations. Additionally, the effects estimated by the LMM models, together with their 95% confidence intervals, are reported in the Appendix. Statistical analyses, data aggregation and RI computation were run in RStudio (version: 4.5.1, RRID: SCR_000432) [45], through the lme4 [46] and psych [47] R packages.

2.2. Results

Table 1. Mean and range of variance components, RI, and performance indices computed on VAS and COMFORTneo scores over the sensitivity analyses on actual data.

Variance components	VAS	COMFORTneo
σ^2_{stim}	.069 [.035 - .134]	.067 [.049 - .106]
σ^2_{exCat1}	.027 [.001 - .075]	.035 [.004 - .066]
σ^2_{exCat2}	.511 [.459 - .570]	.551 [.457 - .636]

σ^2_{imCat}		.214 [.147 - .289]	.196 [.170 - .240]
σ^2_{rater}		.029 [.001 - .135]	.019 [.001 - .093]
$\sigma^2_{rater:exCat1}$		.005 [.001 - .012]	.006 [.001 - .028]
$\sigma^2_{rater:exCat2}$		.036 [.001 - .087]	.027 [.001 - .089]
$\sigma^2_{rater:imCat}$		.014 [.001 - .037]	.011 [.001 - .027]
$\sigma^2_{residual}$		.094 [.043 - .189]	.089 [.062 - .113]
RI		VAS	COMFORTneo
N		5 [2 - 8]	5 [2 - 8]
α		.946 [.835 - .977]	.944 [.850 - .973]
ICC2		.695 [.475 - .856]	.734 [.513 - .846]
ICC3		.789 [.695 - .865]	.800 [.707 - .860]
G		.957 [.875 - .980]	.962 [.861 - .982]
D		.949 [.844 - .977]	.957 [.827 - .981]
PI	ML model	VAS	COMFORTneo
R2 ~	KNN	.656 [.364 - .834]	.637 [.306 - .831]
	LSSVM	.684 [.434 - .836]	.680 [.406 - .843]
	RF	.675 [.398 - .843]	.671 [.337 - .812]
	SVM	.706 [.437 - .872]	.695 [.408 - .867]
RMSE ~	KNN	1.520 [.984 - 2.409]	.779 [.477 - 1.185]
	LSSVM	1.436 [.984 - 2.185]	.726 [.525 - .969]
	RF	1.753 [1.198 - 2.633]	.876 [.617 - 1.109]
	SVM	1.419 [.933 - 2.191]	.720 [.490 - .970]
MAE ~	KNN	1.180 [.724 - 1.929]	.595 [.377 - .891]
	LSSVM	1.143 [.726 - 1.702]	.579 [.418 - .792]
	RF	1.512 [1.006 - 2.132]	.744 [.516 - .968]

	SVM	1.104 [.714 - 1.679]	.565 [.391 - .794]

Note: PI = performance indices. Variance components and R2 values are reported as a percentage of the overall model variance. The values of α, ICC2, ICC3, G, and D coefficients range from 0 to 1, with higher values indicating greater ratings' reliability. RMSE and MAE measure prediction error in the same units as the dependent variable. Level noise was assessed by σ^2_{rater} variance component while $\sigma^2_{rater:exCat1}$, $\sigma^2_{rater:exCat2}$, and $\sigma^2_{rater:imCat}$ quantified pattern noise.

Descriptive statistics for variance components, RIs and performance indices calculated across the sensitivity analyses on actual data are shown in Table 1. On average, across 700 iterations, noise accounted for 8.4% (range: 0.1–21.4) and 6.3% (range: 2.2–19.2) of the variance in VAS and COMFORTneo ratings, respectively. *ExCat2* and *imCat* variance components explained the largest proportion of variance in both VAS and COMFORTneo ratings. On average, SVM models showed the best performance, regardless of the rating scale. LSSVM models, a noise-robust variant of SVM, did not provide a concrete performance advantage over other algorithms.

2.2.1 Association between N, RI and ML model performance

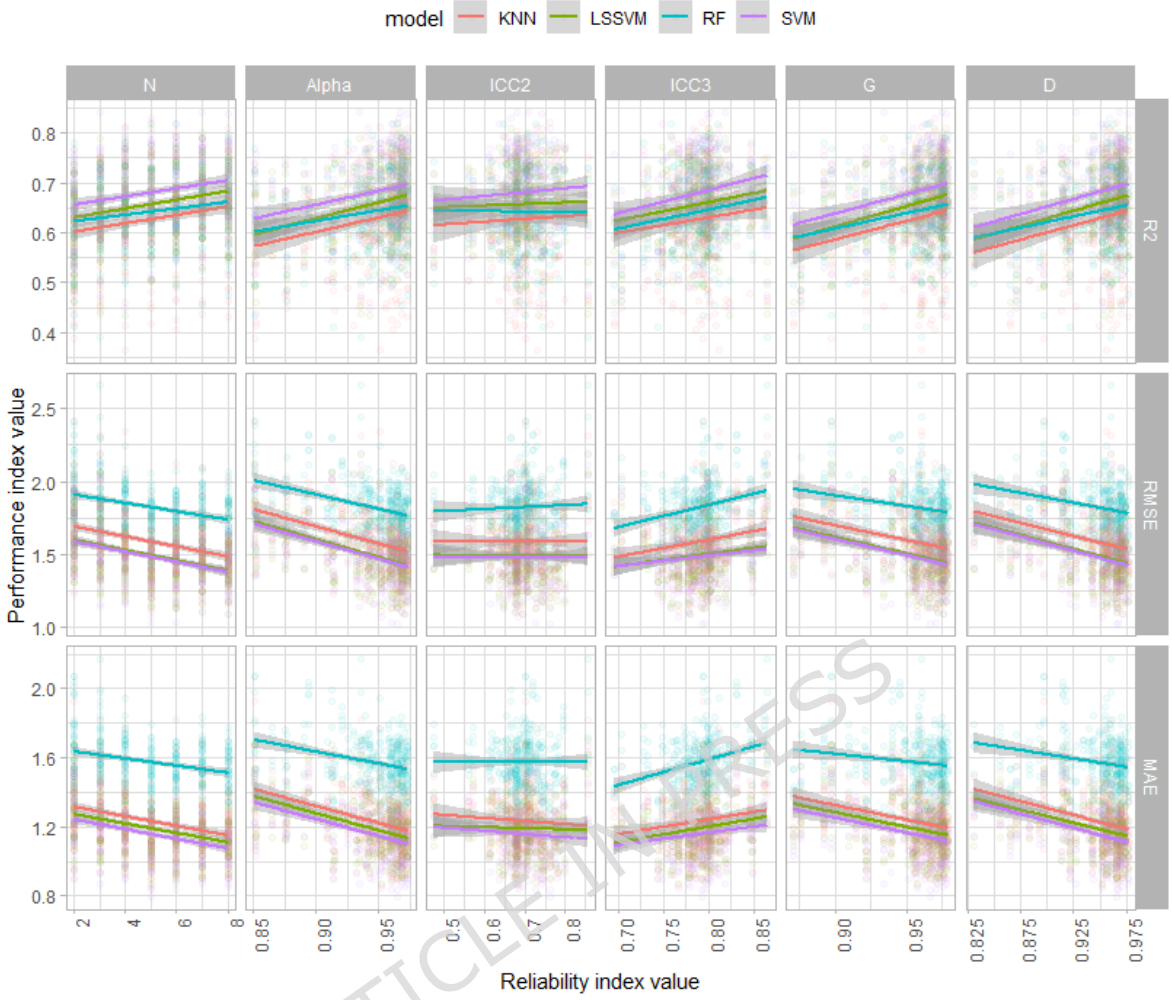
The results of the sensitivity analyses showed that N, α, G and D were significantly associated with increases in R2 as well as decrements in MAE and RMSE values (see Table 2). ICC2, and ICC3 were significantly related to R2, RMSE, and MAE values. However, these relationships were inconsistent, especially between ICC3 coefficients and performance indices. Indeed, high ICC3 scores were associated with greater RMSE and MAE values, indicating an anomalous positive correlation between ratings' reliability and prediction error measures, regardless of the model and rating scale (see Figure 2 and Table 2).

Table 2. Effect estimates of N and RI on performance indices of KNN, LSSVM, RF, and SVM models predicting VAS and COMFORTneo scores.

		VAS			COMFORTneo		
IV	ML model	R2	RMSE	MAE	R2	RMSE	MAE
N	KNN	.219*	-.128*	-.133*	.193*	-.082*	-.090*
	LSSVM	.235*	-.137*	-.133*	.246*	-.103*	-.102*
	RF	.168*	-.092*	-.076*	.161*	-.053*	-.036*
	SVM	.221*	-.138*	-.143*	.220*	-.096*	-.108*
α	KNN	.080*	-.045*	-.049*	.071*	-.020*	-.025*
	LSSVM	.090*	-.051*	-.050*	.085*	-.027*	-.027*
	RF	.063*	-.033*	-.027*	.064*	-.011*	-.007*
	SVM	.085*	-.050*	-.052*	.080*	-.025*	-.030*
ICC2	KNN	.035*	-.007	-.015*	.054*	.003	.001
	LSSVM	.024*	-.007	-.010	.055*	.002	.001
	RF	.009	-.001	-.004	.053*	.010*	.012*
	SVM	.030*	-.006	-.014*	.050*	.004	.001
ICC3	KNN	.046*	.019*	.019*	.045*	.005	-.002
	LSSVM	.058*	.012*	.018*	.045*	.004	.003
	RF	.055*	.022*	.025*	.052*	.008*	.008*
	SVM	.069*	.011*	.015*	.046*	.004	.001
G	KNN	.083*	-.035*	-.037*	.070*	-.018*	-.022*
	LSSVM	.094*	-.041*	-.038*	.085*	-.024*	-.025*
	RF	.071*	-.023*	-.016*	.066*	-.009*	-.005*
	SVM	.092*	-.041*	-.041*	.079*	-.022*	-.028*
D	KNN	.083*	-.039*	-.043*	.074*	-.017*	-.021*
	LSSVM	.090*	-.043*	-.042*	.088*	-.024*	-.024*
	RF	.066*	-.026*	-.021*	.067*	-.008*	-.003
	SVM	.090*	-.044*	-.046*	.081*	-.021*	-.026*

Note: * $p < .05$. N was normalized to the 0–1 range. N = 0 corresponds to two raters, and N = 1 corresponds to eight raters. All other IVs were standardized across iterations. Analyses performed on real data across 700 sensitivity analysis iterations. 95% confidence intervals of the effects estimated by the LMM models are reported in Appendix 2.

Figure 2. Association between N, RI, and performance indices of KNN, LSSVM, RF, and SVM models predicting VAS scores.



Note: the figure shows the impact of panel size (N) and reliability indices on the performance (R2, RMSE, MAE) of ML models in predicting VAS scores. Alpha, G and D coefficients show a stronger and more consistent correlation with model performance indices than standard ICC coefficients. Similar associations have been found for COMFORTneo scores.

2.2.2 Aggregation methods comparison

Figure 3 shows a graphical comparison between aggregates of COMFORTneo scores computed by mean-aggregation (in red) and model-aggregation (in blue) on a random subset of 100 stimuli. The analyses indicated that *debias* was associated with improvements in model performance, increasing R2 and reducing MAE and RMSE values. As shown in Table 3, the effect estimates for *debias*, β = .266 [.182 - .334],

were slightly greater than those for N, β = .207 [.161 - .245]. This suggests that model-aggregation performed consistently better than mean-aggregation even with the maximum number of evaluators considered in the study (N = 8). Thus, our data shows that models trained using model-aggregated scores from two raters perform similarly to those trained using mean-aggregated scores from eight raters (see Figure 4 and Appendix 4 for further details). This means that researchers can achieve reliable ground truths with significantly fewer human annotators by applying statistical debiasing post-processing, thereby reducing data collection costs (by 75% in this study). It is important to note that the effects of debias and N were not cumulative, and the effect of debias was reduced in large samples of raters (see Figures 4 and 3). Phrased differently, model aggregation shows its greatest advantage with samples with relatively low numbers of human raters (i.e., N < 6).

Figure 3. Comparison between mean-aggregation and model-aggregation scores computed on ratings of 100 infant images.

Note: the figure shows COMFORTneo scores assigned by nine independent raters (in gray) to 100 stimuli as well as score aggregates computed by mean-aggregation (in red) and model-aggregation (in blue). Stimuli are ordered by mean-aggregation score values.

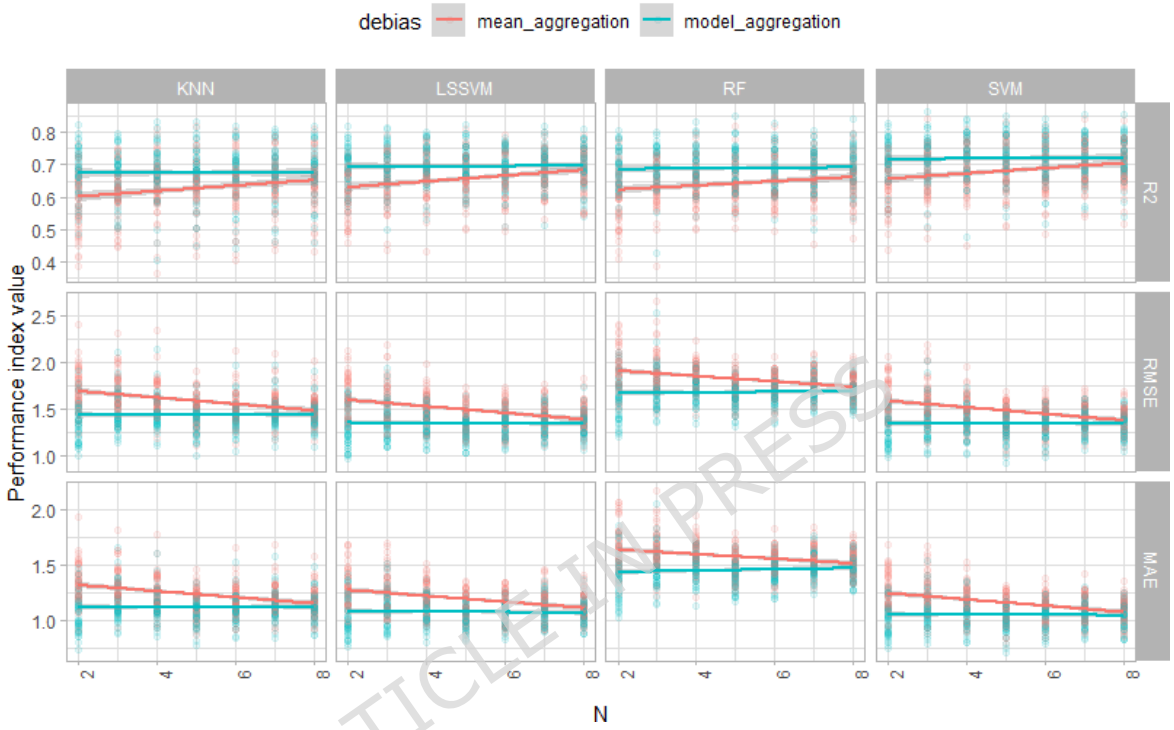
Table 3. Summaries of the LMM models assessing the effect of N, noise and debias on performance indices of KNN, LSSVM, RF, and SVM models predicting VAS and COMFORTneo scores.

		VAS			COMFORTneo		
ML_model	IV	R2	RMSE	MAE	R2	RMSE	MAE
KNN	Intercept	.421*	.520*	.269*	.329*	-.173*	-.438*
	debias	.318*	-.167*	-.167*	.230*	-.153*	-.155*
	N	.219*	-.128*	-.133*	.193*	-.082*	-.090*
	debias:N	-.217*	.136*	.138*	-.185*	.124*	.128*
LSSVM	Intercept	.541*	.464*	.235*	.489*	-.237*	-.465*
	debias	.293*	-.168*	-.168*	.334*	-.154*	-.153*
	N	.235*	-.137*	-.133*	.245*	-.104*	-.102*
	debias:N	-.236*	.131*	.128*	-.207*	.132*	.132*
RF	Intercept	.509*	.643*	.488*	.485*	-.069*	-.241*
	debias	.284*	-.134*	-.131*	.182*	-.128*	-.121*
	N	.168*	-.092*	-.075*	.161*	-.053*	-.036*
	debias:N	-.144*	.106*	.102*	-.134*	.101*	.098*
SVM	Intercept	.659*	.457*	.211*	.591*	-.247*	-.477*
	debias	.286*	-.166*	-.168*	.203*	-.149*	-.158*
	N	.221*	-.138*	-.143*	.220*	-.096*	-.108*
	debias:N	-.200*	.138*	.137*	-.170*	.122*	.130*

Note: * p < .05. N was normalized to the 0–1 range. N = 0 corresponds to two raters, and N = 1 corresponds to eight raters. Debias was dummy coded as 0 = mean-

aggregation and 1 = model-aggregation. Analyses were run on real data across 700 sensitivity analysis iterations. 95% confidence intervals of the effects estimated by the LMM models are reported in Appendix 3.

Figure 4. Association between model-aggregation and performance indices of KNN, LSSVM, RF, and SVM models predicting VAS scores.



Note: the figure shows the effect of model-aggregation over mean-aggregation on the performance (R2, RMSE, MAE) of ML models in predicting VAS scores. Model-aggregation provided substantial improvements in model performance, particularly with small panels of evaluators. Similar associations have been found for COMFORTneo scores.

2.3. Discussion

The results of Experiment 1 broadly confirm the hypotheses of the study: around 5-8% of the variance in the data can be considered noise due to the human raters, either as level noise or pattern noise. The RIs computed on the dataset are significantly associated with ML performance. The most effective predictors are G and D coefficients from

generalizability theory, as they showed the strongest and most consistent association with ML model performance. Employing REMs for the aggregation of human ratings increases the performance of the ML models compared to simple mean aggregation, especially in panels with low numbers of raters (N < 6). The advantage of model aggregation is up to 9% of the variance explained by the ML model. As additional insight we found that the SVM models outperformed the other ML methods across all conditions of the sensitivity analysis and all three performance indices. Knowing that the dataset used in Experiment 1 represents only a single, medium-sized dataset showing high inter-rater reliability (ICC2 > .80), we conduct simulations in Experiment 2 to investigate whether the advantage of model aggregation and the usefulness of the coefficients G and D as indicators of the level of noise present in the data holds true across a range of similar data scenarios.

3. Experiment 2

Simulated data was generated to assess the validity of the results of Experiment 1 and to understand whether noise and interindividual variability in ratings were the main cause of the results found in the real data. We expect noise to significantly affect both RI and the performance of ML models, with larger amounts of noise decreasing the performance of the ML model more. The effect of noise should be less severe in the presence of large panel sizes.

3.1. Methods

3.1.1. Data simulations

Simulated data were designed to mimic the data collected and analyzed in Experiment 1 and were generated by the following steps. First, 335 GT pain ratings were generated: these were uniformly distributed in the 0-10 range, similarly to VAS scores (see section *pain measures and assessment*). Second, n*p feature matrices were computed, simulating vectors of 1920 features (p) from 335 stimuli (n). Each feature was moderately correlated with GT scores, with Pearson's r values ranging between .381 and .980. Third, artificial ratings from a panel of 40 raters were generated by adding random Gaussian noise (M = 0, SD = 1.5) to GT scores. Fourth, pattern noise was simulated for implicit and explicit categories of stimuli: 100 implicit noise categories (*imCat*) were identified by k-means algorithms, based on clusters of data found in the feature matrix. Explicit noise categories (*exCat*) split the sample into ten random subgroups of equal size. Pattern noise was simulated independently for each rater adding random values from a reference Gaussian noise distribution (M = 0, SD = [0.3-3.0]) to each pattern noise level. In this way, virtual raters showed quasi-constant rating behaviors in response to specific subgroups of stimuli with an absolute deviation from GT scores (i.e., *noise*) between 3% to 30% as the scale range. These values were selected to simulate scenarios extending from highly trained expert agreement (SD=0.3) to layperson guessing (SD=3.0) to test the robustness of the model-aggregation under extreme conditions.

SVM models with a linear kernel and C = 1 were trained on simulated data. Optimal kernel and C-value were identified by preliminary hyperparameter-tuning. Note that since the focus of Experiment 2 is not

to compare the performance of ML models, only SVM models were trained, as they performed best in Experiment 1.

3.1.2. Sensitivity analysis

A sensitivity analysis was performed on the simulated data using the same procedure as for the real data. In the simulated data, the amount of artificial pattern noise varied across sensitivity analysis iterations, with *noise* (i.e., the SD term of the reference Gaussian noise distribution mentioned above) ranging from 3% to 30% of the rating scale range and increasing in 3% steps with each run. Thus, the sensitivity analysis had a 7 (number of raters) * 5 (folds) * 10 (noise levels) *2 (aggregation methods: mean- vs model- aggregation) structure, resulting in 700 iterations in total. In addition to the RI presented above, at each iteration Pearson's r correlation coefficient between GT scores and aggregated scores was computed (GT_cor), as a similarity measure. Note that GT_cor is only available in simulation studies where the true amount of artificial noise is known, not in real world applications where it is necessary to rely on RIs.

3.1.3. Statistical analyses: the effect of noise in ratings

Simulated data were analyzed using the same LMMs used in Experiment 1. *Noise, noise:N, debias*:*noise* and *debias*:*noise*:*N* interactions were additionally included in the model as fixed effects. *Noise* was normalized to the 0-1 range.

3.2. Results

Table 4. Mean and range of variance components, RI and performance indices computed on simulated ratings over the sensitivity analysis.

Variance components	Simulated data
σ^2_{stim}	.006 [.001 - .023]
σ^2_{exCat}	.006 [.001 - .062]
σ^2_{imCat}	.527 [.232 - .781]
σ^2_{rater}	.003 [.001 - .033]
$\sigma^2_{rater:exCat}$	.154 [.001 - .405]
$\sigma^2_{rater:imCat}$	.165 [.001 - .374]
$\sigma^2_{residual}$	.139 [.061 - .225]
RI	Simulated data
α	.809 [.127 - .967]
ICC2	.525 [.065 - .792]
ICC3	.530 [.068 - .792]
G	.818 [.416 - .967]
D	.817 [.413 - .967]
GT_cor	.901 [.684 - .982]
PI	Simulated data
R2 ~ SVM	.807 [.291 - .988]
RMSE ~ SVM	1.394 [.483 – 3.383]
MAE ~ SVM	1.118 [.396 – 2.639]

Note: PI = performance indices, GT_cor = Pearson's correlation between GT and model-aggregated scores. Variance components and R2 values are reported as a percentage of the overall model variance. The values of α, ICC2, ICC3, G, and D coefficients range from 0 to 1, with higher values indicating greater ratings' reliability. RMSE and MAE measure prediction error in the same units as the dependent variable. Level noise was assessed by σ^2_{rater} variance component while $\sigma^2_{rater:exCat1}$, $\sigma^2_{rater:exCat2}$, and $\sigma^2_{rater:imCat}$ quantified pattern noise.

Descriptive statistics for variance components, RI, and performance indices computed during the sensitivity analysis are reported in Table 4. Noise accounted for 32.2% (range = 0.1 – 69.4) of the variance in simulated ratings. In addition, the amount of noise artificially generated

in the data (i.e., from 3% to 30% of the rating scale range) was strongly correlated with the overall amount of noise detected by variance decomposition analyses, r = .985, p < .001, highlighting the effectiveness of this method in quantifying noise in ratings. GT_cor values, the correlation between GT and aggregated scores, were significantly higher for scores computed via model-aggregation (M = .941, SD = .048) than for scores computed via mean-aggregation (M = .901, SD = .070), t(349) = 8.75, p < .001.

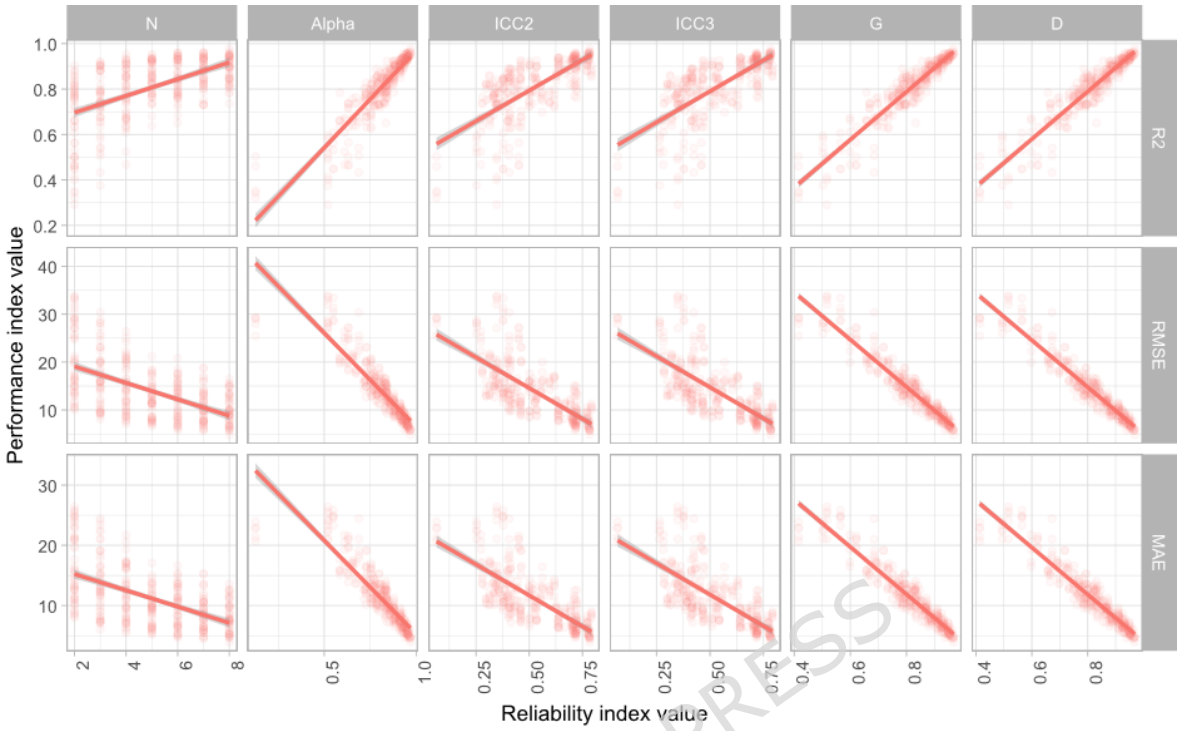
3.2.1. Association between noise, N, RI and ML model performance

Table 5. Effect estimates of N, and RI on the performance of SVM models predicting simulated data.

IV	R2	RMSE	MAE
N	1.528*	-.698*	-.688*
α	.673*	-.351*	-.343*
ICC2	.574*	-.325*	-.326*
ICC3	.571*	-.324*	-.325*
G	.701*	-.379*	-.373*
D	.703*	-.379*	-.373*
GT_cor	.691*	-.356*	-.351*

Notes: * p < .05. N was normalized to the 0–1 range. N = 0 corresponds to two raters, and N = 1 corresponds to eight raters. All other IVs were standardized across iterations. Analyses were run on simulated data across 700 sensitivity analysis iterations. 95% confidence intervals of the effects estimated by the LMM models are reported in Appendix 5.

Figure 5. Association between N, RI, and performance indices of SVM models predicting simulated data.



Note: the figure shows the impact of panel size (N) and reliability indices on the performance (R2, RMSE, MAE) of SVM models in predicting simulated data. As in Experiment 1, Alpha, G and D coefficients show a stronger correlation with model performance indices than standard ICC coefficients.

The results of the statistical analyses on simulated data are reported in Table 5, Table 6 and Figure 5. They indicated significant associations between N, RI and performance indices. N, α, G, and D seemed particularly effective in predicting RMSE, MAE, and R2 values. The association of ICC2 and ICC3 with performance indices was weaker, but in the same direction as the other RIs. GT_cor was related to RMSE, MAE, and R2 values, indicating that greater similarity between aggregated and GT scores resulted in better model performance.

3.2.2. Aggregation methods comparison

Noise led to significant decreases in all performance indices and its effect was less severe in large panel sizes, as evidenced by the significant *noise:N* interaction (see Table 6 and Figure 5). Model-aggregation was associated with significant increments in R2 and decreases in RMSE and MAE values. The significant *debias:N* interaction indicated that the effects of model-aggregation were reduced in large samples of raters, at least in terms of R2 values. Additionally, model aggregation appeared to be particularly effective in reducing prediction error in the presence of high levels of noise and small sample sizes, as evidenced by the significant effects of *debias:noise* and *debias:noise:N* on RMSE and MAE values.

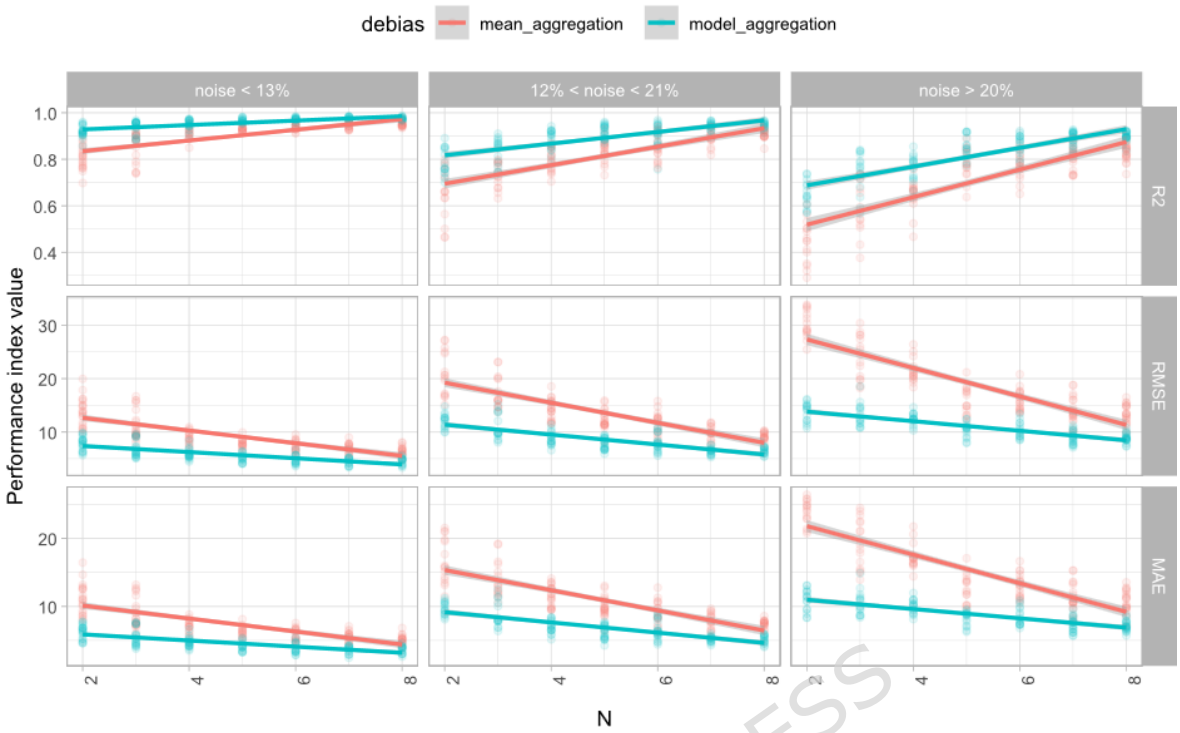
Table 6. Summary of the LMM model assessing the effect of N, noise and debias on performance indices of SVM models predicting simulated data.

IV	R2	RMSE	MAE
Intercept	.870*	2.349*	2.122*
N	1.484*	-.604*	-.600*
noise	-2.191*	1.107*	1.110*
debias	.794*	-.486*	-.482*
noise:N	.561*	-.186*	-.176*
debias:N	-.184*	.087	.082
debias:noise	-.225	-.198*	-.205*
debias:noise:N	.101	.245*	.255*

Notes: * p < .05. N was normalized to the 0–1 range. N = 0 corresponds to two raters, and N = 1 corresponds to eight raters. Debias was dummy coded as 0 = mean-aggregation and 1 = model-aggregation. Analyses were run on simulated data across 700 sensitivity analysis iterations. 95% confidence intervals of the effects estimated by the LMM models are reported in Appendix 6.

Figure 6. Association between model-aggregation and performance indices of SVM models predicting simulated data with different noise

levels.



Note: the figure shows the effect of model-aggregation over mean-aggregation on the performance (R2, RMSE, MAE) of SVM models in predicting simulated data. As in Experiment 1, model-aggregation significantly improved model performance, regardless of the amount of artificial noise in the data.

3.3. Discussion

The results of Experiment 2 highlight the negative effect of noise in human ratings on ML model performance. Great interrater variability in ratings was associated with large prediction errors (i.e., RMSE and MAE values) and reduced the amount of variance (i.e., R2 values) in pain ratings explained by the ML model. The similarity between the artificial GT scores and the score aggregates used to train ML models was a strong predictor of model performance indices, highlighting again the importance of rating reliability in the development of effective ML models. The amount of noise identified by variance decomposition

methods almost perfectly matched the amount of noise artificially generated in the data, demonstrating the effectiveness and robustness of this method for quantifying the contributions of noise to the overall variability in scores. Model-aggregation was effective in improving ML model performance, regardless of the amount of noise in the data. Nevertheless, its effects were less evident in large panel sizes. As in Passarotto et al. [5], data simulations showed that model aggregation can identify GT from individual scores, effectively accounting for interindividual variability in ratings. This suggests that the results of Experiment 1 may be due to score aggregation producing ratings closely related to GT, increasing the reliability and robustness of the ML models trained on them and improving their predictive ability.

4. General discussion

The study examined the relationship between noise in ratings and model performance in the context of supervised ML and appearance tasks. It aimed at evaluating the relationship between panel size, reliability indices, and ML model performance and to explore the use of a debiasing procedure based on REMs to improve ML model performance.

The results showed a significant relationship between panel size and ML model performance in both simulated and real data, with models trained on data from larger panels of raters performing better. RI such as α, ICC2, ICC3, G, and D were also related to model performance. Nevertheless, ICC2 and ICC3 seemed less robust, as their association with performance indices were weaker and inconsistent across datasets

and rating scales. While ICC2 and ICC3 were negatively related to RMSE and MAE values in the simulated data, the opposite trend was observed in actual data (Figure 2, and Table 2). Simulated data confirmed that artificial noise in ratings was a major cause of poor ML model performance. Moreover, the similarity between aggregated and GT scores was positively related to performance indices. Noise-robust LSSVM models showed no performance advantage over alternative models. LSSVM optimizes a least-squares cost function which, while generally robust, penalizes large errors squared. If noise consists of many consistent deviations rather than distinct, extreme outliers as in the present case, LSSVM models are unlikely to provide any substantial improvement in model performance. Debiasing procedures based on REMs lead to a significant improvement in ML model performance, increasing R2 and decreasing RMSE and MAE value, especially with small panels of raters. Their effectiveness is likely due to their ability to identify GT from individual ratings more accurately than alternative methods, as evidenced in Experiment 2 and previous studies [5].

The relationship between noise in ratings and ML model performance in the context of performance tasks is known. For instance, a recent survey identified inter-rater variability in anatomical annotation as a primary source of label noise, which significantly hampers the generalization of deep learning models [48]. This phenomenon extends to remote sensing, where Peng et al. reported that human annotation errors in satellite imagery act as instance-dependent noise [49]. The authors found that even a 1% increase in label noise can result in a 0.5% drop in

classification accuracy, closely paralleling our own findings on the sensitivity of ML to noise. Furthermore, in materials science inter-rater variability in the microstructural analysis of steel has been identified as a critical limit to automating defect detection [50]. Here, we extend these results to appearance tasks, demonstrating that noise is likely caused by the unreliable and inconsistent nature of human judgements. As expected, we showed that increasing the number of raters can compensate for the effects of noise, due to its random nature. G and D appeared to be the most robust coefficients for measuring the reliability of ratings, given their versatility in dealing with the complex noise structures found in human ratings. ICC2 and ICC3 appeared to be less flexible and consistent, showing negative and unexpected associations with model performance (i.e., high reliability scores corresponding to greater prediction error). The strength of G and D probably lies in the penalization of variance components due to noise related to the panel size (see Section 2.1.6), which is not found in the ICC formulae. Furthermore, the ICC3 formula does not take into consideration the variance to the raters (i.e., σ^2_{rater} in formulae 8-9), which can substantially bias the resulting reliability estimate by ignoring a potentially large source of measurement error. Thus, while ICC coefficients measure within-sample agreement and consistency, G and D measure absolute reliability of ratings, which is more relevant to the development of ML models. To our knowledge, this is the first study to directly investigate the relationship between reliability indices and predictive performance of ML regression algorithms.

With this study, we demonstrated the effectiveness of a compact and scalable debiasing procedure based on REMs that allowed us to assess and account for noise in ratings, thereby improving ML model performance. The procedure is theoretically sound [4,5] compared to alternative solutions that underestimate the psychometric dimension of human ratings such as removing observations or penalizing divergent raters. By incorporating hierarchical metadata, our model can take advantage of similarities between stimuli to disentangle rater bias (i.e. level noise) from bias associated with specific stimulus categories (i.e. pattern noise), which is advantageous over simple weighted averaging. Moreover, this procedure is versatile as it allows us to account for noise in different stimuli categories, based on the physical properties of the stimuli or arbitrarily defined by the researchers. Additionally, the results suggest that simple mean-aggregation may not be an effective solution for handling complex rating schemes that may reflect subjective perceptions and experiences within a given field, particularly when there are small numbers of evaluators. Rating debiasing based on REMs can enable reliable and generalizable score aggregates to be estimated with only a few sets of ratings. The proposed method demonstrates that knowledge about the data (i.e., stimuli categories) can be used to increase the reliability of ratings for ML applications, as an alternative to or in addition to increasing the size of the panel and number of evaluators.

For example, REMs can be used to exploit data from multicenter biomedical trials or materials testing consortia in order to dynamically

estimate and correct site-related offsets as random effects, as demonstrated in previous studies [51]. It can also facilitate the development of ML models that perform similarly well to those developed using ratings from larger samples of evaluators, saving researchers time and resources. Moreover, model-aggregation requires minimal computational costs, as it predominantly comprises REMs, which are relatively straightforward to implement.

According to our results, model-aggregation is recommendable in most scenarios as it consistently outperforms alternative methods such as mean-aggregation, although the greatest advantages are observed in small samples of evaluators ($N < 6$). A fundamental requirement of this approach as well as aggregation methods in general is the homogeneity of the sample of evaluators, which should provide a comprehensive and consistent representation of the population of interest. It also requires knowledge about potential sources of noise, intended as explicit stimuli categories or intrinsic patterns in the data, which are needed to model variability in ratings, estimate variance components, compute reliability indices and aggregate data. Although model-aggregation can provide great advantages in small samples of evaluators, low sparsity in the data and possibly fully crossed rating designs are needed to properly estimate individual rating schema. This allows raters to evaluate stimuli naturally without requiring additional training beyond instructions on how to use the assessment instrument correctly. Ultimately, repeatedly calculating reliability indices and applying predictive methods (e.g., G study [13])

can guide data collection and help determine whether additional evaluators are required to achieve reliable score aggregates.

This study comes with some limitations. The proposed method requires two or more sets of ratings with substantial overlap between them in order to estimate and account for noise. In sparse crowdsourcing scenarios where each rater evaluates only a single unique image, model-aggregation cannot effectively disentangle rater bias from stimulus signal. Thus, it may not be applicable to massive databases, where extensive nested rating procedures are required. ML classification models were not considered in this study due to three main reasons: first, we assumed that pain was a continuous rather than a categorical variable, in line with the literature [30]. Second, the number of stimuli was not sufficient to treat VAS and COMFORTneo as ordinal scales, as this would have led to undersized subsets of ratings (i.e., less than 50 observations per subgroup). Third, we preferred not to dichotomize COMFORTneo and VAS scores into high-low pain classes based on arbitrary thresholds. The choice of ML models tested in the study was limited by the computational load required by the sensitivity analyses. Future studies should investigate how generalizable the reported results are to different statistical models, particularly deep learning architectures trained on large-scale datasets. The study did not consider variability due to feature extraction and image preprocessing approaches. However, this is unlikely to have biased the reported results, given that the same preprocessing pipeline was used throughout the analyses.

The data simulations assumed that GT scores in appearance tasks can be estimated by averaging data from large panels of raters, in line with Kahneman et al. [4]. While this assumption might hold true within well-defined populations, with similar demographic and cultural backgrounds, the generalizability of GT across heterogeneous populations might be limited, as appearance tasks strongly depend on subjective perception and experience in a domain. Thus, multiple sets of reliable and equally valid GT scores for the same stimuli may be found, allowing for the development of different ML models. Future studies may compare debiasing procedures based on REMs to alternative methods and investigate their effectiveness in the presence of varying degrees of data sparsity. Noise was simulated using Gaussian distributions, which oversimplified the heteroscedastic structure of human error. Exploring skewed or scale-dependent noise distributions is an area of future research. Finally, Experiment 1 considered a case scenario involving limited disagreement between evaluators (ICC >.8), and its results may not generalize to noisier evaluation contexts.

In conclusion, our results highlight the importance of ratings' reliability in ML model development. This aspect is often underestimated and given secondary importance compared to features and ML algorithms selection. Given their relevance to ML model performance, researchers are encouraged to regularly report detailed information about the rating procedure and ratings' reliability, providing standardized reliability metrics. Rating aggregation based on REMs can be a viable solution to improve rating reliability and ML model performance, especially when

data from only a limited number of raters are available. The proposed framework can be used in any field that relies on human annotation to reduce annotation costs and improve the robustness of automated evaluation systems.

5. References

1. Ashton, R.H. A Review and Analysis of Research on the Test–Retest Reliability of Professional Judgment. *J. Behav. Decis. Mak.* **2000**, *13*, 277–294, doi:10.1002/1099-0771(200007/09)13:3%3C277::AID-BDM350%3E3.0.CO;2-B.
2. Kruglanski, A.W.; Ajzen, I. Bias and Error in Human Judgment. *Eur. J. Soc. Psychol.* **1983**, *13*, 1–44, doi:10.1002/ejsp.2420130102.
3. Amidei, J.; Piwek, P.; Willis, A. Agreement Is Overrated: A Plea for Correlation to Assess Human Evaluation Reliability.; Tokyo, Japan, 2019.
4. Kahneman, D.; Sibony, O.; Sunstein, C.R. *Noise: A flaw in human judgment;* 1st edition.; William Collins: London, 2021; ISBN 978-0-00-830900-8.
5. Passarotto, E.; Altenmüller, E.; Müllensiefen, D. Music Performance Assessment: Noise in Judgments and Reliability of Measurements. *Psychol. Aesthet. Creat. Arts* **2023**, doi:10.1037/ACA0000574.
6. Franco, G. Agreement of Medical Decisions in Occupational Health as a Quality Requirement. *Int. Arch. Occup. Environ. Health* **2006**, *79*, 607–611, doi:10.1007/s00420-006-0084-9.
7. Kingdom, F.A.A.; Prins, N. *Psychophysics: A Practical Introduction*; Academic Press, 2016; ISBN 978-0-08-099381-2.
8. Houben, S.; Stallkamp, J.; Salmen, J.; Schlipsing, M.; Igel, C. Detection of Traffic Signs in Real-World Images: The German Traffic Sign Detection Benchmark. In Proceedings of the The 2013 International Joint Conference on Neural Networks (IJCNN); August 2013; pp. 1–8.
9. Stallkamp, J.; Schlipsing, M.; Salmen, J.; Igel, C. Man vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition. *Neural Netw.* **2012**, *32*, 323–332, doi:10.1016/j.neunet.2012.02.016.
10. Artstein, R.; Poesio, M. Inter-Coder Agreement for Computational Linguistics. *Comput. Linguist.* **2008**, *34*, 555–596, doi:10.1162/coli.07-034-R2.
11. Hallgren, K.A. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor. Quant. Methods Psychol.* **2012**, *8*, 23–34, doi:10.20982/tqmp.08.1.p023.

12. Kim, D.Y.; Wallraven, C. Label Quality in AffectNet: Results of Crowd-Based Re-Annotation Available online: https://arxiv.org/abs/2110.04476v1 (accessed on 20 October 2025).
13. Brennan, R.L. *Generalizability Theory*; Springer Science & Business Media, 2001; ISBN 978-0-387-95282-6.
14. Chernoff, H. Large-Sample Theory: Parametric Case. *Ann. Math. Stat.* **1956**, *27*, 1–22.
15. Navajas, J.; Niella, T.; Garbulsky, G.; Bahrami, B.; Sigman, M. Aggregated Knowledge from a Small Number of Debates Outperforms the Wisdom of Large Crowds. *Nat. Hum. Behav.* **2018**, *2*, 126–132, doi:10.1038/s41562-017-0273-4.
16. Zhu, X.; Wu, X. Class Noise vs. Attribute Noise: A Quantitative Study. *Artif. Intell. Rev.* **2004**, *22*, 177–210, doi:10.1007/s10462-004-0751-8.
17. Frenay, B.; Verleysen, M. Classification in the Presence of Label Noise: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 845–869, doi:10.1109/TNNLS.2013.2292894.
18. Brodley, C.E.; Friedl, M.A. Identifying Mislabeled Training Data. *J. Artif. Intell. Res.* **1999**, *11*, 131–167, doi:10.1613/jair.606.
19. Raykar, V.C.; Yu, S.; Zhao, L.H.; Valadez, G.H.; Florin, C.; Bogoni, L.; Moy, L. Learning From Crowds. *J. Mach. Learn. Res.* **2010**, *11*, 1297–1322.
20. Song, H.; Kim, M.; Park, D.; Shin, Y.; Lee, J.-G. Learning from Noisy Labels with Deep Neural Networks: A Survey 2022.
21. Beltramini, A.; Milojevic, K.; Pateron, D. Pain Assessment in Newborns, Infants, and Children. *Pediatr. Ann.* **2017**, *46*, e387–e395, doi:10.3928/19382359-20170921-03.
22. Duhn, L.J.; Medves, J.M. A SYSTEMATIC INTEGRATIVE REVIEW OF INFANT PAIN ASSESSMENT TOOLS. *Adv. Neonatal Care* **2004**, *4*, 126, doi:10.1016/j.adnc.2004.04.005.
23. Eriksson, M.; Campbell-Yeo, M. Assessment of Pain in Newborn Infants. *Semin. Fetal. Neonatal Med.* **2019**, *24*, 101003, doi:10.1016/j.siny.2019.04.003.
24. Gkikas, S.; Tsiknakis, M. Automatic Assessment of Pain Based on Deep Learning Methods: A Systematic Review. *Comput. Methods Programs Biomed.* **2023**, *231*, 107365, doi:10.1016/j.cmpb.2023.107365.
25. Zamzmi, G.; Goldgof, D.; Kasturi, R.; Sun, Y.; Ashmeade, T. Machine-Based Multimodal Pain Assessment Tool for Infants: A Review 2019.
26. Webb, R.; Ayers, S.; Endress, A. The City Infant Faces Database: A Validated Set of Infant Facial Expressions. *Behav. Res. Methods* **2018**, *50*, 151–159, doi:10.3758/s13428-017-0859-9.
27. Brahnam, S.; Chuang, C.-F.; Shih, F.Y.; Slack, M.R. Machine Recognition and Representation of Neonatal Facial Displays of Acute Pain. *Artif. Intell. Med.* **2006**, *36*, 211–222, doi:10.1016/j.artmed.2004.12.003.
28. Bergamasco, L.; Lattanzi, M.; Gavelli, M.; Pastrone, C.; Olmo, G.; Borsotti, L.; Parodi, E. Pain Assessment in Neonatal Clinical Practice via Facial Expression Analysis and Deep Learning. In Proceedings of

the Bioinformatics and Biomedical Engineering; Rojas, I., Ortuño, F., Rojas, F., Herrera, L.J., Valenzuela, O., Eds.; Springer Nature Switzerland: Cham, 2024; pp. 249–263.
29. Martínez, A.; Pujol, F.A.; Mora, H. Application of Texture Descriptors to Facial Emotion Recognition in Infants. *Appl. Sci.* **2020**, *10*, 1115, doi:10.3390/app10031115.
30. van Dijk, M.; Roofthooft, D.W.E.; Anand, K.J.S.; Guldemond, F.; de Graaf, J.; Simons, S.; de Jager, Y.; van Goudoever, J.B.; Tibboel, D. Taking up the Challenge of Measuring Prolonged Pain in (Premature) Neonates: The COMFORTneo Scale Seems Promising. *Clin. J. Pain* **2009**, *25*, 607–616, doi:10.1097/AJP.0b013e3181a5b52a.
31. Jiang, L.; Wu, M.; Fu, W.; Wang, Y.; Gu, Y.; Zhang, F.; Gong, W.; Qin, Y.; Xu, Y.; Feng, R.; et al. Construction and Validation of a Pain Facial Expressions Dataset for Critically Ill Children. *Sci. Rep.* **2025**, *15*, 17214, doi:10.1038/s41598-025-02247-w.
32. Haines, N.; Southward, M.W.; Cheavens, J.S.; Beauchaine, T.; Ahn, W.-Y. Using Computer-Vision and Machine Learning to Automate Facial Coding of Positive and Negative Affect Intensity. *PLOS ONE* **2019**, *14*, e0211735, doi:10.1371/journal.pone.0211735.
33. Haines, N.; Bell, Z.; Crowell, S.; Hahn, H.; Kamara, D.; McDonough-Caplan, H.; Shader, T.; Beauchaine, T.P. Using Automated Computer Vision and Machine Learning to Code Facial Expressions of Affect and Arousal: Implications for Emotion Dysregulation Research. *Dev. Psychopathol.* **2019**, *31*, 871–886, doi:10.1017/S0954579419000312.
34. Fang, J.; Wu, W.; Liu, J.; Zhang, S. Deep Learning–Guided Postoperative Pain Assessment in Children. *PAIN* **2023**, *164*, 2029, doi:10.1097/j.pain.0000000000002900.
35. Romeo, L.; Olugbade, T.; Pontil, M.; Bianchi-Berthouze, N. Multi-Rater Consensus Learning for Modeling Multiple Sparse Ratings of Affective Behaviour. *IEEE Trans. Affect. Comput.* **2024**, *15*, 859–871, doi:10.1109/TAFFC.2023.3297270.
36. Inc, T.M. MATLAB Version: 9.13.0 (R2022b) 2022.
37. Martínez, A.; Pujol, F.A.; Mora, H. Application of Texture Descriptors to Facial Emotion Recognition in Infants. *Appl. Sci.* **2020**, *10*, 1115, doi:10.3390/app10031115.
38. Celona, L.; Manoni, L. Neonatal Facial Pain Assessment Combining Hand-Crafted and Deep Features. In Proceedings of the New Trends in Image Analysis and Processing – ICIAP 2017; Battiato, S., Farinella, G.M., Leo, M., Gallo, G., Eds.; Springer International Publishing: Cham, 2017; pp. 197–204.
39. Nanni, L.; Lumini, A.; Brahnam, S. Local Binary Patterns Variants as Texture Descriptors for Medical Image Analysis. *Artif. Intell. Med.* **2010**, *49*, 117–125, doi:10.1016/j.artmed.2010.02.006.
40. Naufal Mansor, M.; Rejab, M.N. A Computational Model of the Infant Pain Impressions with Gaussian and Nearest Mean Classifier. In Proceedings of the 2013 IEEE International Conference on Control System, Computing and Engineering; November 2013; pp. 249–253.
41. Gelman, A.; Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*; Analytical Methods for Social

Research; Cambridge University Press: Cambridge, 2006; ISBN 978-0-521-86706-1.
42. Bartko, J.J. The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychol. Rep.* **1966**, *19*, 3–11, doi:10.2466/pr0.1966.19.1.3.
43. Kuhn; Max Building Predictive Models in R Using the Caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26, doi:10.18637/jss.v028.i05.
44. Sorber, L. Lsorber/Neo-Ls-Svm 2024.
45. RStudio Team *RStudio: Integrated Development Environment for R*; RStudio, PBC.: Boston, MA, 2020;
46. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *ArXiv E-Prints* **2014**, *arXiv:1406*, doi:10.18637/jss.v067.i01.
47. Revelle, W. Psych: Procedures for Psychological, Psychometric, and Personality Research 2024.
48. Shi, J.; Zhang, K.; Guo, C.; Yang, Y.; Xu, Y.; Wu, J. A Survey of Label-Noise Deep Learning for Medical Image Analysis. *Medical Image Analysis* **2024**, 95, 103166, doi:10.1016/j.media.2024.103166.
49. Peng, L.; Wei, T.; Chen, X.; Chen, X.; Sun, R.; Wan, L.; Chen, J.; Zhu, X. Human-Annotated Label Noise and Their Impact on ConvNets for Remote Sensing Image Scene Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2025**, 18, 1500–1514, doi:10.1109/JSTARS.2024.3502461.
50. Durmaz, A.R.; Potu, S.T.; Romich, D.; Möller, J.J.; Nützel, R. Microstructure Quality Control of Steels Using Deep Learning. *Front. Mater.* **2023**, 10, doi:10.3389/fmats.2023.1222456.
51. Kahan, B.C.; Harhay, M.O. Many Multicenter Trials Had Few Events per Center, Requiring Analysis via Random-Effects Models or GEEs. *J Clin Epidemiol* **2015**, 68, 1504–1511, doi:10.1016/j.jclinepi.2015.03.016.