scientific reports

Check for updates

OPEN

Comparison between AI and human expert performance in acute pain assessment in sheep

Marcelo Feighelstein[1]✉, Stelio P. Luna[2], Nuno O. Silva[2], Pedro E. Trindade[3], Ilan Shimshoni[1], Dirk van der Linden[4] & Anna Zamansky[1]✉

This study explores the question whether Artificial Intelligence (AI) can outperform human experts in animal pain recognition using sheep as a case study. It uses a dataset of N = 48 sheep undergoing surgery with video recordings taken before (no pain) and after (pain) surgery. Four veterinary experts used two types of pain scoring scales: the sheep facial expression scale (SFPES) and the Unesp-Botucatu composite behavioral scale (USAPS), which is the 'golden standard' in sheep pain assessment. The developed AI pipeline based on CLIP encoder significantly outperformed human facial scoring (AUC difference = 0.115, $p < 0.001$) when having access to the same visual information (front and lateral face images). It further effectively equaled human USAPS behavioral scoring (AUC difference = 0.027, $p = 0.163$), but the small improvement was not statistically significant. The fact that the machine can outperform human experts in recognizing pain in sheep when exposed to the same visual information has significant implications for clinical practice, which warrant further scientific discussion.

> *"Deep Blue was intelligent the way your programmable alarm clock is intelligent. Not that losing to a 10$ million alarm clock made me feel any better."*
>
> Garry Kasparov, 1997.

The use of artificial intelligence (AI) in healthcare by utilizing machine learning (ML) algorithms and data analysis techniques is a real game-changer, resulting in better patient outcomes, better use of resources, and lower operating costs[1,2]. In pain assessment, AI can play an important role in automated non-invasive analysis of behavioral parameters, such as facial expressions and body language. Unsurprisingly, in recent years an increasing amount of works have addressed automation of pain assessment in infants (see Zamzmi et al.[3] for a review). Only recently the first AI-based mobile app for pain assessment in non-verbal patients, PainChek, based solely on facial expression analysis was released[4,5]. Earlier this year it was evaluated for the first time in the context of procedural pain assessment and monitoring in clinical practice, demonstrating high accuracy (area under the curve 0.964 and 0.966, respectively), and precision above 0.89[6].

The interest in automated approaches for animal pain recognition has also drastically increased in recent years. Broome et al.[7] provides a review of more than twenty studies addressing video-based automated recognition of affect and pain in animals, with the majority of works focusing on the latter. Automated pain assessment and recognition has been investigated mostly for rodents[8,9], horses[10–12], and most recently cats[13,14], rabbits[15] and dogs[16]. Sheep have also been addressed in this context, see, e.g.,[17–19], however accuracy reached was quite low (around 67%), partially due to the challenging nature of data collected in farm settings.

Since pain is an internal state that is difficult to measure, the establishment of ground truth is a major challenge in pain research. In the human domain, self-reporting is considered one of the most unobtrusive and non-invasive methods for establishing ground truth in pain[20] and emotion research[21]. However, in humans not able to verbally communicate their pain, and in animals, a ground truth of the pain experience is lacking.

Behavior scoring by human experts is the most common approach for pain assessment in animals.

The first animal grimace scales were developed for rodents and they are now available for many mammalian species[22], including rats[23], rabbits[24], horses[25], pigs[26], ferrets[27], sheep[28,29] and cats[30,31]. Numerous instruments

[1]Department of Information Systems, University of Haifa, Haifa, Israel. [2]School of Veterinary Medicine and Animal Science, Sao Paolo State University (Unesp), São Paulo, Brazil. [3]Department of Population Pathobiology, North Carolina State University, Raleigh, USA. [4]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK. ✉email: feighels@gmail.com; annazam@is.haifa.ac.il; annazam@gmail.com

based on behavior have also been validated for domestic species, like cats[32], dogs[33], rabbits[34], pigs[35], goats[36], sheep[37], horses[25] , donkeys[38] and cattle[39].

Yet even validated pain assessment methods are limited by the observer's previous training and ability to interpret the pain responses accurately[40], as well as by observers' various biases, like gender, fatigue, experience and time consumption[41,42]. Adami et al.[43] recently evaluated the inter-observer reliability of three feline pain scales commonly used in clinical practice: the Glasgow Feline Composite Measure Pain Scale (CMPS-Feline[44]), the Colorado State University Feline Acute Pain Scale (CSU-FAPS[45]) and the Feline Grimace Scale (FGS[31]). The reliability was found to range in most cases from poor to fair/moderate, suggesting that subjectivity is a considerable limitation of these tools specifically designed to quantify pain in cats. It is thus important to highlight the intrinsic subjectivity of such methods, as well as their potential variability of outcome between assessors with different backgrounds and level of expertise. Many veterinarians acknowledge difficulties recognizing pain[46] and consider their knowledge in assessment and treatment of pain inadequate[47]. Difficulties with pain assessment is also conceived a significant barrier of veterinarians to adequate treatment of chronic pain[48]. Due to the inherent limitations of a subjective manual scoring, there is no question that human scoring methods can be digitally enhanced to be less susceptible to human error, subjectivity and bias. The question is, therefore: is automated pain assessment mature enough to be a game changer in the domain of animal pain assessment? And can machines outperform human experts in this task?

When answering these questions, the devil is, of course, in the details. No matter how we measure performance, we need to have a way to objectively establish ground truth, which does not rely on human scoring, which in itself is being scrutinized. As highlighted in[7], this is standardly achieved in strict experimental conditions, where pain is either induced or timed (using time moments, e.g., before and after surgical procedures). The former can refer to experimental induction of clinical short term reversible moderate pain using models known from human volunteers. In[11], e.g., two ethically regulated methods for experimental pain induction were used: a blood pressure cuff placed around one of the forelimbs of horses, or the application of capsaicin (chili extract) on the skin of the horse. The latter refers to timing data collection before and after a clinical procedure. For instance, in[13] videos of female cats undergoing ovariohysterectomy were recorded at different time points pre- and post-surgery.

Due to the obvious difficulties in collecting such data in the context of animal pain, datasets collected in strict experimental settings are extremely scarce. The dataset collected in the study for validating the Unesp-Botucatu composite scale (USAPS) to assess acute postoperative abdominal pain in sheep and defining a cut-off point for analgesic intervention[37] presents an interesting opportunity in this regard, which we explore in this paper.

Another important issue when comparing performance of human vs. machine is whether both are exposed to the same visual information in the same way. The SFPES-based human scoring uses two facial images: front and lateral view, and this is also the input to our AI model. The behavioral USAPS scale uses a video of the animal, and thus it could be the case that human experts may have more visual nformation of the body language than the machine.

The dispute between machine and humans in the task of sheep pain recognition is now all set. We, therefore, investigate the question: can a machine outperform human experts in sheep pain recognition. More precisely, we hypothesize that a machine learning algorithm can outperform human experts in sheep pain recognition when the latter is measured e.g., using the SFPES scoring (using the appropriate cut-off point). The developed algorithm uses a deep learning pipeline, which uses a CLIP encoder for feature extraction and a Naive Base classification model for pain recognition.

Methods

The dataset

The dataset used in this study was collected in a previous study validating the Unesp-Botucatu composite scale to assess acute postoperative abdominal pain in sheep[37]. The study was approved by the Ethics Committee on Animal Use from the School of Veterinary Medicine and Animal Science, São Paulo State University (Unesp), Botucatu, São Paulo, Brazil, under protocol 0027/2017 and followed the recommendations of ARRIVE[49], adapted to the experimental design. Details about housing, management, anesthetic, surgical and analgesic procedures can be found in the previous study[37]. We understand that database reuse for new analysis contributes to the four Rs of animal experimentation (reduce, replace, refine, and responsibility)[50,51].

The dataset is composed of video recordings of 48 sheep (*Ovis Aires*) of three breeds (17 Bergamacia, 18 Lacaune, and 13 Dorper). The animals were submitted to abdominal surgery[52] and video recordings were taken one hour before surgery (M1) and at the predicted time of greatest pain, between three and four hours after the end of surgery (M2). In addition, frontal and lateral photographs of the sheep faces were taken at the same time points. In the original study these videos and photos were randomly and blindly analyzed twice by four observers within one-month interval to calculate repeatability. We constructed the full image dataset for our study containing a total of 96 images with frontal and lateral facial images (48 sheep x 2 stages x 2 sides): 96 'pain' (48 lateral, 48 frontal) and 96 'no pain' (48 lateral, 48 frontal).

These images were divided into two classes: No Pain (stage M1; before surgery) and Pain (stage M2; after surgery).

The reduced dataset. Establishing the 'ground truth' using time points may be insufficient for making sure at time point M1 sheep do not experience pain, while at M2 they do experience it, and thus this 'ground truth' may not be accurate and may impact the measured performance of both humans and machine. To further investigate this issue, we created a reduced dataset integrating USAPS measurement into 'ground truth' establishment. More specifically, we removed 4 samples having an average score of all observers indicating "No Pain" label ($USAPS < 4$) at time point M2 (after surgery) and 5 samples having an average score of all observers indicating

Fig. 1. Example of frontal images: Sheep 1: no pain; pain; Sheep 17: no pain; pain.

Fig. 2. Example of lateral images: Sheep 1: no pain; pain; Sheep 17: no pain; pain.

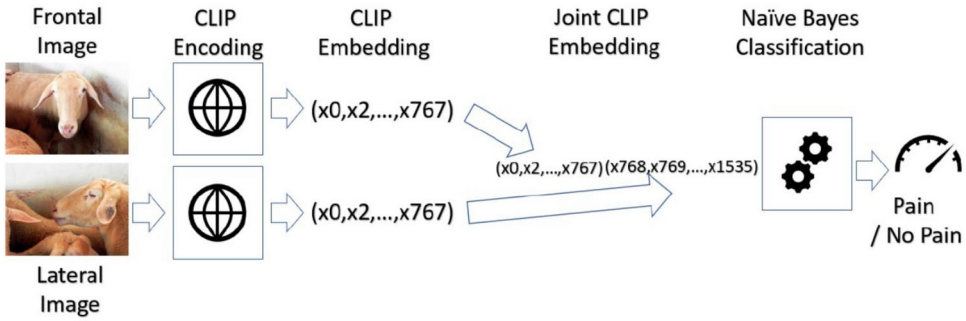

Fig. 3. Pipeline description.

"Pain" ($USAPS \geq 4$) at time point M1 (before surgery). Overall, removing 9 samples, we remained with a reduced dataset of N'=39 individuals.

Examples of frontal and lateral images are shown in Figs. 1 and 2.

Pain recognition by human experts

Our ground truth was established by the timepoint labels of the images: M2 (class Pain) and M1 (class No Pain), to which both human and machine scoring were compared using metrics described below. The first human scoring method was based on the sheep pain facial expression scale developed in[53]. It uses both frontal and lateral face images and scores five facial areas using a three-point scale (0 = not present, 1 = partially present, 2 = present): orbital tightness, cheek tightness, ear position, lip and jaw profile, and nostril and philtrum position. A total pain score is determined by adding the individual scores for each of the five areas for each set of photographs, with the maximum possible score being 12 (i.e. a score of 2 for each of the facial areas and lateral and frontal view of ear position). The calculation of the Youden index which is the intersection point of simultaneous greatest sensitivity and specificity (sensitivity + specificity -1) determined by the receiver operating characteristic curve[37] led to the definition of the Cut-Off point for analgesia as 4 (this result had not been published until now).

The second human scoring method was the USAPS based on body behavioral scoring. The USAPS was validated in[37] to assess acute postoperative abdominal pain in sheep. The USAPS items refer to interaction, activity, locomotion, appetite, head position, and posture, and each of them is scored on a scale between 0 and 2, with a maximal overall score of 12; The above mentioned Cut-Off point 4 for analgesia was used for establishing the Pain and No Pain classes. Four independent experts performed both of the scoring tasks, reaching above moderate inter-observer reliability (≥ 0.53)[37], each expert repeating each scoring two times (phases). A total of 768 observations were collected (48 sheep x 2 classes (pain or no pain) x 4 observers x 2 phases). For moving from scoring to recognition (class Pain/No Pain), the scores were then calculated using the appropriate cut-off point (≥ 4 for USAPS and SPFES) on each score. Calculations were also performed for the USAPS cut-off point ≥ 5 to avoid the diagnostic uncertainty zone as indicates sheep truly suffering pain (true positives). To summarize, the way we obtain the two human scores to which we refer as USAPS and SPFES is by (i) aggregation of experts scoring each image (on a scale 0-12), (ii) transforming to pain/no pain (binary score) using appropriate cut-off points.

Method	Accuracy	Recall	Precision	F1	Sensitivity	Specificity
USAPS Cut-Off 4	0.7956	0.8776	0.7539	0.8111	0.8776	0.7135
USAPS Cut-Off 5	0.8177	0.8411	0.8034	0.8219	0.8411	0.7943
SPFES Cut-Off 4	0.7083	0.8672	0.6581	0.7483	0.8672	0.5495
Machine	0.8229	0.8125	0.8298	0.8211	0.8125	0.8333

Table 1. Machine performance and its comparison to humans.

	ML	$USAPS_4$	$USAPS_5$	SFPES
AUC	0.823	0.796	0.818	0.708

Table 2. AUCs Comparison; ML is the machine learning algorithm. $USAPS_4$ and $USAPS_5$ is the Unesp-Botucatu Sheep Acute Pain Scale using Cut-Off points 4 and 5 respectively; SPFES is the Sheep Pain Facial Expression Scale.

Method	Accuracy	Recall	Precision	F1	Sensitivity	Specificity
USAPS	0.8365	0.9199	0.7884	0.8491	0.9199	0.7532
SPFES	0.7276	0.9038	0.6682	0.7684	0.9038	0.5512
Machine	0.7949	0.8462	0.7674	0.8049	0.8462	0.7436

Table 3. Comparison using the reduced dataset (using USAPS Cut-Off Point 4).

Pain recognition by machine

An AI pipeline consisting of two components was developed for automated pain recognition. The pipeline is depicted in Fig. 3. It uses a CLIP encoder for feature extraction of both frontal and lateral facial images of sheep on a certain pain state and the Naive Bayes classifier[54] for pain recognition.

The CLIP[55] encoding is a process of mapping images into a high-dimensional embedding space, where each image is represented by a unique embedding vector. The CLIP encoder achieves this by pre-training a neural network on a large dataset of image and text pairs using a contrastive loss function.

Once obtained the CLIP 768-dimensional embedding vectors of the frontal and lateral facial images of a sheep, we concatenate them into a single 1536-dimensional vector representing the embedding of both images.

The Naive Bayes classification model[54] is a probabilistic algorithm used for classification tasks in machine learning, which is computationally efficient and can work well even with small amounts of training data.

We evaluate the performance of the classification model using leave-one-animal-out cross-validation with no animal overlap. Due to the relatively low numbers of sheep (N = 48) and of image samples (n=48 × 2 classes × 2 sides) in the dataset, this method is appropriate[8,11]. By separating the images of individuals used for training and testing respectively, we enforce generalization to unseen subjects and ensure that no specific features of an individual are used for classification.

In the training process we used feature selection[56] to improve the classification performance by reducing the dimensionality of the input space and eliminating redundant or irrelevant features that may cause overfitting or increase the computational complexity of the model.

Performance metrics

We evaluate the ML pipeline performance (and compare it to human) using standard metrics commonly used in the literature: accuracy, precision, recall, F1, sensitivity and specificity[13].

Statistical analysis

For a statistical analysis of the performance, we compared areas under the receiver operating characteristic curve (AUCs) with DeLong test[57]. The AUC represents an index to evaluate the classification performance, that varies from 0 to 100. Accuracy is considered low when values are between 0.50 and 0.70, moderate between 0.70 and 0.90 and high when above 0.90. Data were analyzed using Jamovi software (https://www.jamovi.org; version 2.3.28.0; Jamovi project (2023)), using Test ROC from the psychoPDA package (version 1.0.5).

A Shapiro-wilk test of normality indicated all four considered data distributions were not normally distributed (Shapiro-wilk W=0.59, 0.63, 0.64, 0.64 resp. with a $p<0.001$).

Results

Table 1 presents the performance metrics of the machine vs. human scoring based on USAPS and SPFES. The machine outperformed human scoring in terms of accuracy, precision, specificity and F1.

Table 2 presents the AUC comparison between the machine and the two human scoring methods. Pair-wise comparisons indicated that the machine significantly outperforms SPFES (AUC difference = 0.115, $p<0.001$).

The machine further effectively equals both USAPS (Cut-Off 4) (AUC difference= 0.027, p=0.163), and USAPS (Cut-Off 5) (AUC difference = 0.005, p=0.787), but the small improvement was not statistically significant.

Table 3 presents the results of the comparison using the reduced dataset with USAPS Cut-Off point 4. In terms of accuracy, we see a small drop in machine performance, and a larger drop in human SPFES performance, with machine still outperforming SPFES in terms of accuracy and F1.

Discussion

The answer to our question whether machine outperforms human experts in recognizing pain in sheep when being exposed to the same visual information was affirmative. The improvement of the machine over facial scoring (SPFES) was found significant, showing a better diagnostic performance. Moreover, the machine was higher than both methods of human scoring (USAPS and SFPES) in accuracy, precision, recall, specificity and sensitivity.

The problem of automation of sheep pain recognition has already been addressed[17,19,58] with the aim to automate the SPFES scale. The pipeline presented in[19] automatically recognizes facial action units and uses them to predict pain level.

The approach for automating sheep pain recognition taken in[17,19,58] automate the SPFES scale, using landmarks to localize facial regions of interest, and then extracts histograms of oriented gradients features from these regions, applying a support vector machine (SVM) model to assess the facial action units. Their pipeline reached an overall accuracy of just 67%, whereas the accuracy of our model is above 82%. Thus the AI pipeline presented in this study significantly outperforms existing AI solutions for sheep pain recognition. The reason behind this finding is probably related to the limitation for human detection of some facial action units and that SPFES exhibited only moderate level of evidence (based on methodological quality, number of studies, and studies' findings) in a recent systematic review[59]. Unsurprisingly the human assessment SPFES results were the worse in the current study. However, perhaps a more important contribution of our study is presenting a framework where performance of human scoring can be evaluated against machine scoring: using the same data, and based on the same visual input. Measuring the performance in this framework using the AUC metric, the machine outperforms human experts using both USAPS and SPFES in pain recognition.

The 'ground truth' used in this framework are the time points before and after surgery, which are used for the definition of the classes No Pain/Pain respectively for measuring pain recognition performance. However, pain is an individual-based sensation and unlike with humans, we cannot easily communicate with animals. Therefore, the behavioral changes are apparently the best way to diagnose clinical pain in animals[34,36–38].

According to the above one may claim the use of time points may be insufficient for making sure sheep do not experience pain before surgery, while after surgery they experience it, and thus this 'ground truth' may not be accurate and may impact the measured performance of both humans and the machine. Our experiment with the reduced dataset of N=33 individuals was performed to investigate this issue. Table 3 presents the results, showing a small drop in accuracy in the machine performance, with a larger drop in accuracy of human performance, with the machine still far outperforming human facial scoring. Thus our conclusion that assessing pain using facial expression was more accurate with AI than with human estimation remains valid under these new, stricter conditions. The question of what the machine is detecting in facial pain expressions beyond what humans see is still open. It is probably beyond the action units, as results of the machine by using facial units was not so promising[58].

Another aspect of the AI model presented here is that it uses two images - both front and side. However, this was imposed by our aim to match the visual information presented to the human when scoring with SFPES. Therefore, we also ran experiments with just one side, reaching accuracy of above 70% with frontal view, and a slightly lower performance (67%) with lateral view.

The importance of front and lateral views for both machine and human assessments can be explaind by the fact that only in the lateral view cheek muscle tightening and abnormal lip and jaw profile can be viewed, while only the frontal view allows the observation of abnormal nostril and muzzle shape. Either view provides information on orbital tightening, and both views are probably necessary to assess abnormal ear position. That explains why the last item was assessed in both views by humans and total maximum score was 12.

It should be noted that while the machine outperforms humans when humans use SFPES, the latter is not the 'golden standard' in the field of sheep pain assessment[59]. The dependence on good quality images with two views is one of the most important limitations for this method and pros and cons of in-person or remote automated monitoring have been previously addressed[34]. USAPS uses body behavioral information and is considered a more accurate method than SFPES[60]. Although comparing the machine to USAPS may not be fair, as the machine only has access to frontal and side images, while a human using USAPS observes the animal's behavior over a period of time, Table 1 still shows that the machine outperforms human experts also in this case, although the improvement was not found significant. This indicates a great potential for the development of future AI pipelines looking at behavior and including the temporal dimension. Our recent study on rabbit pain[15] is a first step in this direction.

Another important point to address is that only the extreme time points (no pain and possible intense pain) were assessed by machine, therefore it is necessary to include other time points (after analgesia and 24h after surgery) as performed in the behavioral study[37], to check if machine does well in diagnosing mild and moderate pain as well. A more systematic investigation of explainability of the obtained models along the lines of[14] is an additional immediate future direction. This type of investigation can provide further insights into the specific facial features utilized by the models to detect pain and potentially enhance human methods of pain recognition in sheep. A pragmatical challenge for future research and development is to include our findings into an application capable of automatic recognition of pain in animals like the application available for human

assessment of animal pain body behavior in all domestic species (Vetpain) and the Feline Grimace Scale (https://www.felinegrimacescale.com).

The implications of the findings of this study may leave many veterinarians speechless, as, like Garry Kasparov in 1997, they may be about to face their own 'Deep Blue moment'. It is too early to say that, and much more research is needed with more data and exploring other models and architectures. Also, novel and more accurate pain assessment instruments may be developed in the future. However, we need to be mindful of how slow the process of a scientific validation of such instruments is. The pace of AI development is significantly higher, compelling us to proclaim (with caution): "Human Experts, Make Way for AI!"

Data availability

The data used in this study is available upon request from the corresponding author.

Received: 25 October 2023; Accepted: 18 December 2024
Published online: 03 January 2025

References

1. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Fut. Healthc. J.* **6**(2), 94 (2019).
2. Bajwa, J., Munir, U., Nori, A. & Williams, B. Artificial intelligence in healthcare: Transforming the practice of medicine. *Fut. Healthc. J.* **8**(2), 188 (2021).
3. Zamzmi, G. et al. A review of automated pain assessment in infants: Features, classification tasks, and databases. *IEEE Rev. Biomed. Eng.* **11**, 77–96 (2017).
4. Atee, M., Hoti, K. & Hughes, J. Painchek™ use in clinical practice: An artificial intelligence (AI) assisted-pain assessment tool for aged care residents with dementia. In: 17th IASP World Congress on Pain 2018 (2018).
5. Hoti, K., Chivers, P. T. & Hughes, J. D. Assessing procedural pain in infants: A feasibility study evaluating a point-of-care mobile solution based on automated facial analysis. *The Lancet Digital Health* **3**(10), 623–634 (2021).
6. Hughes, J. D., Chivers, P. & Hoti, K. The clinical suitability of an artificial intelligence-enabled pain assessment tool for use in infants: Feasibility and usability evaluation study. *J. Med. Internet Res.* **25**, 41992 (2023).
7. Broome, S. et al. Going deeper than tracking: A survey of computer-vision based recognition of animal pain and emotions. *Int. J. Comput. Vision* **131**(2), 572–590 (2023).
8. Andresen, N. et al. Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *PLoS ONE* **15**(4), 0228059 (2020).
9. Tuttle, A. H. et al. A deep neural network to assess spontaneous pain from mouse facial expressions. *Mol. Pain* **14**, 1744806918763658 (2018).
10. Lencioni, G. C., de Sousa, R. V., de Souza Sardinha, E. J., Corrêa, R. R. & Zanella, A. J. Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling. *PLoS ONE* **16**(10), 0258672 (2021).
11. Broomé, S., Gleerup, K.B., Andersen, P.H. & Kjellstrom, H. Dynamics are important for the recognition of equine pain in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12667–12676 (2019).
12. Pessanha, F., Salah, A. A., Loon, T. V. & Veltkamp, R. Facial image-based automatic assessment of equine pain. *IEEE Trans. Affect. Comput.*[SPACE]https://doi.org/10.1109/TAFFC.2022.3177639 (2022).
13. Feighelstein, M. et al. Automated recognition of pain in cats. *Sci. Rep.* **12**(1), 9575 (2022).
14. Feighelstein, M. et al. Explainable automated pain recognition in cats. *Sci. Rep.* **13**(1), 8973 (2023).
15. Feighelstein, M. et al. Deep learning for video-based automated pain recognition in rabbits. *Sci. Rep.* **13**(1), 14679 (2023).
16. Zhu, H., Salgırlı, Y., Can, P., Atılgan, D. & Salah, A.A. Video-based estimation of pain indicators in dogs. arXiv preprint arXiv:2209.13296 (2022).
17. Mahmoud, M., Lu, Y., Hou, X., McLennan, K. & Robinson, P. Estimation of pain in sheep using computer vision. Handbook of Pain and Palliative Care: Biopsychosocial and environmental approaches for the life course, 145–157 (2018).
18. Pessanha, F., McLennan, K. & Mahmoud, M. Towards automatic monitoring of disease progression in sheep: A hierarchical model for sheep facial expressions analysis from video. In: 2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020), pp. 387–393 (2020).
19. McLennan, K. & Mahmoud, M. Development of an automated pain facial expression detection system for sheep (ovis aries). *Animals* **9**(4), 196 (2019).
20. Labus, J. S., Keefe, F. J. & Jensen, M. P. Self-reports of pain intensity and direct observations of pain behavior: When are they correlated?. *Pain* **102**(1–2), 109–124 (2003).
21. Barrett, L. F. Feelings or words? Understanding the content in self-report ratings of experienced emotion. *J. Pers. Soc. Psychol.* **87**(2), 266–281 (2004).
22. Mogil, J. S., Pang, D. S., Dutra, G. G. S. & Chambers, C. T. The development and use of facial grimace scales for pain measurement in animals. *Neurosci. Biobehav. Rev.* **116**, 480–493 (2020).
23. Sotocina, S. G. et al. The rat grimace scale: A partially automated method for quantifying pain in the laboratory rat via facial expressions. *Mol. Pain* **7**, 1744–8069 (2011).
24. Keating, S. C., Thomas, A. A., Flecknell, P. A. & Leach, M. C. Evaluation of EMLA cream for preventing pain during tattooing of rabbits: Changes in physiological, behavioural and facial expression responses. *PloS one*[SPACE], https://doi.org/10.1371/journal.pone.0044437 (2012).
25. Dalla Costa, E. et al. Development of the horse grimace scale (hgs) as a pain assessment tool in horses undergoing routine castration. *PLoS ONE* **9**(3), 92281 (2014).
26. Di Giminiani, P. et al. The assessment of facial expressions in piglets undergoing tail docking and castration: Toward the development of the piglet grimace scale. *Front. Veter. Sci.* **3**, 100 (2016).
27. Reijgwart, M. L. et al. The composition and initial evaluation of a grimace scale in ferrets after surgical implantation of a telemetry probe. *PLoS ONE* **12**(11), 0187986 (2017).
28. McLennan, K. M. et al. Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Appl. Anim. Behav. Sci.* **176**, 19–26 (2016).
29. Häger, C. et al. The sheep grimace scale as an indicator of post-operative distress and pain in laboratory sheep. *PLoS ONE* **12**(4), 0175839 (2017).
30. Holden, E. et al. Evaluation of facial expression in acute pain in cats. *J. Small Anim. Pract.* **55**(12), 615–621 (2014).
31. Evangelista, M. C. et al. Facial expressions of pain in cats: The development and validation of a feline grimace scale. *Sci. Report* **9**(1), 1–11 (2019).
32. Brondani, J. T. et al. Validation of the english version of the unesp-botucatu multidimensional composite pain scale for assessing postoperative pain in cats. *BMC Vet. Res.* **9**(1), 1–15 (2013).

33. Reid, J. et al. Development of the short-form glasgow composite measure pain scale (cmps-sf) and derivation of an analgesic intervention score. *Anim. Welf.* **16**(S1), 97–104 (2007).
34. Haddad Pinho, R. et al. Validation of the rabbit pain behaviour scale (rpbs) to assess acute postoperative pain in rabbits (oryctolagus cuniculus). *PLoS One* **17**(5), 0268973 (2022).
35. Luna, S. P. L. et al. Validation of the unesp-botucatu pig composite acute pain scale (upaps). *PLoS One* **15**(6), 0233552 (2020).
36. Fonseca, M. W. et al. Development and validation of the unesp-botucatu goat acute pain scale. *Animals* **13**(13), 2136 (2023).
37. Silva, N. et al. Correction: Validation of the unesp-botucatu composite scale to assess acute postoperative abdominal pain in sheep (usaps). *PLoS ONE* **17**, 0268305. https://doi.org/10.1371/journal.pone.0268305 (2022).
38. Oliveira, M. G. et al. Validation of the donkey pain scale (dops) for assessing postoperative pain in donkeys. *Front. Veter. Sci.* **8**, 671330 (2021).
39. de Oliveira, F. A. et al. Validation of the unesp-botucatu unidimensional composite pain scale for assessing postoperative pain in cattle. *BMC Veter. Res.* **10**, 1–14 (2014).
40. De Sario, G. D. et al. Using ai to detect pain through facial expressions: A review. *Bioengineering* **10**(5), 548 (2023).
41. Robinson, M. E. & Wise, E. A. Gender bias in the observation of experimental pain. *Pain* **104**(1–2), 259–264 (2003).
42. Contreras-Huerta, L. S., Baker, K. S., Reynolds, K. J., Batalha, L. & Cunnington, R. Racial bias in neural empathic responses to pain. *PLoS ONE* **8**(12), 84001 (2013).
43. Adami, C., Filipas, M., John, C., Skews, K. & Dobson, E. Inter-observer reliability of three feline pain scales used in clinical practice. *J. Feline Med. Surg.* **25**(9), 1098612–231194423 (2023).
44. Reid, J., Scott, E., Calvo, G. & Nolan, A. Definitive glasgow acute pain scale for cats: Validation and intervention level. *Veterin. Record.* [SPACE], https://doi.org/10.1136/vr.104208 (2017).
45. Shipley, H., Guedes, A., Graham, L., Goudie-DeAngelis, E. & Wendt-Hornickle, E. Preliminary appraisal of the reliability and validity of the colorado state university feline acute pain scale. *J. Feline Med. Surg.* **21**(4), 335–339 (2019).
46. Weber, G., Morton, J. & Keates, H. Postoperative pain and perioperative analgesic administration in dogs: Practices, attitudes and beliefs of Queensland veterinarians. *Aust. Vet. J.* **90**(5), 186–193 (2012).
47. Williams, V., Lascelles, B. & Robson, M. Current attitudes to, and use of, peri-operative analgesia in dogs and cats by veterinarians in New Zealand. *N. Z. Vet. J.* **53**(3), 193–202 (2005).
48. Bell, A., Helm, J. & Reid, J. Veterinarians' attitudes to chronic pain in dogs. *Veter. Record* **175**(17), 428–428 (2014).
49. Kilkenny, C., Browne, W., Cuthill, I. C., Emerson, M. & Altman, D. G. Animal research: Reporting in vivo experiments: The arrive guidelines. *Br. J. Pharmacol.* **160**(7), 1577 (2010).
50. Banks, R. The Four Rs of research. *Contemp. Top. Lab. Anim. Sci.* **34**(1), 50–51 (1995).
51. Russell, W.M.S. & Burch, R.L. The principles of humane experimental technique. Methuen, (1959).
52. Teixeira, P. et al. Ovariectomy by laparotomy, a video-assisted approach or a complete laparoscopic technique in santa ines sheep. *Small Rumin. Res.* **99**(2–3), 199–202 (2011).
53. McLennan, K. M. et al. Development of a facial expression scale using footrot and mastitis as models of pain in sheep. *Appl. Anim. Behav. Sci.* **176**, 19–26. https://doi.org/10.1016/j.applanim.2016.01.007 (2016).
54. Vikramkumar, Vijaykumar, B., Trilochan: Bayes and naive bayes classifier. arXiv:abs/1404.0933 (2014).
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.* Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748–8763 (2021). PMLR.
56. Li, J. et al. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **50**(6), 1–45 (2017).
57. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**(3), 837–845 (1988).
58. Lu, Y., Mahmoud, M. & Robinson, P. Estimating sheep pain level using facial action unit detection. In: 2017 12th IEEE International conference on automatic face & gesture recognition (FG 2017), IEEE, pp. 394–399 (2017).
59. Evangelista, M. C., Monteiro, B. P. & Steagall, P. V. Measurement properties of grimace scales for pain assessment in nonhuman mammals: A systematic review. *Pain* **163**(6), 697–714 (2022).
60. Tomacheuski, R. M., Monteiro, B. P., Evangelista, M. C., Luna, S. P. L. & Steagall, P. V. Measurement properties of pain scoring instruments in farm animals: A systematic review using the cosmin checklist. *PLoS ONE* **18**(1), 0280830 (2023).

Acknowledgements

The first and last authors were supported by the Joint SNSF-ISF Research Grant Program (grant number 1050/24).

Author contributions

MF, SL, NS, PT and AZ conceived the study; MF ran the experiments; all authors analyzed the data and participated in writing the manuscript.

Additional information

Correspondence and requests for materials should be addressed to M.F. or A.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.