



OPEN Automated acute pain prediction in domestic goats using deep learning-based models on video-recordings

Ludovica Chiavaccini^{1✉}, Anjali Gupta¹, Nicole Anclade¹, Guido Chiavaccini², Chiara De Gennaro¹, Alanna N. Johnson¹, Diego A. Portela¹, Marta Romano¹, Enzo Vettorato¹ & Daniela Luethy³

Facial expressions are essential in animal communication, and facial expression-based pain scales have been developed for different species. Automated pain recognition offers a valid alternative to manual annotation with growing evidence across species. This study applied machine learning (ML) methods, using a pre-trained VGG-16 base and a Support Vector Machine classifier to automate pain recognition in caprine patients in hospital settings, evaluating different frame extraction rates and validation techniques. The study included goats of different breed, age, sex, and varying medical conditions presented to the University of Florida's Large Animal Hospital. Painful status was determined using the UNESP-Botucatu Goat Acute Pain Scale. The final dataset comprised images from 40 goats (20 painful, 20 non-painful), with 2,253 'non-painful' and 3,154 'painful' images at 1 frame per second (FPS) extraction rate and 7,630 'non-painful' and 9,071 'painful' images at 3 FPS. Images were used to train deep learning-based models with different approaches. The model input was raw images, and pain presence was the target attribute (model output). For the single train-test split and 5-fold cross-validation, the models achieved approximately 80% accuracy, while the subject-wise 10-fold cross-validation showed mean accuracies above 60%. These findings suggest ML's potential in goat pain assessment.

Keywords Artificial intelligence, Acute pain, Deep learning, Facial expression, Goats, Pain measurement

In animals, as in humans, facial expressions are a primary nonverbal means of regulating peer interactions and delivering information about emotional states. The use of facial expressions to interpret emotional conditions, like pain, has grown in humans and non-human species in the past two decades. Several facial expression-based pain scales have been developed for non-human animals like mice¹, rats², rabbits³, ferrets⁴, bovines⁵, sheep^{6,7}, piglets⁸, equids^{9–12} and cats¹³. Most of these pain scales were developed retrieving action units (AUs) from static images, various-length videos or during real-time scoring. The annotation was always manually made by human experts. Detailed descriptions and comparison on how the various animal facial expression-based pain scales were developed and validated can be found elsewhere^{14,15}. Interestingly, many of these investigations were conducted before establishing a formal codification system for facial expressions in the relevant species, such as the Facial Action Coding System (FACS)¹⁶. FACS is a comprehensive, anatomically-based system that taxonomizes all visible human facial movements. FACS has recently been developed for some non-human animals as well, including orangutans¹⁷, chimpanzees¹⁸, macaques¹⁹, marmosets²⁰, horses²¹, dogs and cats^{22,23}. While better standardizing the recognition of AUs, the use of AnimalFACS for facial expression is not without limitations, including the limited number of species for which it is available and its dependence on manual annotation, which requires extensive human training and certification; this can be time-consuming and expensive.

Automation has the potential to provide a complementary advancement to this process. Automated pain recognition (APR) is an external observation method in which image capture sensors are associated with pain

¹Department of Comparative, Diagnostic and Population Medicine, College of Veterinary Medicine, University of Florida, 2015 SW 16th Avenue, PO BOX 100123, Gainesville, FL 32610-0123, USA. ²Livorno, Italy. ³Department of Clinical Studies – New Bolton Center, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA. ✉email: lchiavaccini@ufl.edu

algorithms using Artificial Intelligence (AI)²⁴. Two main approaches to APR have been described to this point in animals: (i) landmark (LM)-based or key points and (ii) Convolutional Neural Networks (CNNs). Classical machine learning relies on hand-crafted features²⁵. The process is a mathematical representation of previous findings by human experts regarding pain-related patterns using facial or bodily landmarks, grimace scale elements, or pose representations. Most animal-based studies used this system, sourcing previously annotated data^{26–30}. Instead, the deep learning (DL) approach can use either standard (pre-trained) or customized CNNs together with a conventional classifier to learn features that might not be readily discernible to the human eye. Deep learning-based models learn features directly from raw data, eliminating the need for manual feature extraction, which reduces time, effort and minimizes the introduction of human bias that can occur during manual annotation in LM-based methods. Furthermore, unlike LM-based techniques, DL-based models are less affected by variation in facial alignment²⁷, making it a suitable option for evaluating animals in real-life clinical settings across various environments and lighting conditions. However, DL methods face challenges when dealing with small or very homogenous dataset and the accuracy of the models should be improved through data cleaning (noise reduction)^{27,31}. Expanding on the research conducted by Finka et al.³², Feighelestein et al.²⁷ compared LM-based and DL methods in automated pain recognition for cats, achieving similar accuracies of approximately 72%. Data imbalance and the issues of overfitting and underfitting are key limitations in both traditional machine learning (ML) and deep learning (DL) models. Overfitting occurs when a model becomes too specialized to the training data, which reduces its ability to generalize to new, unseen data. Conversely, underfitting happens when a model performs poorly even on the training set, indicating that it has failed to capture the underlying data patterns. To assess a model's performance on unseen data, several validation techniques are commonly used. The simplest method is the “single-train split,” where data, such as images, are randomly divided into training, validation, and testing sets prior to analysis³³. Another, more robust technique, is “k-fold cross-validation.” In k-fold cross-validation, the data is partitioned into multiple folds, usually between five and 10, and different combinations of these folds are used at various stages of training, validation, and testing. This process offers a more comprehensive evaluation compared to a single-train split. For smaller datasets, a more conservative approach called “subject-wise cross-validation” has been proposed^{25,27}. In this method, each subject is included in only one of the sets (training or testing) at a time, and subjects are rotated between sets throughout different rounds of the validation process, so that the system never sees the same subject twice.

Recently, there has been a growing awareness of the importance of identifying and relieving stress and pain in farm animals. In addition to the ethical considerations, it is now recognized that pain and distress affect the release of nociceptive neuroendocrine transmitters with consequent negative influence on surgical outcomes³⁴, productivity, the welfare of an animal and its quality of life³⁵. However, pain management in farm animals is often inadequate^{36,37}, due to the difficulty in recognizing and quantifying pain in these species. Behavioral pain scales have been developed and validated for some farm animal species^{38–40}. However, at the time of writing, only one behavioral pain scale was validated for domesticated goats⁴¹. Moreover, pain ethograms have yet to be developed in this species, limiting the possibility of developing an LM-based approach to APR. Given the increasing interest in goats as pets and models for biomedical research and surgical training and teaching^{42,43}, there is a crucial need to develop an objective, standardized and generalizable instrument for pain assessment in numerous clinical contexts.

The aim of this study was to explore APR in caprine patients in hospital setting using DL, and classifying facial images of goats into pain and no-pain categories (binary outcome). Given that Rashid et al.⁴⁴ observed that most pain AUs lasted 0.33 s in horses, we tested different frame extraction rates [1 and 3 frames per second (FPS)], balancing between the model temporal data richness and the computational requirements. Different validation techniques proposed in previous animal studies, such as single-train split validation³³, 5-fold cross-validation, and subject-wise cross-validation²⁷, were tested to assess the potential and limitations of this approach.

Results

Dataset

The raw dataset included footage from 74 adult goats of different breeds, sexes and ages. The estimated correlation (intra-rater correlation coefficient) between individual UNESP-Botucatu Goat Acute Pain Scale (UGASP) scoring was 0.96 (95% Confidence Interval: 0.93–0.98), indicating excellent consistency between ratings for the same goat across the two testing periods. Images of three non-painful and five painful goats were discarded due to the presence of other conditions that could have influenced their facial expressions. Additionally, one goat from the painful group was excluded as it appeared sedated. A further 25 painful goats were also excluded because they scored below the established cutoff of 3 out of 10 on the UGASP. Consequentially, the final dataset included images from 40 goats, totaling 2,253 images of ‘non painful’ and 3,154 images of ‘painful’ goats using a 1 FPS extraction rate. With a 3 FPS extraction rate, the dataset contained 7,630 images of ‘non-painful’ and 9,071 images of ‘painful’ goats. Tables 1 and 2 present the demographic information of the participating goats, including breed, age, sex, cause of hospitalization, whether they received any analgesic drugs within 24 h before video recording, the type and dose of the drugs administered, the timing of the last analgesic dose, and their UGASP scores.

Preliminary testing

The pilot dataset comprised 11,778 training images and 1,308 validation images. Both models demonstrated acceptable performance, achieving > 90% accuracy in discriminating pain in validation. However, the VGG-16 model with fine-tuning consistently outperformed, reaching a validation accuracy of 98.4%. The VGG-16 model was then selected for further analysis (Fig. 1).

ID	Breed	Age (years)	Sex	Cause of hospitalization	On treatment	Drugs and dose	Last treatment	UGASP
1	Toggenburg	5	F	Gestational check	No	n/a	n/a	0
2	Nubian	5	M	Pre-castration	No	n/a	n/a	0
3	Nigerian dwarf	7	F	Thymoma recheck	No	n/a	n/a	0
4	Mixed breed	1	F	Health check	No	n/a	n/a	0
5	Alpine cross	2	M	Teaching herd	No	n/a	n/a	0
6	Alpine cross	2	F	Teaching herd	No	n/a	n/a	0
7	Saanen	4	F	Teaching herd	No	n/a	n/a	0
8	Alpine cross	3	F	Teaching herd	No	n/a	n/a	0
9	Nubian	7	F	Pre-mastectomy	No	n/a	n/a	0
10	Nigerian dwarf	5	M	Pre-dental cleaning	No	n/a	n/a	0
11	La Mancha	4	F	Teaching herd	No	n/a	n/a	0
12	La Mancha	7	F	Teaching herd	No	n/a	n/a	0
13	Mixed breed	2	M	Pre-castration	No	n/a	n/a	0
14	La Mancha	1	F	Teaching herd	No	n/a	n/a	0
15	La Mancha	6	F	Teaching herd	No	n/a	n/a	0
16	Nigerian dwarf	0.5	M	Pre-castration	No	n/a	n/a	0
17	Oberhasli	3	F	Pregnancy check	No	n/a	n/a	0
18	Pygmy goat	6	MC	Pre-elective PU	No	n/a	n/a	0
19	Mixed breed	6	MC	Recheck neck wound	No	n/a	n/a	0
20	La Mancha	10	F	Teaching herd	No	n/a	n/a	0

Table 1. Demographic of goats in the class ‘non-painful’ and UGASP scoring performed at check-up examination. F, female; M, male; MC, male castrated; UGASP, UNESP-Botucatu Goat Acute Pain Scale.

ID	Breed	Age (years)	Sex	Cause of hospitalization	On treatment	Drugs and dose	Last treatment (hours)	UGASP
1	Pygmy goat	5	F	C-section	Yes	Morphine 0.2 mg/kg IV	5	5.5
2	La Mancha	6	F	C-section	No	n/a	n/a	6
3	Nigerian dwarf	1	MC	Hepatopathy	No	n/a	n/a	4
4	Nubian	2	M	Amputation	Yes	PI/PS block -bupivacaine 0.5%	5	4
5	Nigerian dwarf	11	F	Mastectomy	Yes	Morphine 0.13 mg/kg epidurally	7	4
6	Pygmy goat	6	F	C-section	Yes	Flunixin meglumine 1.1 mg/kg IV	7	5
7	Nigerian dwarf	0.5	F	Dog attack	No	n/a	n/a	5
8	Mixed breed	3	M	Tube cystotomy revision	No	n/a	n/a	5
9	Saanen	3	F	Dog attack	Yes	Flunixin meglumine 1.1 mg/kg IV	9	4
10	La Mancha	3	F	Mastectomy	No	n/a	n/a	5
11	Boer goat	7	F	Mastectomy	Yes	Flunixin meglumine 1.1 mg/kg IV	11	6
12	Nigerian dwarf	3	MC	Tube cystotomy	No	n/a	n/a	7
13	Nigerian dwarf	7	MC	Acute urinary obstruction	No	n/a	n/a	4
14	Nigerian dwarf	7	MC	Tube cystotomy	No	n/a	n/a	6
15	Mixed breed	8.5	MC	Perineal urethrostomy	Yes	Morphine 0.17 mg/kg IV	6	4
16	Nigerian dwarf	0.5	MC	Segmental ostectomy	Yes	Methadone 0.2 mg/kg IV	6	4
17	Nigerian dwarf	1	MC	Thoracic limb amputation	No	n/a	n/a	4
18	Pygmy goat	6	MC	Perineal urethrostomy	Yes	Morphine 0.1 mg/kg epidurally	6	5
19	La Mancha	6	F	Mastectomy	Yes	Morphine 0.1 mg/kg epidurally	7	4
20	Pygmy goat	5	MC	Tube cystotomy	Yes	Morphine 0.1 mg/kg epidurally	7	6

Table 2. Demographic of goats in the class ‘painful’ and UGASP scoring performed at physical examination or after recovery. Last treatment is intended as hours since the last analgesic treatment was performed prior to video recording. F, female; IV, endovenous; M, male; MC, male castrated; PI/PS, preiliac-parasacral; UGASP, UNESP-Botucatu Goat Acute Pain Scale.

Development of the DL model

Tables 3 and 4 present the performance results of the proposed AI pipelines at different FPS extraction rates. In both the 5-fold cross-validation and the subject-wise 10-fold cross-validation approaches, increasing the FPS rate from 1 to 3 slightly decreased model performance on the dataset, with the 3 FPS extraction rate performing marginally worse than the 1 FPS rate. Tables 5 and 6 present the testing results for the ‘non-painful’ and ‘painful’

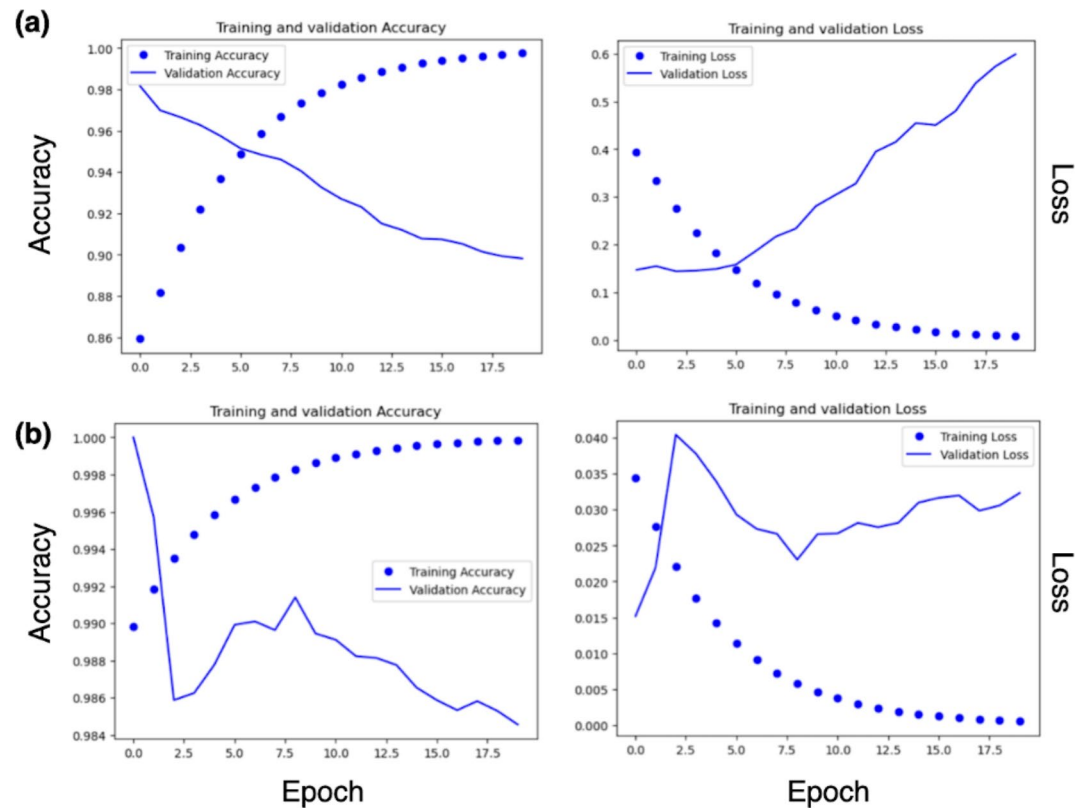


Fig. 1. Performance of the Artificial Neuronal Network (ANN) on a preliminary dataset of eight videos, featuring four goats of different breeds per group (‘painful’ and ‘non-painful’), based on (a) a customized convolutional base and (b) the pre-trained VGG-16 with fine tuning.

Approach	FPS	Validation method	Accuracy	Precision	Recall	F1
DL	1	Single train-test split	0.82	0.97	0.62	0.76
DL	3		0.84	0.99	0.64	0.69
DL	1	5-folds cross-validation	0.81	0.93	0.82	0.87
DL	3		0.79	0.75	0.97	0.85

Table 3. Testing results of the different approaches tested. DL, deep learning; FPS, frames per second extraction rate.

FPS		Precision	Recall	F1
1	Non-painful	0.76	0.98	0.86
	Painful	0.97	0.62	0.76
3	Non-painful	0.78	0.99	0.88
	Painful	0.99	0.64	0.77

Table 4. Testing results for the VGG-16 base and support Vector Machine (SVM) classifier with single train-test split validation separated by class, at 1 and 3 frames per second (FPS) extraction rates.

classes for the single train-test split validation and 5-fold cross-validation models. Under the single train-test split validation, the classifier achieved high precision (0.97–0.99) for detecting painful states across different frames extraction rates, though it showed significant discrepancies when identifying ‘non-painful’ goats. In contrast, the 5-fold cross-validation model provided more balanced precision between the ‘non-painful’ and ‘painful’ classes, with precision scores of 0.97 and 0.80, respectively. Table 4 represents the mean \pm SD of the eight closest results of the subject-wise 10-fold cross-validation model. The subject-wise 10-fold cross-validation results at 1 and 3 FPS extraction rates showed consistency but moderate performance metrics. At 1 FPS, the classifier achieved an accuracy of 0.62 ± 0.10 , with a precision of 0.55 ± 0.20 , recall of 0.60 ± 0.20 and an F1 score

FPS		Precision	Recall	F1
1	Non-painful	0.97	0.69	0.81
	Painful	0.80	0.98	0.88
3	Non-painful	0.93	0.57	0.70
	Painful	0.75	0.97	0.84

Table 5. Testing results for the VGG-16 base and support Vector Machine (SVM) classifier with 5-folds cross-validation separated by class, at 1 and 3 frames per second (FPS) extraction rates.

Approach	FPS	Validation method	Accuracy	Precision	Recall	F1
DL-SW	1	10-fold cross-validation	0.62 ± 0.10	0.55 ± 0.20	0.60 ± 0.20	0.58 ± 0.20
DL-SW	3		0.60 ± 0.08	0.54 ± 0.11	0.64 ± 0.22	0.59 ± 0.15

Table 6. Cross-validation results for the VGG-16 base and support Vector Machine (SVM) classifier, separated by class and evaluated at extraction rates of 1 and 3 frames per second (FPS). In subject-wise (SW) 10-fold cross-validation, the model is trained on one subset and tested on another, with no overlap between training and testing subjects. We repeatedly trained on 33 goats and tested on five (three painful and two non-painful). Results represent mean ± SD of the eight closest models.

of 0.58 ± 0.20. At 3 FPS, the accuracy was slightly lower at 0.60 ± 0.08, with a precision of 0.54 ± 0.11, recall of 0.64 ± 0.22 and an F1 score of 0.59 ± 0.15.

Discussion

This study presents a fully automated AI pipeline for detecting pain in goats by analyzing frames extracted from raw video footage, without requiring manual annotation or preprocessing. We explored various validation techniques previously applied for other species. Our results show an overall accuracy ≥ 79% with single train-test split and k-fold cross-validation methods. However, accuracy dropped to approximately 60% when using subject-exclusive pre-splitting testing. This aligns with the performance reported in previous animal studies with similar size datasets^{31,44,45}.

A discrepancy in precision and recall between positive (painful) and negative (non-painful) cases observed could in part be due to the small sample size, as the DL approach is notoriously “data hungry”. With a small sample size, the model may not see enough examples of each class, particularly if the data is imbalanced. This can lead to the model failing to learn the full range of variability within each class or cause overfitting. Thus investigating whether the performance of the DL approach is improved by enlarging the dataset is an immediate priority for future research. Additionally, the features used to discriminate between ‘painful’ and ‘non-painful’ states might overlap, or ‘painful’ states might exhibit a wider range of expressions compared to ‘non-painful’ states. This ambiguity was previously observed by Häger et al.⁶ and McLennan and Mahmoud⁴⁶ who both reported a high number of false positives in facial pain recognition in sheep and emphasized the need for further refinement and standardization in the field. Further data cleaning and regularization techniques could improve the balance between precision and recall for positive and negative cases. Finally, only two goats, one ‘painful’ and one ‘non-painful’, randomly withdrawn before model development, were used for testing for the single train-test split and 5-fold cross-validation methods. While this number aligns with similar studies in animals^{27,31,33}, the two goats might have not been a good representation of their category.

The first crucial step in implementing APR is data collection, which is a challenge in veterinary science due to limited available datasets. This study addressed this by creating a custom goat dataset explicitly tailored for this purpose. Unlike other studies that used pre-existing annotated databases^{27,44,47–50}, our dataset included goats with pain due to spontaneous acute conditions (urethral blockage) and after surgical procedures (and before treatment). Stimulus-based annotations would have used a within-subject design. In this setting, the animal is evaluated before and after the induction of a standardized painful stimulus^{33,51}. This design offer a potential solution to the challenge of variability in pain perceptions across individuals. However, since pain is a subjective and complex elaboration of the nociceptive experience, identical stimuli cannot guarantee identical pain expression, as demonstrated by Gleerup et al.⁵². On the other hand, exposure to various scenarios allowed the DL-based model trained on diverse data to effectively generalize to new examples. It helped capturing robust features that can be applied in different contexts^{14,53}.

Before the advent of DL, classical machine learning relied heavily on hand-crafted features²⁵. This process involved extracting characteristics from data using prior knowledge to capture pain-related patterns via facial or bodily landmarks, grimace scale elements, or pose representations. For example, Feighelstein et al.²⁷ utilized 48 facial landmarks based on the CatFACS and manually annotated them to develop their automated model. Similarly, Pessanha et al.³⁰ used a previously annotated dataset of sheep affected by mastitis and pregnancy toxemia⁷. While LM-based approaches have the advantage of directly measuring and accounting for morphological variability, they are resource-intensive due to the manual effort required for landmark annotation⁵⁴. The absence of a standardized FACS or annotated dataset for goats further complicated this approach. In contrast, the DL approach is gaining popularity in APR because it reduces the need for annotation and manual feature

crafting. Unlike LM-based methods, DL-based methods are less sensitive to facial alignment²⁷, making them particularly appealing for recognizing animal facial expressions in clinical or farm settings. However, DL-based models require large volumes of data for training³¹, which are rarely readily available in veterinary medicine. To overcome this problem, researchers have used data augmentation techniques^{26,27}. Data augmentation is a set of techniques that enhance the size and quality of machine learning training datasets so that better DL models can be trained with them. Image augmentation algorithms include geometric transformations, color space augmentation, mixing images, random erasing, feature space augmentation. However, with careful selection of the data used for learning, even small samples can yield powerful diagnostic tools. Variability is one of the strengths of our dataset. Previous studies often included animals of similar breeds, age, sex and color^{27,33}. Our dataset encompassed over 12 different caprine breeds equally distributed between positive (painful) and negative (non-painful) cases, animals from 6 months of age to 15 years and intact and sterilized goats.

The pre-splitting subject-exclusive validation used in this study ensures no subject overlap between training and testing sets. This technique has been recommended for preventing the model from learning individual-specific features, which can lead to overfitting, especially in relatively small samples like ours. Subject-exclusive validation was proposed by Feighelstein et al.²⁷ but it has been employed in only a few studies on pain recognition in animals. For example, Broome et al.⁴⁵ achieved 67.6% accuracy for horses, Rashid et al.⁴⁴ reached 60.9% for horses and Feighelstein et al.³¹ reached 65.1% in cats. We reached a mean accuracy of 62% and 60% with 1 FPS and 3 FPS extraction rates respectively when using the subject-wise 10-fold cross-validation. Although these results may seem low, it is important to note that expert accuracy in distinguishing pain from facial expressions in horses has been reported at $58.0 \pm 13.6\%$ ⁴⁵. Similarly, in human medicine, nurses correctly identified postoperative pain scores ≥ 4 and ≥ 7 on a 0–10 scale in human patients with sensitivities of 44.9% and 17.0%, respectively⁵⁵. The next step will be to perform the leave-one-subject-out cross-validation with no subject overlap.

We did not find an advantage in extracting at 3 FPS compared to 1 FPS with the 5-fold cross-validation and the subject-wise 10-fold cross-validation. Our results are in contrast with those of Martvel et al.²⁸, who found that the model's accuracy increased, although slightly, with the FPS extraction rate when applied to two distinct feline datasets. The discrepancy could be due to the two different approaches. Martvel et al.²⁸ used an LM-based approach that typically identify and analyze specific vital points or features (such as facial landmarks) within each frame. The increase in frame extraction rate can benefit this approach by increasing temporal resolution, lowering the chances of missing key features and enhancing feature stability. On the other hand, DL-based methods process the entire frame and learn features directly from the raw pixel data. Increasing the extraction rate in this context can cause the model to overfit the redundant information in the additional frames, especially if the frames are highly similar, like in our case. Increasing the computational load can lead to inefficiencies, practical limitations in training and inference, increasing the chance of introducing noisy or irrelevant data.

This study presents several limitations: (1) First, it was an observational study, with the decision to administer analgesics made solely by the attending clinician. (2) Pain assessment in veterinary medicine is inherently challenging due to the lack of verbal communication. Typically, new pain scales are validated against an established gold standard¹³. In this study, goats were classified as 'painful' if they scored UGASP ≥ 3 and had clinical reasons to be in pain. However, the UGASP has been validated exclusively for young, healthy goats housed in group pens and limited to two specific caprine breeds undergoing a single soft tissue surgery (orchietomy), potentially limiting its reliability in our population. (3) While the operator did not interact with the goat, the goat was not completely blinded to the presence of the operator or environmental stimuli. (4) Finally, animals were filmed at different times during the day and the circadian rhythm may have affected pain expression. The inclusion of a small number of institutionally owned goats in the control group was not considered a limitation, as these animals are not housed on hospital premises and their familiarity with the environment was not considered different from that of the client-owned goats. Finally, one additional consideration should be made on the costs of the described method. Studies demonstrated that camera resolutions of 224×224 pixels and frame rates of 25 FPS are sufficient for processing images and videos in modern CV systems⁴⁷, and they are readily available nowadays. While required software is largely open-source and available at no cost, large datasets and advanced data processing may require hefty computational resources not readily available in all realities.

In conclusion, we introduced DL-based methods to distinguish between 'painful' and 'non-painful' goats with acute painful conditions. By testing various frame extraction rates and validation methods—including 5-fold and subject-wise 10-fold cross-validation—we observed no significant differences in model performance between the 1 FPS and 3 FPS extraction rates. However, the results varied considerably depending on the validation method used. This variability highlights the importance of considering the choice of validation technique when comparing results across studies. While further sophistication is needed to make DL models reliable for individual pain assessment in clinical settings, future work should focus on using more extensive, diverse goat populations and additional behavioral data to enhance classification accuracy.

Methods

The study was evaluated by the University of Florida (UF) Institutional Animal Care and Use Committee (IACUC#20220000709) and was exonerated from IACUC approval as it was deemed strictly observational. Owners provided consent to video-record their animals, along with consent for treatment. All the methods were carried out in accordance with relevant Institutional guidelines and regulations.

Animals

Sixty-five (44 'painful' and 21 'non-painful') client-owned adult goats (≥ 6 months of age) presented to the UF Veterinary Hospital were enrolled independently of breed, color, age, or reason for presentation. Nine goats from the teaching colony at our institution were also included as controls (non-painful goats). Institutional goats

were deemed systemically healthy based on physical examination, blood work screening (complete blood count, serum biochemistry, and electrolytes), and parasite testing (McMaster's fecal analysis) and bacterial testing (*Coxiella burnetii*). Goats were enrolled between February 2023 and June 2024. Exclusion criteria included goats presented with diseases or conditions that could affect facial expressions (beside pain) or whose facial features were likely impaired by surgery or dressings (dehorning). Goats with a history of chronic pain or mental impairment or who were administered sedatives and/or analgesics within four hours from video-recording were also excluded.

Image capture and pain assessment

Upon presentation and after a complete clinical examination, goats were left in individual pens for at least 5–10 min to settle. Goats were then filmed for 2 min using a high-definition camera at 1080p resolution and 12-megapixel frames at 60 FPS (iPhone 12 Pro, Apple Inc., Cupertino, CA) placed at the level of the eyes at a distance of approximately 1 m. Goats were filmed either within the hospital treatment room or while contained in the goat's individual box stall, with attention given to capturing a similar number of frames from both front and profile views. No feed was available if goats were recorded in the hospital treatment room. If goats were recorded in their stalls, they usually had timothy hay available free-choice in the stall, although feeding was determined by the attending clinician and the clinical diagnosis of the goat. The video recording was performed at least four hours after recovery from general anesthesia for goats undergoing surgical procedures, considering evidence that anesthesia can affect pain-induced facial expressions for an extended period in laboratory rodents^{56,57} and client-owned horses⁵⁸. Hospital staff did not interact with the goats and goats were left undisturbed during video recording. The recordings were made either with a tripod placed inside the stall or using a maneuverable stick from outside. Although the operator avoided direct interaction with the goats, the animals were not completely unaware of the operator's presence. All goats suffering from disease were treated appropriately and revisited throughout recovery. No treatment was modified or withheld during the study. After reviewing the recordings, a veterinarian board-certified in anesthesia and pain management (LC), experienced in goat pain assessment, scored pain intensity using the UGASP scale, immediately and again six months later. Intraclass Correlation Coefficient (ICC) was used to assess intra-rater consistency of rating. The UGASP validated scale comprises five main items and ten subitems assessed for their presence or absence, offering a possible final score from 0 to 10⁴¹. A score ≥ 3 was classified as painful, following the authors' recommended cutoff for pain treatment⁴¹. To ascertain discrimination, only goats belonging to the 'non-painful' group and scoring UGASP 0 and goats scoring UGASP ≥ 3 and with clinical reasons to suspect pain were retained for analysis. The final dataset included 40 goats, 20 deemed painful and 20 deemed non-painful.

Preliminary testing

For the initial processing of the binary image classification task (painful vs. non-painful), a preliminary dataset of eight videos randomly selected among the first 25 videos recorded, featuring four goats of different breeds per group, was utilized. The preprocessing involved the extraction of images from Full-HD video at 60 FPS using the FFmpeg (an open-source software project developed and maintained by a community of contributors; <https://ffmpeg.org>), the cropping of an 800 × 800 pixel window around the goats' muzzles, the resizing of the images to 400 × 400 pixels and the removal of the background with Rembg version 2.0 (an open-source tool available at Rembg GitHub repository). Two networks were trained and externally validated: one based on a custom convolutional base with five Conv2D layers (filters: 32, 64, 128, 256, 512; kernel size: 3 × 3; ReLU activation; pool size: 2 × 2) and five MaxPooling2D layers, resulting in an output shape of (10, 10, 512). The second network employed a pre-trained standard convolutional base [Visual Geometry Group (VGG)-16] with fine-tuning of the upper layer (block 5), achieving an output shape of (12, 12, 512). The custom classifier included a flattened layer, a Dense layer with 128 nodes and ReLU activation, a Dropout layer with a 50% rate and a final Dense layer with two nodes and Softmax activation. The training was conducted over 20 epochs (369 steps per epoch) using binary cross-entropy loss, the RMSprop optimizer (learning rate = 1e-5) and accuracy as the evaluation metric.

Development of the DL-based model

In the second phase, we selected the standard convolutional base (VGG-16) pre-trained on the ImageNet dataset, with fine-tuning to study the DL-based model on the entire dataset. The input was raw images (model input), with no alignment or augmentation and the pain presence was the target attribute (model output).

Preprocessing pipeline

Utilizing the Open Source Computer Vision (OpenCV) library 4.10.0, frames were extracted from the videos. Since Martvel et al.²⁸ suggested that FPS rates may affect accuracy, we experimented with 1 and 3 FPS sampling rates. The FPS rate was chosen based on the observation made by Rashid et al.⁴⁴ that most pain AUs last 0.33 s in horses. The input size for the frames was set to (224, 224, 3) to match the model's expected input dimensions. The focus was on selecting clear frames where the goat's face was visible with at least one eye and one ear. The preprocessing pipeline is illustrated in Fig. 2.

Model training

The frames were fed into the VGG-16 model, bypassing the final classification layers. The model was modified by removing its top classification layers (include_top=False) and applying global average pooling (pooling='avg'), resulting in a 4096-dimensional feature vector for each frame. The resulting feature vectors were stored and labeled according to their corresponding image classifications ('painful' or 'non-painful'). After extracting features using VGG-16, we employed a Support Vector Machine (SVM) classifier, a robust and widely used method for classification tasks⁵⁹, with a linear kernel (kernel='linear'), to discern between 'painful' and 'non-

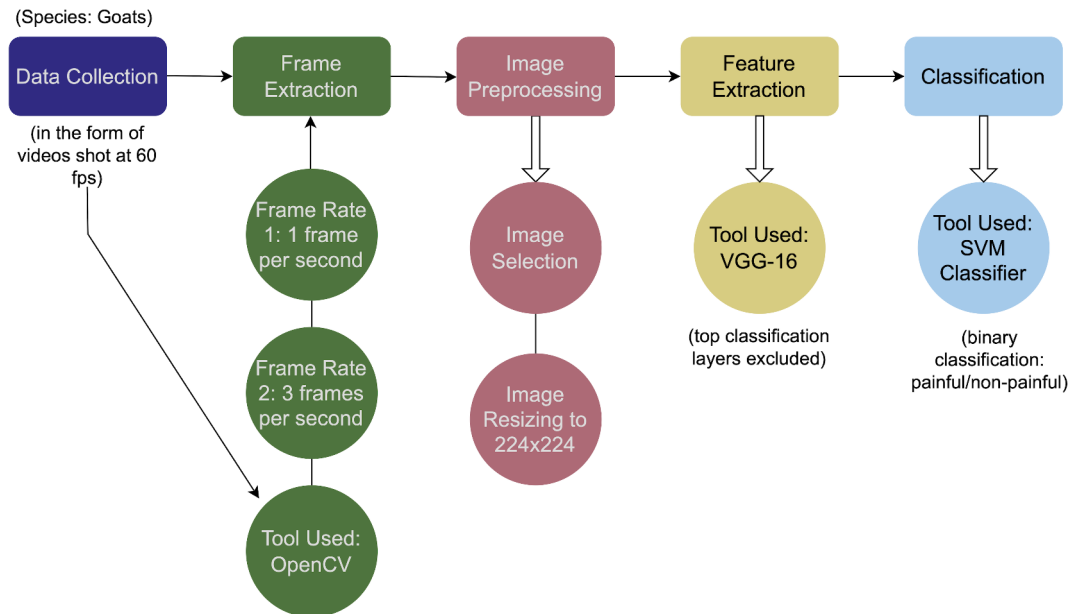


Fig. 2. Preprocessing pipeline for the deep learning (DL)-based model developed on 40 client-owned adult goats (≥ 6 months old) presented to the UF Veterinary Hospital for different conditions.

‘painful’ expressions in goats. The modeling cycle was accomplished using 80% of the images for training and 20% for validation. Images were randomly assigned to training and validation sets. To ensure a more robust evaluation, we also employed 5-folds cross-validation. The performance metrics from each of the five iterations were averaged to give a more reliable estimate of the model’s performance. This helps mitigate the effects of variance in the data splits and provides a more robust evaluation than a single train-test split. Two videos of two independent goats (one ‘painful’ and one ‘non-painful’) that were randomly left out during the training and validation phase were used for testing. All stages of the pipeline were implemented in Python, leveraging Keras with a TensorFlow backend for DL tasks and Scikit-Learn for the SVM classifier.

Model evaluation

For measuring performance of models, we use standard evaluation metrics of accuracy,

precision, recall as previously described^{27,33}. True positives contained images belonging to class ‘painful,’ extracted from video footage of goats post-operatively, before rescue analgesia, or considered to be in pain and recording a UGASP score $\geq 3/10$ and that the model had correctly classified. True negatives contained images belonging to the class ‘non-painful,’ extracted from video footage of non-painful goats that scored UGASP = 0/10 and the model had correctly classified as not being painful. The set of ‘false positives’ contained images belonging to the class ‘non-painful,’ which were incorrectly classified by the model as painful. The set of ‘false negatives’ contained images belonging to the class ‘painful,’ which were incorrectly classified by the model as non-painful.

Accuracy indicates the overall efficiency of the model and was calculated as:

$$\text{Accuracy} = (\text{True positives} + \text{True negatives}) / \text{Total}.$$

The precision indicates whether the data was classified in the correct class or not and is calculated as:

$$\text{Precision} = \text{True positive} / (\text{True positives} + \text{False positives}).$$

Finally recall measures the ability of the classifier to identify all the correct data for each class and it is calculated as:

$$\text{Recall} = \text{True positives} / (\text{True positives} + \text{False negative}).$$

The F1 score combines the latter two metrics into a single value by taking their harmonic mean.

Subject-wise cross-validation

Subject-wise cross-validation involves dividing the dataset based on the single animal to ensure that all data points from a specific animal are either in the training or the hold-out set, but not both. In the last part of the study, we trained the model multiple times on 33 balanced goats (16 classified as ‘painful’ and 17 as ‘non-painful’) and then tested the model on five goats (3 classified as ‘painful’ and 2 as ‘non-painful’), using 10-fold cross-validation. Implementing 10-fold cross-validation within subject-wise cross-validation enables a comprehensive evaluation across different subsets of data, enhancing confidence in model performance metrics. This approach helps to prevent data leakage, provides a more realistic assessment of model performance and has been recommended for small datasets²⁷.

Data availability

The dataset is available from the corresponding author upon reasonable request.

Received: 17 July 2024; Accepted: 31 October 2024

Published online: 07 November 2024

References

- Langford, D. J. et al. Coding of facial expressions of pain in the laboratory mouse. *Nat. Methods*. **7**, 447–449. <https://doi.org/10.1038/nmeth.1455> (2010).
- Sotocinal, 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**, 55. <https://doi.org/10.1186/1744-8069-7-55> (2011).
- 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*. **7**, e44437. <https://doi.org/10.1371/journal.pone.0044437> (2012).
- 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**, e0187986. <https://doi.org/10.1371/journal.pone.0187986> (2017).
- Yamada, P. H. et al. Pain assessment based on facial expression of bulls during castration. *Appl. Anim. Behav. Sci.* **236**, 105258. <https://doi.org/10.1016/j.applanim.2021.105258> (2021).
- Häger, C. et al. The Sheep Grimace Scale as an indicator of post-operative distress and pain in laboratory sheep. *PLoS One*. **12**, e0175839. <https://doi.org/10.1371/journal.pone.0175839> (2017).
- 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).
- Viscardi, A. V., Hunniford, M., Lawlis, P., Leach, M. & Turner, P. V. Development of a piglet grimace scale to evaluate piglet pain using facial expressions following castration and tail docking: A pilot study. *Front. Vet. Sci.* **4**, 51. <https://doi.org/10.3389/fvets.2017.00051> (2017).
- 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**, e92281. <https://doi.org/10.1371/journal.pone.0092281> (2014).
- van Loon, J. P. & Van Dierendonck, M. C. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): A scale-construction study. *Vet. J.* **206**, 356–364. <https://doi.org/10.1016/j.tvjl.2015.08.023> (2015).
- VanDierendonck, M. C. & van Loon, J. P. Monitoring acute equine visceral pain with the Equine Utrecht University Scale for Composite Pain Assessment (EQUUS-COMPASS) and the Equine Utrecht University Scale for Facial Assessment of Pain (EQUUS-FAP): A validation study. *Vet. J.* **216**, 175–177. <https://doi.org/10.1016/j.tvjl.2016.08.004> (2016).
- Orth, E. K. et al. Development of a Donkey Grimace Scale to Recognize Pain in donkeys (*Equus asinus*) Post Castration. *Anim. (Basel)*. **10**, 1411. <https://doi.org/10.3390/ani10081411> (2020).
- Evangelista, M. C. et al. Facial expressions of pain in cats: The development and validation of a feline grimace scale. *Sci. Rep.* **9**, 19128. <https://doi.org/10.1038/s41598-019-55693-8> (2019).
- Chiavaccini, L., Gupta, A. & Chiavaccini, G. From facial expressions to algorithms: A narrative review of animal pain recognition technologies. *Front. Vet. Sci.* **11**, 1436795. <https://doi.org/10.3389/fvets.2024.1436795> (2024).
- McLennan, K. M. et al. Conceptual and methodological issues relating to pain assessment in mammals: The development and utilisation of pain facial expression scales. *Appl. Anim. Behav. Sci.* **217**, 1–15. <https://doi.org/10.1016/j.applanim.2019.06.001> (2019).
- Ekman, P. & Friesen, W. V. *Manual for the Facial Action Code* (Consulting Psychologists, 1978).
- Caeiro, C. C., Waller, B. M., Zimmermann, E., Burrows, A. M. & Davila-Ross, M. OrangFACS: A muscle-based facial movement coding system for orangutans (*Pongo spp.*). *Int. J. Primatol.* **34**, 115–129. <https://doi.org/10.1007/s10764-012-9652-x> (2013).
- Parr, L. A., Waller, B. M., Vick, S. J. & Bard, K. A. Classifying chimpanzee facial expressions using muscle action. *Emotion*. **7**, 172–181. <https://doi.org/10.1037/1528-3542.7.1.172> (2007).
- Correia-Caeiro, C., Holmes, K. & Miyabe-Nishiwaki, T. Extending the MaqFACS to measure facial movement in Japanese macaques (*Macaca fuscata*) reveals a wide repertoire potential. *PLoS One*. **16**, e0245117. <https://doi.org/10.1371/journal.pone.0245117> (2021).
- Correia-Caeiro, C. et al. The common marmoset facial action coding system. *PLoS One*. **17**, e0266442. <https://doi.org/10.1371/journal.pone.0266442> (2022).
- Wathan, J., Burrows, A. M., Waller, B. M. & McComb, K. EquiFACS: The equine facial action coding system. *PLoS One*. **10**, e0131738. <https://doi.org/10.1371/journal.pone.0131738> (2015).
- Caeiro, C. C., Burrows, A. M. & Waller, B. M. Development and application of CatFACS: Are human cat adopters influenced by cat facial expressions? *Appl. Anim. Behav. Sci.* **189**, 66–78. <https://doi.org/10.1016/j.applanim.2017.01.005> (2017).
- Waller, B. M. et al. Paedomorphic facial expressions give dogs a selective advantage. *PLoS One*. **8**, e82686 (2013).
- Walter, S. et al. What about Automated Pain Recognition for Routine Clinical Use? A survey of Physicians and nursing staff on expectations, requirements, and Acceptance. *Front. Med.* **7** <https://doi.org/10.3389/fmed.2020.566278> (2020).
- Broomé, S. et al. Going deeper than Tracking: A survey of computer-vision based recognition of animal pain and emotions. *Int. J. Comput. Vis.* **131**, 572–590. <https://doi.org/10.1007/s11263-022-01716-3> (2023).
- Pessanha, F., Salah, A., van Loon, A., Veltkamp, R. & T. & Facial image-based automatic assessment of equine pain. *IEEE Trans. Affect. Comput.* **14**, 2064–2076. <https://doi.org/10.1109/taffc.2022.3177639> (2023).
- Feighelstein, M. et al. Automated recognition of pain in cats. *Sci. Rep.* **12**, 9575. <https://doi.org/10.1038/s41598-022-13348-1> (2022).
- Martvel, G. et al. Automated pain recognition in cats using facial landmarks: Dynamics matter. *Preprint at (2023)*. <https://www.researchsquare.com/article/rs-3754559/v1>.
- Lu, Y., Mahmoud, M. & Robinson, P. in *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. 394–399 (IEEE). (2017).
- Pessanha, F., McLennan, K. & Mahmoud, M. M. Towards automatic monitoring of disease progression in sheep: A hierarchical model for sheep facial expressions analysis from video. *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 387–393 (2020).
- Feighelstein, M. et al. Explainable automated pain recognition in cats. *Sci. Rep.* **13**, 8973. <https://doi.org/10.1038/s41598-023-35846-6> (2023).
- Finka, L. R. et al. Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. *Sci. Rep.* **9** <https://doi.org/10.1038/s41598-019-46330-5> (2019).
- 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**, e0258672. <https://doi.org/10.1371/journal.pone.0258672> (2021).
- Casoni, D., Mirra, A., Suter, M. R., Gutzwiller, A. & Spadavecchia, C. Can disbudding of calves (one versus four weeks of age) induce chronic pain? *Physiol. Behav.* **199**, 47–55. <https://doi.org/10.1016/j.physbeh.2018.11.010> (2019).
- Newton, H. P. & O'Connor, A. M. The economics of pain management. *Vet. Clin. North. Am. Food Anim. Pract.* **29**, 229–250. <https://doi.org/10.1016/j.cvfa.2012.11.010> (2013).

36. Fajt, V. R., Wagner, S. A. & Norby, B. Analgesic drug administration and attitudes about analgesia in cattle among bovine practitioners in the United States. *J. Am. Vet. Med. Assoc.* **238**, 755–767. <https://doi.org/10.2460/javma.238.6.755> (2011).
37. Alvarez, L., Nava, R. A., Ramírez, A., Ramírez, E. & Gutiérrez, J. Physiological and behavioural alterations in disbudded goat kids with and without local anaesthesia. *Appl. Anim. Behav. Sci.* **117**, 190–196. <https://doi.org/10.1016/j.applanim.2009.01.001> (2009).
38. de Oliveira, F. A. et al. Validation of the UNESP-Botucatu unidimensional composite pain scale for assessing postoperative pain in cattle. *BMC Vet. Res.* **10**, 200. <https://doi.org/10.1186/s12917-014-0200-0> (2014).
39. Luna, S. P. L. et al. Validation of the UNESP-Botucatu pig composite acute pain scale (UPAPS). *PLoS One.* **15**, e0233552. <https://doi.org/10.1371/journal.pone.0233552> (2020).
40. Silva, N. et al. Validation of the Unesp-Botucatu composite scale to assess acute postoperative abdominal pain in sheep (USAPS). *PLoS One.* **15**, e0239622. <https://doi.org/10.1371/journal.pone.0239622> (2020).
41. Fonseca, M. W. et al. Development and validation of the Unesp-Botucatu Goat Acute Pain Scale. *Anim. (Basel).* **13**, 2136. <https://doi.org/10.3390/ani13132136> (2023).
42. Steen, J. et al. A single-cohort retrospective analysis of factors associated with morbidity and mortality in 193 anesthetized domestic goats. *Veterinary Anaesth. Analg.* **50**, 245–254. <https://doi.org/10.1016/j.vaa.2023.02.003> (2023).
43. Fulton, L. K., Clarke, M. S. & Farris, H. E. Jr The goat as a model for biomedical research and teaching. *ILAR J.* **36**, 21–29 (1994).
44. Rashid, M., Silventoinen, A., Gleerup, K. B. & Andersen, P. H. Equine Facial Action Coding System for determination of pain-related facial responses in videos of horses. *PLoS ONE.* **15**, e0231608. <https://doi.org/10.1371/journal.pone.0231608> (2020).
45. Broomé, S., Gleerup, K. B., Andersen, P. H. & Kjellström, H. in *IEEE/CVF conference on computer vision and pattern recognition*. 12667–12676.
46. McLennan, K. & Mahmoud, M. Development of an automated pain facial expression detection system for Sheep (Ovis Aries). *Anim. (Basel).* **9**. <https://doi.org/10.3390/ani9040196> (2019).
47. Andersen, P. H. et al. Towards machine recognition of facial expressions of pain in horses. *Anim. (Basel).* **11**. <https://doi.org/10.3390/ani11061643> (2021).
48. Ask, K., Rhodin, M., Rashid-Engström, M., Hernlund, E. & Andersen, P. H. Changes in the equine facial repertoire during different orthopedic pain intensities. *Sci. Rep.* **14**, 129. <https://doi.org/10.1038/s41598-023-50383-y> (2024).
49. 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**, e0228059. <https://doi.org/10.1371/journal.pone.0228059> (2020).
50. Broomé, S., Ask, K., Rashid-Engström, M., Andersen, H., Kjellström, H. & P. & Sharing pain: Using pain domain transfer for video recognition of low grade orthopedic pain in horses. *PLoS One.* **17**, e0263854. <https://doi.org/10.1371/journal.pone.0263854> (2022).
51. Broomé, S. et al. Going deeper than Tracking: A survey of computer-vision based recognition of animal pain and emotions. *Int. J. Comput. Vision.* **131**, 572–590. <https://doi.org/10.1007/s11263-022-01716-3> (2023).
52. Gleerup, K. B., Forkman, B., Lindegaard, C. & Andersen, P. H. An equine pain face. *Vet. Anaesth. Analg.* **42**, 103–114. <https://doi.org/10.1111/vaa.12212> (2015).
53. Lee, J. G. et al. Deep learning in medical imaging: General overview. *Korean J. Radiol.* **18**, 570–584. <https://doi.org/10.3348/kjr.2017.18.4.570> (2017).
54. Martvel, G., Shimshoni, I. & Zamansky, A. Automated detection of cat facial landmarks. *Int. J. Comput. Vision.* <https://doi.org/10.1007/s11263-024-02006-w> (2024).
55. Fontaine, D. et al. Artificial intelligence to evaluate postoperative pain based on facial expression recognition. *Eur. J. Pain.* **26**, 1282–1291. <https://doi.org/10.1002/ejp.1948> (2022).
56. Miller, A. L., Golledge, H. D. & Leach, M. C. The influence of isoflurane anaesthesia on the rat grimace scale. *PLoS One.* **11**, e0166652 (2016).
57. Miller, A., Kitson, G., Skalkoyannis, B. & Leach, M. The effect of isoflurane anaesthesia and buprenorphine on the mouse grimace scale and behaviour in CBA and DBA/2 mice. *Appl. Anim. Behav. Sci.* **172**, 58–62 (2015).
58. Reed, R. A. et al. Post-anesthetic CPS and EQUUS-FAP scores in surgical and non-surgical equine patients: An observational study. *Front. Pain Res. (Lausanne).* **4**, 1217034. <https://doi.org/10.3389/fpain.2023.1217034> (2023).
59. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. <https://doi.org/10.1007/BF00994018> (1995).

Acknowledgements

The research was supported by the International Veterinary Academy of Pain Management (IVAMP) (grant number: PRO00054545). The authors would like to thank the entire staff of the University of Florida Large Animal Hospital for their patience and continuous support at all stages of this work and in particular Dr. Raiane Moura for helping collecting some of the videos.

Author contributions

L.C., N.A., E.V. and D.L. acquired the data. L.C., A.G. and G.C. conceived the experiment(s). A.G. and G.C. conducted the experiment(s). L.C., G.C., C.D.G., A.N.J., D.P., M.R., E.V. and D.L. analyzed and/or interpreted the results. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.C.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024