Journal of the American Association for Laboratory Animal Science
Copyright 2024
by the American Association for Laboratory Animal Science

Vol 63, No 4
July 2024
Pages 403–411

Deep Learning for Face Detection and Pain Assessment in Japanese macaques (*Macaca fuscata*)

Vanessa N Gris, MV, DSc,[1,2,†] Thomás R Crespo, MSc,[3,†] Akihisa Kaneko, DVM,[1,2] Munehiro Okamoto, DVM, PhD,[1,2] Juri Suzuki, DVM, PhD,[1] Jun-nosuke Teramae, PhD,[3,*] and Takako Miyabe-Nishiwaki, DVM, PhD[1,2,*]

Facial expressions have increasingly been used to assess emotional states in mammals. The recognition of pain in research animals is essential for their well-being and leads to more reliable research outcomes. Automating this process could contribute to early pain diagnosis and treatment. Artificial neural networks have become a popular option for image classification tasks in recent years due to the development of deep learning. In this study, we investigated the ability of a deep learning model to detect pain in Japanese macaques based on their facial expression. Thirty to 60 min of video footage from Japanese macaques undergoing laparotomy was used in the study. Macaques were recorded undisturbed in their cages before surgery (No Pain) and one day after the surgery before scheduled analgesia (Pain). Videos were processed for facial detection and image extraction with the algorithms RetinaFace (adding a bounding box around the face for image extraction) or Mask R-CNN (contouring the face for extraction). ResNet50 used 75% of the images to train systems; the other 25% were used for testing. Test accuracy varied from 48 to 54% after box extraction. The low accuracy of classification after box extraction was likely due to the incorporation of features that were not relevant for pain (for example, background, illumination, skin color, or objects in the enclosure). However, using contour extraction, preprocessing the images, and fine-tuning, the network resulted in 64% appropriate generalization. These results suggest that Mask R-CNN can be used for facial feature extractions and that the performance of the classifying model is relatively accurate for nonannotated single-frame images.

Abbreviations and Acronyms: ANN, artificial neural network; BID, twice daily; DL, deep learning; IM, intramuscularly; NP, No Pain; P, Pain; SC, subcutaneously; ReLU, rectified linear unit; SID, once daily

DOI: 10.30802/AALAS-JAALAS-23-000056

Introduction

Facial expressions provide cues to emotions being experienced by mammals and can yield valuable information about their internal states.[17] Macaques are used extensively for research worldwide,[10,18] and negative experiences can significantly affect their physiologic, psychologic, and behavioral responses during or after an experimental procedure.[42] According to the Association of Primate Veterinarians, pain is a debilitating condition that affects an animal's quality of life and, as a consequence, may negatively impact scientific results and increase the variability of animal-based research data.[5] While legislation to enforce the ethical treatment of research animals has improved over the years, it still varies by country and relies heavily on self-regulation.[37] The ethical debate on animal experimentation and the 3Rs (Reduction, Refinement, Replacement) principle emphasizes the importance of assessing and treating pain to minimize the suffering of research animals. Recognizing pain and evaluating its severity are critical components of this ethical framework as they guide the treatment and assessment frequency for pain. However, the assessment of pain in nonhuman primates is greatly derived from anecdotal evidence due to a lack of comprehensive assessment tools.[11,35,40] Other reasons include the lack of time and resources for intensive monitoring and inherent difficulties in recognizing pain in nonverbal beings. Therefore, an evaluation method that does not require complete human management and does not increase workload is desirable.

Macaques live in societies with frequent competition among group members and may hide behaviors associated with weakness from conspecifics and potential predators. Among captive nonhuman primates, the presence of an observer has been shown to influence the spontaneous behaviors of the animals, making the animal appear to be healthier than its actual status.[21] This suggests that direct human observation may alter an animal's spontaneous actions, thereby influencing the observer's assessment of their condition. Pain evaluation may include facial expressions, as it has been reported that they can be an important indicator of pain in several species.[16,20,27,33] Methods such as grimace scales[13] or geometric morphometrics[20,22] are used to evaluate facial expressions in mammals, but they depend on a human coder. The disadvantages of having people performing this task include the need to extensively evaluate video records, the time-consuming and labor-intensive image observation or annotation,[3] the need to train observers to use the system correctly, and the inherent human bias in the evaluation process. For example, proficiency in the human facial action

Submitted: 13 Jun 2023. Revision requested: 31 Jul 2023. Accepted: 04 Jan 2024.
[1]Primate Research Institute and [2]Center for the Evolutionary Origins of Human Behavior, Kyoto University, Inuyama, Japan; and [3]Department of Advanced Mathematical Sciences, Graduate School of Informatics, Kyoto University, Kyoto, Japan

**Corresponding author. Email: miyabe.takako.2s@kyoto-u.ac.jp or teramae@acs.i.kyoto-u.ac.jp*

[†]These authors contributed equally to this study

system (FACS) requires approximately 50 to 100 h of training, and experts require about 2 h to code each minute of video.[29] Although reports of facial expressions as indicators of pain in macaques are scarce, they are helpful resources and complement other existing indicators.[16,17,22,39]

Automated recognition systems have the potential to objectively assess pain in nonverbal humans[6,7] and other animals.[8] These systems have been particularly useful for identifying pain in horses undergoing castration, allowing for efficient treatment.[2,28] Artificial neural networks (ANNs) are a branch of machine learning loosely inspired by the brain, consisting of thousands of interconnected nodes, organized into layers, that conduct information. ANNs are particularly useful in computer vision tasks, speech recognition, and medical image analysis.[1] To perform a task, ANNs typically use training examples (that is, previously identified data), such as images. In the case of object recognition, the system can be trained with thousands of labeled images (for example, house, truck, tree). By adjusting its parameters, the network can learn to match input images with corresponding output labels. Similarly, the system can be trained to recognize and classify images for the assessment of pain states in animals. However, recognizing pain from animal facial expressions, especially from macaques, is complicated by their subtle signals and potential masking behavior in the presence of observers.[21,22,39] To automate the classification of pain from facial images in macaques, videos must be recorded in the absence of observers; the macaque face must be located in the video frames and then used to train the system to perform the classification task.

The use of deep learning for image analysis provides several advantages, such as significantly reducing the need for manual image capture and annotation while also minimizing evaluation bias. This study aims to evaluate 2 models for facial recognition and frame extraction, as well as a training model for classifying pain in Japanese macaques using deep learning techniques.

Materials and Methods

Animals. A portion of the same videos described in our previous study[22] was used for the dataset. This research was approved by the Animal Welfare and Care Committee of the Primate Research Institute, Kyoto University (nos. 2016-109, 2017-096, 2018-178, 2019-156, and 2020-050), and institutional guidelines for the care and use of nonhuman primates were followed. Animals did not undergo surgery solely for this study, and video recordings were opportunistic. The study group consisted of 22 female Japanese macaques (*Macaca fuscata*), aged 9 ± 4 y and weighing 8.3 ± 1.9 kg. The macaques were captive bred at the Primate Research Institute (currently partially succeeded to the Center for the Evolutionary Origins of Human Behavior, Kyoto University) and were housed indoors at the time of the study. They were housed in singly (n = 12) or in pairs (n = 10) (650 × 1,560 × 800 mm [diameter × width × height]) in rooms with controlled temperature (20 to 27 °C) and a 12:12-h light:dark cycle (lights on at 0700). Their diet consisted of monkey chow twice daily, sweet potatoes 3 times a week, and occasional fresh apples and bananas; water was freely available. Sixteen of the macaques underwent laparotomy for reproductive biology studies. Six did not undergo surgery, and videos were collected to increase the number of images extracted for training (videos showing the absence of pain).

Macaques underwent experimental laparotomy between 2016 and 2020. Laparotomy was performed between 0900 and 1100 for egg collection or implantation. The surgical procedure involved a midline abdominal incision through the skin, fascia, and musculature, with manipulation of the uterus and ovaries based on the specific surgery. According to the survey of primate veterinarians, the degree of pain experienced by macaques after laparotomy ranges from moderate to severe.[35] Subjects were anesthetized with an IM combination of ketamine (5 mg/kg; Daiichi Sankyo Propharma, Tokyo, Japan), medetomidine (0.025 mg/kg medetomidine injection; Meiji Seika Pharma, Tokyo, Japan), and midazolam (0.125 mg/kg midazolam injection; Sandoz K.K., Tokyo, Japan). Anesthesia was maintained with sevoflurane in 100% oxygen using a face mask. The macaques also received amoxicillin (15 mg/kg Amostac; Meiji Seika Pharma), famotidine (0.1 mg/kg; Sawai Pharma, Osaka, Japan), buprenorphine (0.01 mg/kg Lepetan; Otsuka Pharmaceutical, Tokyo, Japan), and carprofen (4 mg/kg Rimadyl; Zoetis, Tokyo, Japan) during the procedure. On the morning after surgery, video recording was performed at 0800. Postoperative analgesia of buprenorphine (0.01 mg/kg, IM, BID; 0900 and 1800) and carprofen (4 mg/kg, SC, SID; 0900) was administered immediately thereafter and again on days 2 and 3 after surgery.

Face detection and frame extraction. Facial images were captured from 30 to 60 min of video footage of the macaques under 2 different conditions: before surgery (No Pain [NP]) and before receiving analgesic medication on the morning after surgery (Pain [P]), by using cameras (GoPro HERO6Black, HERO7Black, and HERO8Black) attached to the cage bars (Figure 1). The observer was not in the room during the recording session. The video recording was taken before the daily administration of analgesics to capture images at the time considered to have minimal analgesic benefit based on the pharmacokinetics of the analgesics.[36] Pain was considered to represent the most informative condition for facial pain changes, while NP was categorized as pain free. Automatic sequential video processing was performed to localize the region of the frame that contained faces to build the dataset.

Two facial location and frame extraction systems were compared: box extraction and contour extraction. For box extraction, we employed RetinaFace,[15] which localizes the face and yields a bounding box around it. For the Contour extraction, we used Mask R-CNN,[23] which marks the specific pixels in the image that belong to the face as compared with using coarse bounding boxes during object localization. Therefore, the image resulting from contour extraction is a polygon outline of the face (Figure 2).

Box extraction. In the first set of experiments, RetinaFace was used to detect and capture the macaque face in the frame. A total of 68 videos were processed, resulting in 70,852 images. The extracted images were labeled based on the macaque's condition (P/NP). Three frames per second were automatically extracted, allowing the capture of different versions of the face without intensive computation or bias. Redundant data were removed

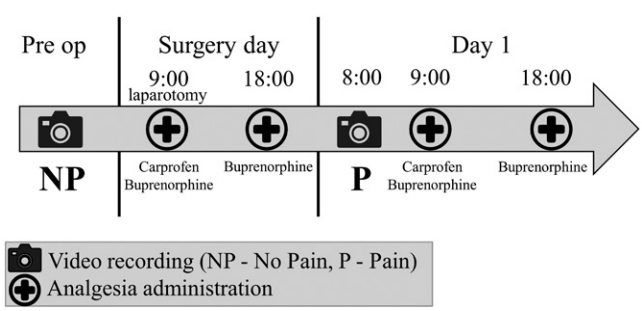

Figure 1. Timeline of video recording, surgery, and the administration of analgesics in Japanese macaques undergoing laparotomy.

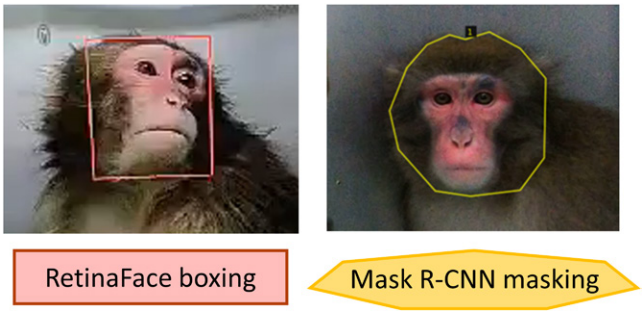

Figure 2. Extraction of facial images of Japanese macaque images using RetinaFace and Mask R-CNN. RetinaFace applies boxing around the face, while Mask R-CNN applies masking based on the object's contour.

based on a high incidence of pairwise similarities as detected by the histogram of oriented gradients (HOG) (Figure 3). HOG is based on feature descriptors, which help to extract useful information while discarding the unnecessary parts. The HOG's 0.9 threshold resulted in 15,987 images, which were then classified by the pretrained neural network ResNet50.[24] After experiments 1 and 2 (E1 and E2, described below), we manually excluded profile and blurred and occluded images or images containing elements other than the face resulting in a dataset of 11,445 pictures.

Contour extraction. In the second set of experiments, Mask R-CNN was used for object recognition and frame extraction. The same 68 videos were analyzed, resulting in 54,542 images. Masking was used to allow capture of only the face in the images, excluding any background or nearby objects. The presence of unnecessary data in a machine-learning model can be detrimental to its performance. For example, if the model uses irrelevant information, such as background and objects in the cage, to classify the images, it may mistakenly assume that all other pictures containing that background and objects belong to the same category. This can lead to inaccurate predictions and decreased accuracy. The extracted images were also converted to grayscale, brightness was equalized, and images were manually selected for suitability (Figures 4 and 5). After redundant data reduction with HOG and manual selection, the dataset comprised 19,216 images.

Neural network training. ResNet50 was used for image classification.[24] A backpropagation algorithm was used to train the multilayer networks, thereby minimizing the loss function that quantifies the difference between the model outputs and correct labels, NP or P, for images representing NP or P. To increase the amount of training data, real-time data augmentation was applied by using minor random modifications of the images, such as rotations, zoom, shifting, and horizontal flipping. We did not use predefined pain indicators, and the classification algorithm relies only on the current image being presented to the ANN.

The experiments were conducted to study the influence of 3 factors: the number of trained layers, image preprocessing, and generalization to nontrained datasets. We assessed the model's overall performance using accuracy, recall (sensitivity), precision (positive predictive value), and F1-score. The accuracy indicates the proportion of pictures correctly classified by the ANN. However, the accuracy does not indicate the degree of bias of the ANN. For example, 50% accuracy may result from classifying images as P or NP 100% of the time. Therefore, ANN performance in binary classification can be described in more detail using recall, precision, and F1-score. Recall indicates how many times the model was able to detect a specific category (that is, of all pain images, what fraction is correctly detected). Precision indicates the fraction of samples classified as pain that are truly pain images. The F1-score summarizes the precision and recall by taking their harmonic mean.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total of images}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Six experiments (E) were conducted after box extraction of the face from the videos using RetinaFace. Each experiment comprises a session of training and testing of images. The first layers of an

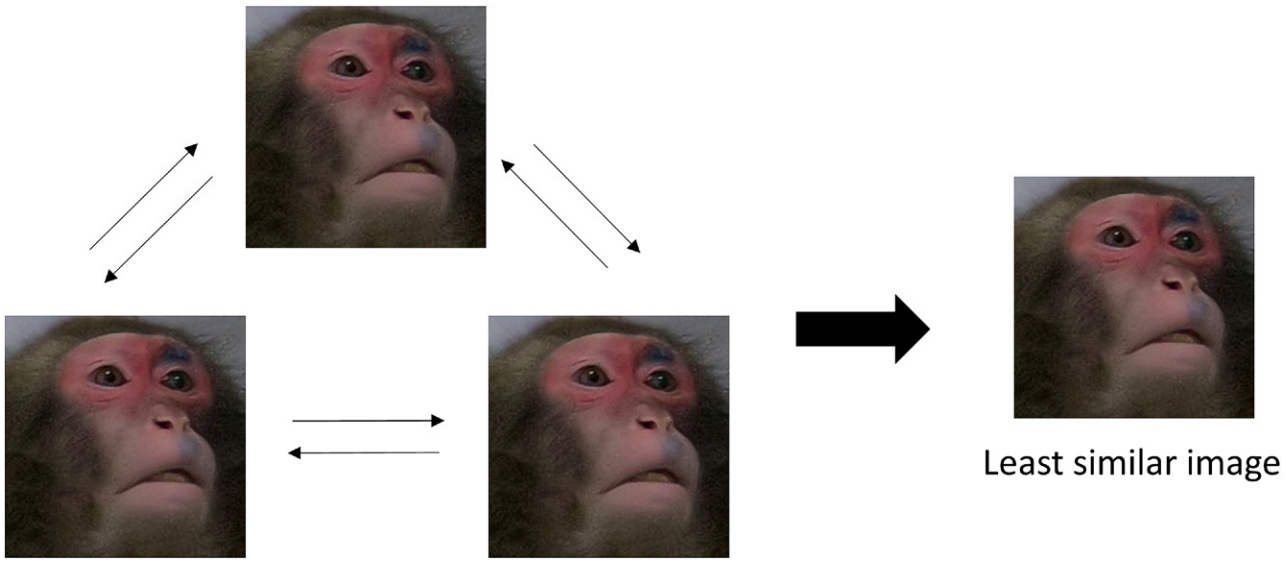

Figure 3. A histogram of oriented gradients (HOG) was used to measure the similarities between images and avoid redundant data. The extracted dataset was reduced from 70,852 images to 15,987 after HOG.

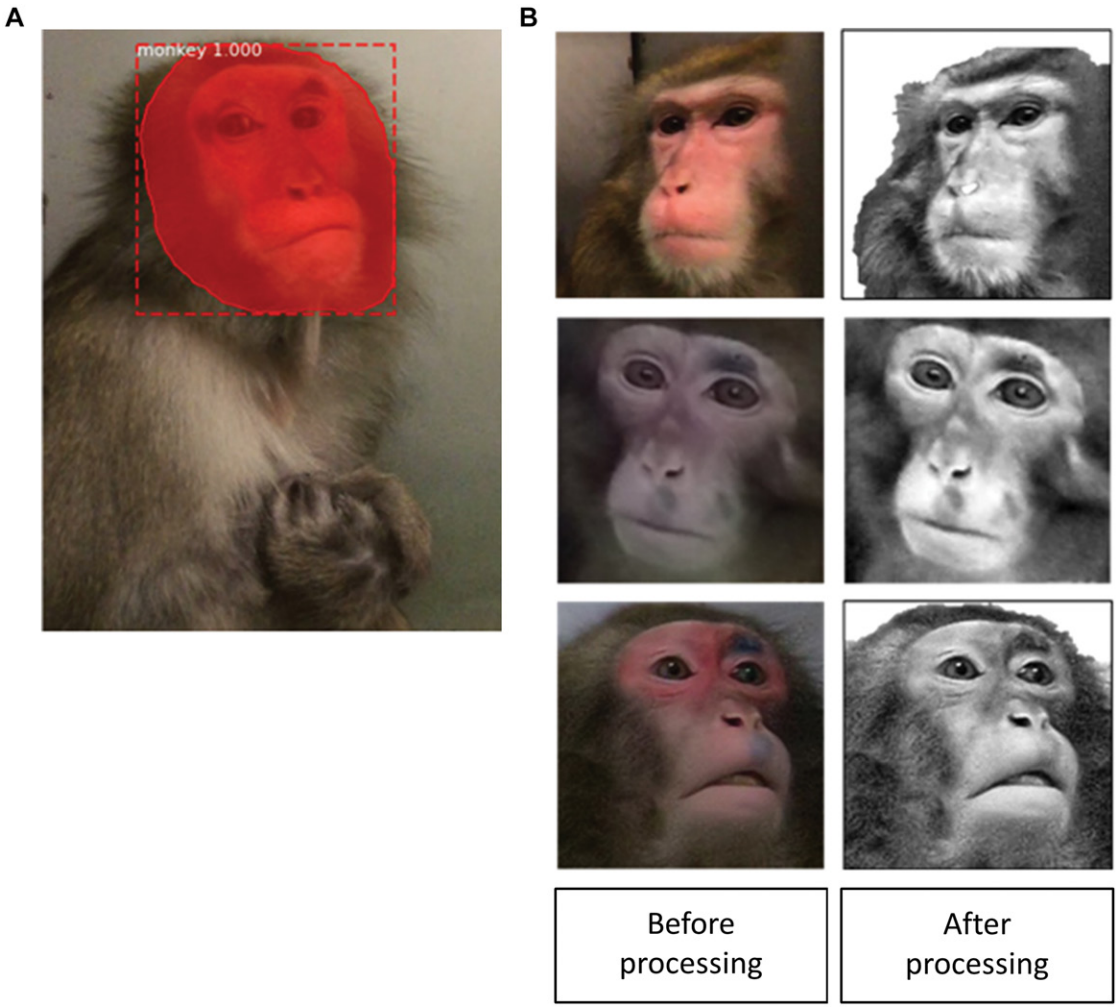

Figure 4. (A) Mask R-CNN was used to capture facial frames. (B) Images were converted to grayscale and brightness equalized to mitigate external interference.

ANN are usually not unlocked during training because they recognize the basic geometrical structures of the objects. The last layer is the one typically modified and refined to recognize a new set of classes, such as P or NP in facial expressions. We compared the number of trained layers in E1 and E2 when using the whole dataset. In E1, only the last layer (a fully connected layer with 256 ReLU units) was modified to classify images, and E2 permitted the training of all 50 layers. For E3, E4, E5, and E6, images were manually selected. We also compared the number of trained layers in E3 (only the last layer was modified) and E4 (all layers were modified). For E5 and E6, the dataset was further refined and contained only paired data (that is, training data included only images of the same individual before and after surgery). E5 and E6 also excluded the datasets of 2 animals to test the model's generalization (that is, an estimate of how well the system can classify novel data). The generalization test set comprised 1,088 images: 544 images each of P and NP classes.

E7 to E30 were conducted after masking extraction of the face from the videos using Mask R-CNN (for an outline of the study, see Figure 6). We compared the number of trained layers in E7 (the last layer modified) and E8 (all layers modified). E9 (the last layer modified) and E10 (all layers modified) used only paired data, and the datasets of 2 animals were not used for training to test the model's generalization. The goal of E11 to E30 was to improve accuracy for generalization. Therefore, we ran 20 trained ANNs using 2-stage training; this allowed fine-tuning of learning. The generalization test set contained 1,586 images (793 images each of P and NP), using the parameters shown in Table 1.

Results

Tables 2, 3, and 4 show classification performances across experiments. The results of the tests using RetinaFace for facial image capture and classification with ResNet50 are as follows: E1, which considered only the modification of the last layer of the ANN and contained all images, resulted in an accuracy of 69%. After excluding unsuitable images, E3 resulted in an accuracy of 70%. The exclusion of unsuitable images and generalization test to 2 novel macaques in E5 resulted in an accuracy of 48% (Table 2).

Results for tests using Mask R-CNN for facial image capture and classification with ResNet50 are as follows: E7, which excluded unsuitable images, resulted in an accuracy of 72%. E9, which excluded both unsuitable images and tested the generalization to 2 novel macaques, resulted in an accuracy of 55% (Table 3). Excluding unsuitable images and fine-tuning the ANN resulted in accuracy between 57% and 64%, with an average ±SD of 60 ± 2% for the generalization test to 2 novel macaques (Table 4).

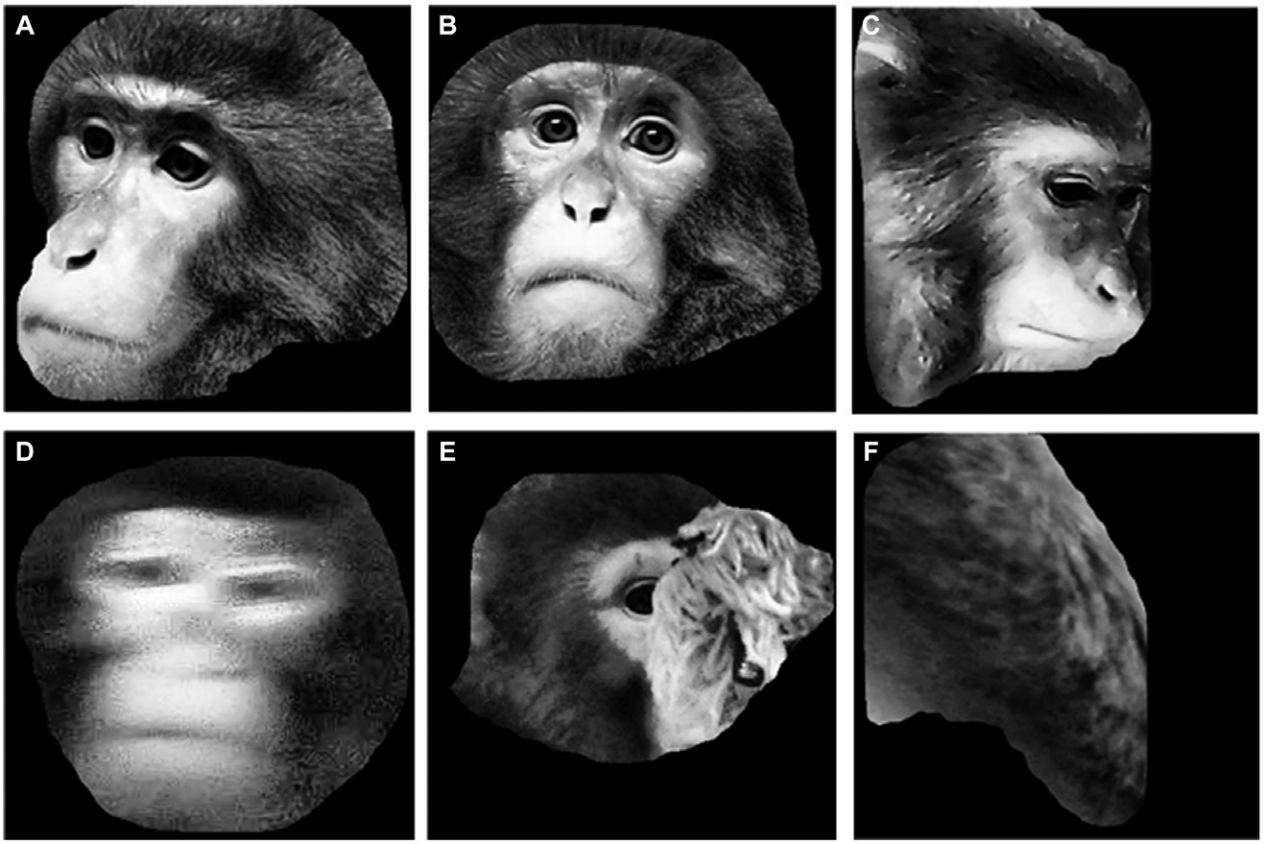

Figure 5. (A, B, C) Example of images extracted with Mask R-CNN. Images of clear faces were included. (D) Blurred images or (E) greater than 50% occluded images or (F) images showing elements other than the face were excluded from the final dataset.

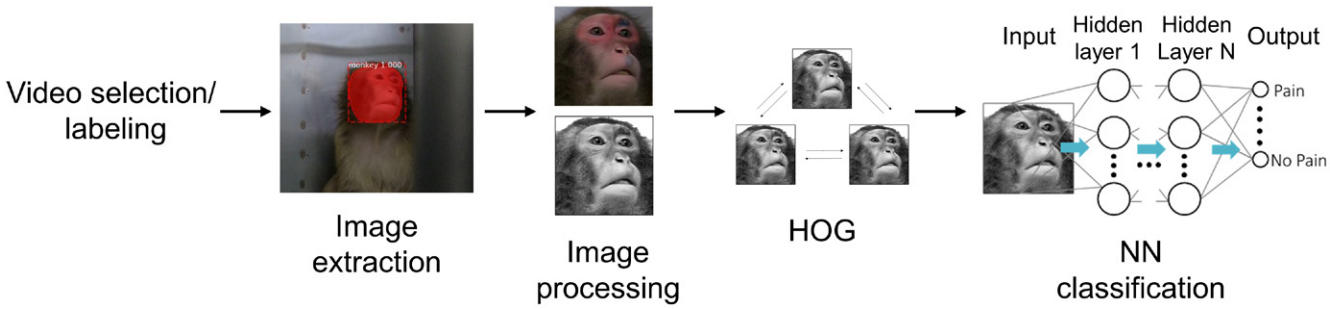

Figure 6. Flowchart of image processing for the classification of pain in facial expressions using Mask R-CNN and ResNet50.

Discussion

Machine learning techniques have been used to decode animal emotions with less risk of anthropocentric biases and comparable performance with human evaluators.[2,28,43] The current study provides information on 2 face detection methods and one ANN model (ResNet50) to classify pain in Japanese macaques without hyperparameter or architecture modification. The methods were tested to identify which performed the classification of facial expressions of pain in Japanese macaques with the greatest accuracy. Using RetinaFace for face detection and image extraction resulted in an overall test accuracy between 48% and 98%, depending on the experiment. E1 used all images extracted by RetinaFace, without manual selection, while E3 used the dataset after manual selection, which excluded profile and blurred or occluded images. Despite rigorous image exclusion, the test accuracy rose from 69% to only 70%, suggesting that the excluded images did not extensively impact the classification system. For E3, modification of all 50 layers was permitted for training, resulting in an accuracy of 94%. However, the high accuracy per se does not mean that the ANN is highly efficient in the classification task. This could result

Table 1. Architecture and hyperparameters of the neural network ResNet50

Architecture	ResNet-50 - 256 FC (0.5 dropout) - 512 FC (0.5 dropout) - 256 FC (0.5 dropout)
Hyperparameters	l2_reg = 0.0001 lr_decay = lr × sqrt(batch_size/(train_size × epochs))
Two-stage training	(decreasing the learning rate at the second stage to fine-tune)
Stage I	All layers updated, lr = 5×10^{-6} EarlyStopping (1) config: monitor = "val_accuracy," patience = 10
Stage II	All layers updated, lr = lr/100 EarlyStopping (2) config: monitor = "val_accuracy," patience = 20

Table 2. Performance of classification of Japanese macaque facial images into No Pain (NP)/Pain (P)

Experiment	Number of P images	Number of NP images	Total	Status	Accuracy (%)	Precision (%)	Recall (%)
E1–All data; the last layer modified	5,182	10,805	11,990	Train			
			3,997	Test	69	59 (NP) 82 (P)	91 (NP) 39 (P)
E2–All data; all layers modified	5,182	10,805	11,990	Train			
			3,997	Test	91	95 (NP) 97 (P)	97 (NP) 95 (P)
E3–Only suitable data; the last layer modified	3,786	7,659	8,584	Train			
			2,861	Test	70	55 (NP) 73 (P)	90 (NP) 27 (P)
E4–Only suitable data; all layers modified	3,786	7,659	8,584	Train			
			2,861	Test	94	98 (NP) 88 (P)	86 (NP) 98 (P)
E5–Only suitable data; the last layer modified; generalization	3,463	5,013	7,388	Train			
			1,088	Test	48	54 (NP) 81 (P)	95 (NP) 21 (P)
E6–Only suitable data; all layers modified; generalization	3,463	5,013	7,388	Train			
			1,088	Test	54	53 (NP) 60 (P)	84 (NP) 24 (P)

After frame extraction with RetinaFace, the classification was performed with the pretrained neural network ResNet50 on ImageNet.

Table 3. Performance of classification of Japanese macaque facial images into No Pain (NP)/Pain (P)

Experiment	Number of P images	Number of NP images	Total	Status	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
E7–Only suitable data; the last layer modified	6,172	13,044	14,022	Train				
			5,194	Test	72	75 (NP) 51 (P) **68**	82 (NP) 41 (P) **69**	79 (NP) 45 (P) **68**
E8–Only suitable data; all layers modified	6,172	13,044	14,022	Train				
			5,194	Test	97	97 (NP) 96 (P) **97**	98 (NP) 93 (P) **97**	98 (NP) 94 (P) **97**
E9–Only suitable data; last layer modified; generalization	6,172	13,044	18,833	Train				
			1,943	Test	55	58 (NP) 32 (P) **47**	90 (NP) 7 (P) **55**	70 (NP) 11 (P) **46**
E10–Only suitable data; all layers modified; generalization	6,172	13,044	18,833	Train				
			1,943	Test	44	52 (NP) 33 (P) **44**	52 (NP) 33 (P) **44**	52 (NP) 33 (P) **44**

After frame extraction with Mask R-CNN, the classification was performed with the pretrained neural network ResNet50 on ImageNet. Bolded numbers are weighted averages.

Table 4. The mean accuracy for 20 trained ANN after fine-tuning was 60% ± 2%

Experiment	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
E11	63	65 (NP) 61 (P)	57 (NP) 69 (P)	60 (NP) 65 (P)
E12	63	63 (NP) 64 (P)	64 (NP) 63 (P)	64 (NP) 63 (P)
E13	64	61 (NP) 67 (P)	73 (NP) 54 (P)	67 (NP) 60 (P)

The best 3 results are shown. Only suitable data, last layer modified, and generalization were used for these tests.

from overfitting training images, as confirmed in E6 in which generalization to novel subjects resulted in 48% accuracy. In E6, the ANN likely recognized and incorporated features that were irrelevant to pain, such as color and background, resulting in low accuracy. Tailoring the dataset by removing background and obstructing objects, converting to grayscale, and normalizing the brightness improved accuracy for E7 and E8 as compared with E3 and E4 from 69 to 72% and 94 to 97%, respectively.

The results of the generalization tests in E9 and E10 had unsatisfactory performance levels of 55 and 44%, respectively. Ensuring that training and test subjects do not overlap is crucial to avoiding classification and learning of individual-specific features by the model.[19] Tests to evaluate generalization are essential to classification systems;[46] therefore, a subset of the data not used for training was used to determine whether the model could be applied to other Japanese macaques. Our results indicate that the tests conducted after box extraction did not perform well for generalization. The dataset from box extraction included a significant amount of "noise," such as background and different illumination, that could have interfered with the

learning. Because these features are unrelated to pain, and yet this information might have been incorporated by the model, it resulted in poor generalization. Contour extraction excluded a significant proportion of these potential interferences. In this study, P is the classification of greater importance because incorrectly classifying P as NP (false negative) is worse than classifying NP as P (false positive). Therefore, P recall and precision values are important components in assessing the model. In a study on pain classification in cats using a pretrained ResNet50 network, the overall accuracy reached 72%. In the present study, the NP recall values in E9 and E10 were close to 100%, while for P was near 0%. In tests of generalization, the model will likely classify images as NP due to an unbalanced training set that has significantly more NP than P images. A model with good accuracy should be able to distinguish features from a small number of pain images. The model's performance was improved by fine-tuning, and the best model achieved a 69% recall and a 65% F1-score for P.

Even when controlling for a small number of images and an unbalanced training dataset, the classification of pain images is difficult. The facial features that indicate pain in Japanese macaques are usually subtle and vary in intensity.[22] Also, our system does not use predefined regions of the face, action unit annotation, or geometric features to indicate pain areas but learns only from full-face images and their associated labels. Therefore, we view our model results as satisfactory for this dataset. We stress that our classification algorithm relies only on the current image being presented to the ANN. However, a potential avenue for further research is to use images within a predefined time window, classify each image separately, and, if the fraction of images classified as pain exceeds a user-defined threshold, classify the image as P. This approach is similar to recent research on pain categorization based on the facial expressions of mice.[45] We hypothesize that this approach will provide higher accuracy because facial expressions change over time, and some frames are more representative of pain facial expressions than others. Furthermore, this method would prevent false positives that can occur when a brief facial expression similar to a pain expression is misclassified as P. However, this method may require larger datasets because multiple images are used for a single final classification. A more sophisticated approach to detect pain could be the use of Convolutional Long Short-Term Memory (C-LSTM) ANN, as used to detect pain in horses.[9] C-LSTM integrates both temporal and spatial information, thereby using facial expressions and behavior for the classification and outperforming convolutional neural networks (CNN) and CNN followed by an LSTM NN.[9]

Recently, automated recognition of pain has been extensively applied in horses to determine the presence[9,28] and level[28] of pain. Most efforts to identify facial indicators of pain rely on the FACS, which decomposes expressions into individual facial muscle movements or "action units" (AUs). AUs have been identified in research species and provide the anatomic foundation for the development of grimace scales and other tools used to evaluate pain.[17,25,27,43] AUs can be classified and used to train an ANN, tested either alone or together for detecting the presence and intensity of pain.[28,31,32,41] Although FACS was recently published for Japanese macaques,[14] a grimace scale has not yet been developed for this species. The classification model may benefit from specific facial features that contribute to the detection of pain in primates, such as orbital tightening, cheek tightening, and eyebrow lowering.[22,39] When detecting pain in human faces, fusing the best-performing AUs associated with pain achieved a slightly better accuracy (78%) than extracting features from the whole face (75%).[32] Focusing on specific areas of the face is likely to reflect pain more accurately and could improve classification accuracy. For example, using the Mouse Grimace Scale for mice, an automated method achieved an overall accuracy of 89% for pain classification after anesthesia and surgery.[4] In sheep, a multilevel approach with detection of faces, localization of facial landmarks, normalization, and extraction of facial features provided an overall accuracy of 67% of AUs classification.[31]

The position of the ears is a common indicator of pain in mammals.[25,27,43] In Japanese macaques; however, the ears are covered by fur, making them hard to see. The boxing and masking extraction methods were able to include the ear area, but this area probably had no significant impact on the classification results. Ears that are forward or flattened, as compared with being in a neutral position, have been associated with silent threatening and affiliative behaviors.[38,44] However, information on ear changes associated with pain has not been reported. Pain expression can vary widely among species, which complicates the extrapolation of these external cues. Currently, lip tightening and squeezed eyes are considered potential pain indicators in macaques, while ears were not found to be associated with pain.[17,39]

In addition to facial expression, behaviors are also important when evaluating pain. Smart devices have been used to record behavior patterns and activity changes associated with pain in humans[12] and animals.[47] Smartwatches and wearable sensors can provide information in real-time and facilitate the medical approach to the condition. Devices that have contact with the patient's body may be ill-suited for captive wild species and induce stress or be damaged. Therefore, video recording is still among the least expensive and most viable options for objective and continuous monitoring in captive or naturalistic scenarios. Markerless motion capture was developed from video-recorded macaques, facilitating the study of macaque behavior with accuracy comparable to that of humans.[26] In experimental surgical settings, data processing can reduce observation and training bias by monitoring the body parts that indicate the patient status.

Limitations of this study include the limited number of images for training and testing, which differs from human pain and object detection datasets that are more easily accessible in open libraries, containing many images compared with those used in animal studies. In addition, housing macaques in pairs with their conspecifics could have influenced the facial expressions of some individuals. The experience of pain can vary among individuals, and different surgeries can result in different types of pain. Sedation may also affect facial expressions and impact pain scoring, as observed in rats anesthetized with isoflurane.[34] Because our recordings began the day after the surgery, sevoflurane anesthesia was not likely to have affected the frames captured. Finally, the DL approach used in this study uses "black-box" reasoning, which means that the model's decision-making process may not be easily understood by humans, limiting its use in clinical applications.[8,30]

Assessing pain in research macaques is essential for animal welfare and helps to reduce bias in research outcomes. However, manual annotation of facial expressions and behaviors is labor- and time-intensive. Our study has shown that ANN-based algorithms can be used for automated facial recognition and classification of pain in Japanese macaques. Further studies might improve overall performance by expanding the training set, focusing on specific areas of the face, and using sequential models that consider video dynamics for classification.[9]

Acknowledgments

We thank Prof. Daniel Mills, Prof. Hirofumi Akari, Dr. Michael Huffman, and Dr. Andrew MacIntosh for their helpful comments. We also thank the editor and reviewers for their suggestions to improve the manuscript. We thank the staff of the Center for Human Evolution Modeling Research, KUPRI, for their care of the macaques.

Conflict of Interest

The authors have no competing interest to declare.

Funding

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT scholarship no. 180464) and the Leading Graduate Program in Primatology and Wildlife Science, Kyoto University. This work was partially supported by The Kusunoki 125, the Kyoto University 125th Anniversary Fund.

References

1. **Alam M, Samad MD, Vidyaratne L, Glandon A, Iftekharuddin KM.** 2020. Survey on deep neural networks in speech and vision systems. Neurocomputing **417**:302–321. https://doi.org/10.1016/0Sj.neucom.2020.07.053.
2. **Andersen PH, Broomé S, Rashid M, Lundblad J, Ask K, Li Z, Hernlund E, Rhodin M, Kjellström H.** 2021. Towards machine recognition of facial expressions of pain in horses. Animals (Basel) **11**:1643. https://doi.org/10.3390/ani11061643.
3. **Anderson DJ, Perona P.** 2014. Toward a science of computational ethology. Neuron **84**:18–31. https://doi.org/10.1016/j.neuron.2014.09.005.
4. **Andresen N, Wöllhaf M, Hohlbaum K, Lewejohann L, Hellwich O, Thöne-Reineke C, Belik V.** 2020. 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.
5. **APV.** 2019. Association of primate veterinarians' guidelines for assessment of acute pain in nonhuman primates. J Am Assoc Lab Anim Sci **58**:748–749.
6. **Bellieni CV.** 2012. Pain assessment in human fetus and infants. AAPS J **14**:456–461. https://doi.org/10.1208/s12248-012-9354-5.
7. **Birnie KA, Hundert AS, Lalloo C, Nguyen C, Stinson JN.** 2019. Recommendations for selection of self-report pain intensity measures in children and adolescents: A systematic review and quality assessment of measurement properties. Pain **160**:5–18. https://doi.org/10.1097/j.pain.0000000000001377.
8. **Broomé S, Feighelstein M, Zamansky A, Carreira Lencioni G, Haubro Andersen P, Pessanha F, Mahmoud M, Kjellstrom H, Salah AA.** 2023. 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.
9. **Broome S, Gleerup KB, Andersen PH, Kjellstrom H.** 2019. Dynamics are important for the recognition of equine pain in video. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach (CA): IEEE. https://doi.org/10.1109/CVPR.2019.01295.
10. **Carlsson H-E, Schapiro SJ, Farah I, Hau J.** 2004. Use of primates in research: A global overview. Am J Primatol **63**:225–237. https://doi.org/10.1002/ajp.20054.
11. **Carstens E, Moberg GP.** 2000. Recognizing pain and distress in laboratory animals. ILAR J **41**:62–71. https://doi.org/10.1093/ilar.41.2.62.
12. **Ciman M.** 2022. iSenseYourPain: Ubiquitous chronic pain evaluation through behavior-change analysis, p 137–149. In: Wac K, Wulfovich S, editors. Quantifying quality of life: incorporating daily life into medicine. Cham: Springer International Publishing.
13. **Cohen S, Beths T.** 2020. Grimace scores: Tools to support the identification of pain in mammals used in research. Animals (Basel) **10**:1726. https://doi.org/10.3390/ani10101726.
14. **Correia-Caeiro C, Holmes K, Miyabe-Nishiwaki T.** 2021. 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.
15. **Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S.** 2020. RetinaFace: Single-shot multi-level face localisation in the wild, p 5202–5211. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle (WA): IEEE. 10.1109/CVPR42600.2020.00525.
16. **Descovich K.** 2017. Facial expression: An under-utilised tool for the assessment of welfare in mammals. ALTEX **34**:409–429. https://doi.org/10.14573/altex.1607161.
17. **Descovich KA, Richmond SE, Leach MC, Buchanan-Smith HM, Flecknell P, Farningham DAH, Witham C, Gates MC, Vick SJ.** 2019. Opportunities for refinement in neuroscience: Indicators of wellness and post-operative pain in laboratory macaques. ALTEX **36**:535–554. https://doi.org/10.14573/altex.1811061.
18. **DiVincenti L Jr.** 2013. Analgesic use in nonhuman primates undergoing neurosurgical procedures. J Am Assoc Lab Anim Sci **52**:10–16.
19. **Feighelstein M, Shimshoni I, Finka LR, Luna SPL, Mills DS, Zamansky A.** 2022. Automated recognition of pain in cats. Sci Rep **12**:9575. https://doi.org/10.1038/s41598-022-13348-1.
20. **Finka LR, Luna SP, Brondani JT, Tzimiropoulos Y, McDonagh J, Farnworth MJ, Ruta M, Mills DS.** 2019. Geometric morphometrics for the study of facial expressions in non-human animals, using the domestic cat as an exemplar. Sci Rep **9**:9883. https://doi.org/10.1038/s41598-019-46330-5.
21. **Gaither AM, Baker KC, Gilbert MH, Blanchard JL, Liu DX, Luchins KR, Bohm RP.** 2014. Videotaped behavior as a predictor of clinical outcome in rhesus macaques (*Macaca mulatta*). Comp Med **64**:193–193.
22. **Gris VN, Broche N, Kaneko A, Okamoto M, Suzuki J, Mills DS, Miyabe-Nishiwaki T.** 2022. Investigating subtle changes in facial expression to assess acute pain in Japanese macaques. Sci Rep **12**:19675. https://doi.org/10.1038/s41598-022-23595-x.
23. **He K, Gkioxari G, Dollár P, Girshick R.** 2017. Mask R-CNN. 2017 IEEE International Conference on Computer Vision (ICCV). Venice (Italy): IEEE. https://doi.org/10.1109/ICCV.2017.322.
24. **He K, Zhang X, Ren S, Sun J.** 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas (NV): IEEE. https://doi.org/10.1109/CVPR.2016.90.
25. **Keating SCJ, Thomas AA, Flecknell PA, Leach MC.** 2012. 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.
26. **Labuguen R, Matsumoto J, Negrete SB, Nishimaru H, Nishijo H, Takada M, Go Y, Inoue KI, Shibata T.** 2021. MacaquePose: A novel "in the wild" macaque monkey pose dataset for markerless motion capture. Front Behav Neurosci **14**:581154. https://doi.org/10.3389/fnbeh.2020.581154.
27. **Langford DJ, Bailey AL, Chanda ML, Clarke SE, Drummond TE, Echols S, Glick S, Ingrao J, Klassen-Ross T, LaCroix-Fralish ML, Matsumiya L, Sorge RE, Sotocinal SG, Tabaka JM, Wong David, van den Maagdenberg AMJM, Ferrari MD, Craig KD, Mogil JS.** 2010. Coding of facial expressions of pain in the laboratory mouse. Nat Methods **7**:447–449. https://doi.org/10.1038/nmeth.1455.
28. **Lencioni GC, de Sousa RV, de Souza Sardinha EJ, Corrêa RR, Zanella AJ.** 2021. 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.
29. **Littlewort GC, Bartlett MS, Lee K.** 2009. Automatic coding of facial expressions displayed during posed and genuine pain. Image Vis Comput **27**:1797–1803. https://doi.org/10.1016/j.imavis.2008.12.010.
30. **London AJ.** 2019. Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Cent Rep **49**:15–21. https://doi.org/10.1002/hast.973.
31. **Lu Y, Mahmoud M, Robinson P.** 2017. Estimating sheep pain level using facial action unit detection. 2017 12th IEEE International

Conference on Automatic Face & Gesture Recognition (FG 2017). Washington (DC): IEEE.
32. **Lucey P, Cohn J, Lucey S, Matthews I, Sridharan S, Prkachin KM.** 2009. Automatically detecting pain using facial actions. 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam (the Netherlands): IEEE.
33. **McLennan KM, Rebelo CJB, Corke MJ, Holmes MA, Leach MC, Constantino-Casas F.** 2016. 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.
34. **Miller A, Kitson G, Skalkoyannis B, Leach M.** 2015. 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. https://doi.org/10.1016/j.applanim.2015.08.038.
35. **Miyabe-Nishiwaki T, Gris VN, Muta K, Nishimura R, Mills DS.** 2021. Primate veterinarians' knowledge and attitudes regarding pain in macaques. J Med Primatol **50**:259–269. https://doi.org/10.1111/jmp.12537.
36. **Nunamaker EA, Halliday LC, Moody DE, Fang WB, Lindeblad M, Fortman JD.** 2013. Pharmacokinetics of 2 formulations of buprenorphine in macaques (*Macaca mulatta* and *Macaca fascicularis*). J Am Assoc Lab Anim Sci **52**:48–56.
37. **Ogden BE, Pang W, Agui T, Lee BH.** 2016. Laboratory animal laws, regulations, guidelines and standards in China Mainland, Japan, and Korea. ILAR J **57**:301–311. https://doi.org/10.1093/ilar/ilw018.
38. **Partan SR.** 2002. Single and multichannel signal composition: Facial expressions and vocalizations of rhesus macaques (*Macaca mulatta*). Behaviour **139**:993–1027. https://doi.org/10.1163/15685390260337877.
39. **Paterson EA, O'Malley CI, Moody C, Vogel S, Authier S, Turner PV.** 2023. Development and validation of a cynomolgus macaque grimace scale for acute pain assessment. Sci Rep **13**:3209. https://doi.org/10.1038/s41598-023-30380-x.
40. **Paterson EA, Turner PV.** 2022. Challenges with assessing and treating pain in research primates: A focused survey and literature review. Animals (Basel) **12**:2304. https://doi.org/10.3390/ani12172304.
41. **Rashid M, Silventoinen A, Gleerup KB, Andersen PH.** 2020. 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.
42. **Reinhardt V.** 2004. Common husbandry-related variables in biomedical research with animals. Lab Anim **38**:213–235. https://doi.org/10.1258/002367704323133600.
43. **Sotocinal SG, Sorge RE, Zaloum A, Tuttle AH, Martin LJ, Wieskopf JS, Mapplebeck JC, Wei P, Zhan S, Zhang S, McDougall JJ, King OD, Mogil JS.** 2011. The rat grimace scale: A partially automated method for quantifying pain in the laboratory rat via facial expressions. Mol Pain **7**:1744-8069-7–55. https://doi.org/10.1186/1744-8069-7-55.
44. **Thierry B, Bynum EL, Baker S, Kinnaird MF, Matsumura S, Muroyama Y, O'Brien TG, Petit O, Watanbe K.** 2000. The social repertoire of Sulawesi macaques. Primate Rep **16**:203–226. https://doi.org/10.2354/psj.16.203.
45. **Tuttle AH, Molinaro MJ, Jethwa JF, Sotocinal SG, Prieto JC, Styner MA, Mogil JS, Zylka MJ.** 2018. A deep neural network to assess spontaneous pain from mouse facial expressions. Mol Pain **14**:1744806918763658. https://doi.org/10.1177/1744806918763658.
46. **Werner P, Lopez-Martinez D, Walter S, Al-Hamadi A, Gruss S, Picard RW.** 2022. Automatic recognition methods supporting pain assessment: A survey. IEEE Trans Affect Comput **13**:530–552. https://doi.org/10.1109/TAFFC.2019.2946774.
47. **Yamazaki A, Edamura K, Tanegashima K, Tomo Y, Yamamoto M, Hirao H, Seki M, Asano K.** 2020. Utility of a novel activity monitor assessing physical activities and sleep quality in cats. PLoS ONE **15**:e0236795. https://doi.org/10.1371/journal.pone.0236795.