RESEARCH ARTICLE

Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis

Niek Andresen[1☯], Manuel Wöllhaf[1☯], Katharina Hohlbaum[2☯]*, Lars Lewejohann[2,3], Olaf Hellwich[1‡], Christa Thöne-Reineke[2‡], Vitaly Belik[4‡]*

1 Department of Computer Vision & Remote Sensing, Technische Universität Berlin, Berlin, Germany, **2** Institute of Animal Welfare, Animal Behavior, and Laboratory Animal Science, Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany, **3** German Centre for the Protection of Laboratory Animals (Bf3R), German Federal Institute for Risk Assessment (BfR), Berlin, Germany, **4** System Modeling Group, Institute for Veterinary Epidemiology and Biostatistics, Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* katharina.hohlbaum@fu-berlin.de (KH); vitaly.belik@fu-berlin.de (VB)

Abstract

Assessing the well-being of an animal is hindered by the limitations of efficient communication between humans and animals. Instead of direct communication, a variety of parameters are employed to evaluate the well-being of an animal. Especially in the field of biomedical research, scientifically sound tools to assess pain, suffering, and distress for experimental animals are highly demanded due to ethical and legal reasons. For mice, the most commonly used laboratory animals, a valuable tool is the Mouse Grimace Scale (MGS), a coding system for facial expressions of pain in mice. We aim to develop a fully automated system for the surveillance of post-surgical and post-anesthetic effects in mice. Our work introduces a semi-automated pipeline as a first step towards this goal. A new data set of images of black-furred laboratory mice that were moving freely is used and provided. Images were obtained after anesthesia (with isoflurane or ketamine/xylazine combination) and surgery (castration). We deploy two pre-trained state of the art deep convolutional neural network (CNN) architectures (ResNet50 and InceptionV3) and compare to a third CNN architecture without pre-training. Depending on the particular treatment, we achieve an accuracy of up to 99% for the recognition of the absence or presence of post-surgical and/or post-anesthetic effects on the facial expression.

Introduction

The Directive 2010/63/EU stipulates to fully apply the 3-R-principle of Russel and Burch [1] (Replace, Reduce, Refine) with regard to animal experimentation. Refinement measures shall

Citation: Andresen N, Wöllhaf M, Hohlbaum K, Lewejohann L, Hellwich O, Thöne-Reineke C, et al. (2020) Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. PLoS ONE 15(4): e0228059. https://doi.org/10.1371/journal.pone.0228059

Editor: Chang-Qing Gao, Central South University, The Third Xiang Ya Hospital, CHINA

Received: August 8, 2019

Accepted: January 6, 2020

Published: April 15, 2020

Data Availability Statement: Data underlying the results presented in the study will be available from: www.scienceofintelligence.de/research/data/black-mice.

Funding: This study was supported by the German Research Foundation and the Open Access Publication Fund of the Universität Berlin: Deutsche Forschungsgemeinschaft (German Research Foundation) - 390523135 [Andresen], Deutsche Forschungsgemeinschaft (German Research

Foundation) - 390523135 [Wöllhaf], Deutsche Forschungsgemeinschaft (German Research Foundation) - 390523135 [Hohlbaum], German Federal Ministry of Education and Research (grant number: 031A262A) [Hohlbaum], Deutsche Forschungsgemeinschaft (German Research Foundation) - 390523135 [Lewejohann], Deutsche Forschungsgemeinschaft (German Research Foundation) - 390523135 [Hellwich], Deutsche Forschungsgemeinschaft (German Research Foundation) - 390523135 [Thöne-Reineke], and German Federal Ministry of Education and Research (grant number: 031A262A) [Thöne-Reineke].

Competing interests: The authors have declared that no competing interests exist.

be applied to foster the well-being of the laboratory animals. Refinement measures aim to minimize pain, suffering, and distress accompanying the experiment and actively promote well-being. Moreover, refinement significantly contributes to the quality of research findings [2]. Therefore, in the scope of animal welfare and good science, it is crucial to develop scientifically sound tools which help to systematically evaluate and improve the well-being of laboratory animals.

Impaired well-being of animals can be assessed by behavioral, biochemical, physiological, and physical parameters [3]. In recent years, methods to analyze the facial expressions of pain, analogous to the facial action coding system (FACS) for humans [4], were developed for various animal species, e.g., mice, rats, rabbits, cats, horses, cows, sheep, and piglets [5–14]. The so-called Grimace Scales are thought to measure the presence or absence of grimacing associated with pain. Especially for mice, the most commonly used laboratory animals, the Grimace Scale became a valuable tool and was applied in pain research as well as cage-side clinical assessment [15–17]. However, prerequisite to applying the Mouse Grimace Scale (MGS) is that a person is present and either generates live scores or acquires images or videos to be scored retrospectively. Thus, during periods in which the animals are not monitored, the well-being of a mouse cannot be assessed. Another important aspect is the fact that mice are prey animals and often hide signs of weakness, injury, and pain in the presence of humans [18]. Therefore, it is decisive to find a way to automatically monitor well-being of a mouse in the absence of humans. Since the facial expression has proven to be useful in mice, it is worthwhile to develop an automated facial expression recognition software. Automation can also bring substantial time savings for the MGS application as it is labor intensive to train persons in using the MGS and to manually generate MGS scores. MGS training including group discussion with a trainer is essential and should not be skipped because it improves inter-rater reliability [19].

In the field of automatic facial expression recognition in animals, we are only at the very beginning. In contrast, a lot of advances have already be made in automated facial expression recognition in humans [20]. Recently, machine learning techniques have experienced a revival, mostly due to the progress in deep learning (multi-layered neural networks combined with advanced optimization techniques) [21]. They allow classification and predictions on the data without *a priori* feature design. Statistical physics recently provided evidence for their unexpected effectivity [22]. Deep learning is used in different fields of research, relating to image recognition, from urban dynamic [23, 24] to general tracking of humans and animals [25], which is important in neuroscience, behavioral biology and digital pathology [26]. These networks are successfully used in many applications involving the acquisition of information about features of an image that traditionally only humans or animals could perceive. Two influential examples are AlexNet [27] and R-CNN [28]. A number of highly efficient computational frameworks for deep learning and neural networks were recently released, e.g., TensorFlow, Theano, and Caffe. They are easy to deploy and make the technology accessible for a vast audience of researchers [29]. Furthermore, these libraries are equipped with a range of pre-trained models allowing the efficient learning of high level concepts using transfer learning strategies.

The goal of our study was to develop an automated facial expression recognition software for mice that is able to assess whether the well-being of a mouse is impaired by post-anesthetic and/or post-surgical effects. Therefore, we use a pipeline with two components. First, a detector that locates and extracts a face of a mouse in the processed image and secondly, a deep neural network model for the classification of facial expressions. We used an own image data set of adult female and male black-furred mice of the strain C57BL/6JRj which were either untreated or received anesthesia (with isoflurane or ketamine/xylazine combination) with or without surgery (castration) [30, 31].

In line with the 3-R-principle the data comes from previous experiments. This means that the proposed method has to work on the same data that has sufficient quality for human processing, but is not unified or otherwise optimized for automatic processing.

Materials and methods

Ethics statement

In the present study, we reused images from mice which were obtained in previous animal experiments [30, 31]. Animal experimentation was performed according to the guidelines of the German Animal Welfare Act and the Directive 2010/63/EU for the protection of animals used for scientific purposes. Maintenance of mice and all animal experimentation were approved by the Berlin State Authority ("Landesamt für Gesundheit und Soziales", permit number: G0053/15).

Animals

Images of mice were obtained from previous studies performed at Freie Universität Berlin (Berlin, Germany) using 61 female and 65 male adult C57BL/6JRj mice purchased from Janvier Labs (Saint-Berthevin Cedex, France) [30, 31]. At the age of 10–13 weeks mice either underwent inhalation anesthesia with isoflurane (26 male and 26 female C57BL/6JRj mice) [30], injection anesthesia with the combination of ketamine and xylazine (26 male and 22 female C57BL/6JRj mice) [31], or no treatment (13 male and 13 female C57BL/6JRj mice). Additionally, 19 male C57BL/6JRj mice at the age of 18–42 weeks, which had previously received injection anesthesia or no treatment, were castrated (isoflurane anesthesia, meloxicam, lidocaine, and prilocaine) at the age of 18–42 weeks in order to be re-socialized in groups of 3–4 animals.

Mice were housed in a conventional (non-SPF) facility. The mice were free of all viral, bacterial, and parasitic pathogens listed in the FELASA recommendations [32]. Female mice were group-housed with 3–5 mice in Makrolon type IV cages (55 × 33 × 20 cm). Male mice had to be single-housed in Makrolon type III cages (42 × 26 × 15 cm) due to aggressive behavior toward conspecifics. The cages contained fine wooden bedding material (LIGNOCEL® 3–4 S, J. Rettenmaier & Söhne GmbH + Co. KG, Rosenberg, Germany) and nest material (nestlets: Ancare, UK agents, Lillico, United Kingdom; additionally, cocoons were provided for castrated mice: ZOONLAB GmbH, Castrop-Rauxel, Germany). A red plastic house (length: 100 mm, width: 90 mm, height: 55 mm; ZOONLAB GmbH, Castrop-Rauxel, Germany) and metal tunnels (length: 125 mm, diameter: 50 mm; one tunnel in Makrolon type III cages, two tunnels in Makrolon type IV cages) were provided as cage enrichment. The animals were maintained under standard conditions (room temperature: 22 ± 2 °C; relative humidity: 55 ± 10%) on a light:dark cycle of 12:12 h of artificial light with a 5 min twilight transition phase (lights on from 6:00 a.m. to 6:00 p.m.). The mice were fed pelleted mouse diet ad libitum (Ssniff rat/mouse maintenance, Spezialdiäten GmbH, Soest, Germany) and had free access to tap water. Both the technician and veterinarian were female. Combined tunnel and cup handling were used, i.e., the mice were carefully caught in a tunnel and then transferred to the hand, in order to minimize handling induced stress and anxiety [33]. After the study, females as well as castrated males were re-homed and intact male mice were used for educational purposes.

Animal experimental procedures

Inhalation anesthesia with isoflurane. Inhalation anesthesia was induced with 4% isoflurane (Isofluran CP®, CP-Pharma Handelsgesellschaft mbH, Burgdorf, Germany) in 100% oxygen in an anesthetic chamber (with sliding cover, Evonik Plexiglas, 240 × 140 × 120 mm). The chamber was not prefilled with isoflurane. After the mouse had lost the righting reflex, it was transferred to a heating pad and anesthesia was maintained with 1.75–2.5% isoflurane in 100% oxygen via nose cone. Body temperature was measured using a rectal probe during anesthesia and the heating pad temperature, that could be adjusted from 30-42˚C, was increased or decreased accordingly to keep body temperature constant. Artificial tears (Artelac® Splash MDO®, Bausch & Lomb GmbH, Berlin, Germany) were administered to both eyes to prevent the eyes from drying out. Anesthesia lasted for approximately 45 minutes. During anesthesia pedal withdrawal and lid reflex were regularly tested and vital parameters (i.e., respiratory rate, heart rate, and oxygen saturation) were monitored. According to the design of our previous study, mice were anesthetized either once or six times at an interval of three to four days [30].

Injection anesthesia with the combination of ketamine and xylazine. For injection anesthesia a stock solution with 160 μL Ketavet® 100 mg/mL (Zoetis Deutschland GmbH, Berlin, Germany), 160 μL Rompun® 2% (Bayer Vital GmbH, Leverkusen, Germany), and 1680 μL physiologic saline solution was prepared in a syringe. A dosage of 80 mg/kg ketamine and 16 mg/kg xylazine [34], warmed to body temperature, was administered intraperitoneally at a volume of 10 μL/g body weight using 27 $\frac{3}{4}$ Gauge needles. Then the mouse was transferred to a Makrolon Typ III cage placed on a heating pad. After the loss of righting reflex, the mouse was transferred to a heating pad. Body temperature was measured using a rectal probe during anesthesia and the heating pad temperature, that could be adjusted from 30-42˚C, was increased or decreased accordingly to keep body temperature constant. Artificial tears (Artelac® Splash MDO®, Bausch & Lomb GmbH, Berlin, Germany) were administered to both eyes to protect them from drying out. Overall duration of anesthesia was 60–84 minutes. Pedal withdrawal and lid reflex were regularly tested and vital parameters (i.e., respiratory rate, heart rate, and oxygen saturation) were monitored during anesthesia. According to the design of our previous study, mice were anesthetized either once or six times at an interval of three to four days [31].

Castration. Mice were anesthetized with isoflurane in 100% oxygen (induction: 4% isoflurane in an induction chamber, maintenance: 1.5–2.5% isoflurane via nose cone, 17 min mean duration of anesthesia). After anesthesia induction, meloxicam (1 mg/kg body weight, s.c.; Metacam 2 mg/ml Injektionslösung für Katzen, Boehringer Ingelheim Vetmedica GmbH, Ingelheim, Germany) was subcutaneously administered and lidocaine/prilocaine ointment (Emla Creme, AstraZeneca GmbH, Wedel, Germany: 1 g contained 25 mg lidocaine and 25 mg prilocaine) was applied to the scrotum. During anesthesia mice lay on a heating pad, whose temperature could be adjusted from 30-42˚C. Five minutes after anesthesia induction, surgical castration was performed in supine position according to Behrens et al. [35]. The testicles were removed one by one. In brief, testicles were pushed down into the scrotal sacs and an 1 cm incision was made through the skin at a right angle to the midline of the scrotal sac. After a testicle was pushed out, the *vas deferens* with the blood vessels running along it was ligated with absorbable suture (3–0) and the testicle was removed. When both testicles were removed, the skin was stitched with a single button suture. Recovery from surgery was systematically monitored using a clinical score sheet and further indicators of well-being such the MGS, nesting and burrowing behaviour, body weight, and analysis of fecal stress hormone metabolites. Except for two out of 19 mice, these parameters suggested that the mice may had already recovered from surgery on day two after castration. Both mice received an additional treatment with 1 mg/kg meloxicam s.c. on day one. In a case report, that has not been published yet, we critically discuss the analgesic regime used (1mg/kg meloxicam s.c., EMLA cream) since especially MGS scores indicated that inappropriate analgesia was achieved. In short: The onset and duration of action of the EMLA cream when applied to the scrotum of mice is not known. For preventing rabbits from pain during ear tattooing, a contact time of 20 min is recommended [7]. Due to differences in the thickness and permeability of the mouse scrotum and rabbit ear, the onset may be more rapid in mice. Apart from the use of EMLA cream, preemptive analgesia could be provided by the application of analgesics via the drinking water in the dark phase prior to surgery. Analgesia may also be improved by higher doses of meloxicam or other non-steroidal anti-inflammatory drugs (NSAIDs) than those currently recommended [36, 37] or a combination of buprenorphine and a NSAID [38].

No treatment. Mice of the control groups in our previous studies received neither anesthesia nor surgery (i.e., no treatment), hence no post-anesthetic/surgical effects on their facial expressions were expected.

Dataset

The present study is based on a large data set of images of C57BL/6JRj mice. Images were obtained in previous studies, in which the impact of procedures frequently performed in animal experimentation on the well-being and stress levels of mice was systematically assessed by using the MGS, among other animal-based parameters [30, 31]. According to the treatment of the mice, our data set was divided into the three subsets: KXN (ketamine/xylazine anesthesia), IN (inhalation anesthesia with isoflurane), and C (castration).

Image acquisition. Images were generated as described previously in Hohlbaum et al. [39]. All images were taken in observation cages (22 × 29 × 39 cm) (Fig 1) with three white walls to contrast the black mice and one clear wall. Cages were custom-made in our facility and visually varied slightly, e.g., walls were attached to each other with dark- or light-colored material. The bottom of the cage was covered with approximately 0.5 cm bedding material and soiled bedding was scattered on top in order to minimize stress caused by the novel environment. Food pellets normally supplied as diet and a water bowl were provided.

After the procedures were performed, the mouse was gently transferred into an observation cage and was allowed to habituate to the new environment for 30 min. Then a series of images

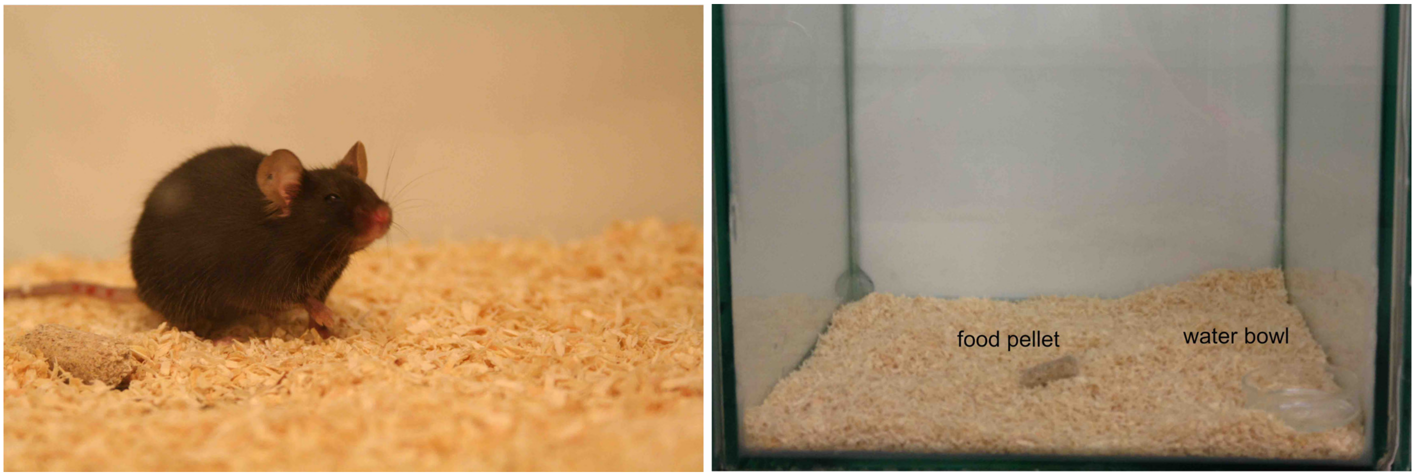


Fig 1. Image of the data set and observation cage. Example image of a black-furred laboratory mouse (C57BL/6JRj strain) of the dataset (left). Observation cage used for monitoring the mice after the procedures (right). Images of the mice were taken when mice were moving freely around in the observation cage.

https://doi.org/10.1371/journal.pone.0228059.g001

was taken within a few minutes (approximately 1-2 minutes, but in some cases even longer) using a high definition camera (Canon EOS 350D, Canon Inc., Tokyo, Japan). Baseline images were acquired prior to the procedures. Post-procedure images were taken at various points in time as follows:

- Inhalation anesthesia with isoflurane [30]: 30 min, 150 min post-procedure
- Injection anesthesia with the combination of ketamine and xylazine [31]: 30 min, 150 min, day 2, day 9 post-procedure
- Castration: 30 min, 150 min, 300 min, day 2, day 3, day 7 post-procedure.

Images of untreated mice were generated at the same times as images of corresponding treated groups. Since they were considered not to show any post-anesthetic or post-surgical distress, they were added to baseline images in the present study.

Human evaluation of mouse images. In order to systematically evaluate the post-anesthetic and/or post-surgical effects on the facial expression of the mice, images were scored by humans on the MGS. The MGS developed by Langford et al. [5] measures characteristic changes in facial expressions of pain in mice. The five facial action units of the MGS (i.e., orbital tightening, nose bulge, cheek bulge, ear position, and whisker change) are scored on a 3-point-scale (0 = not present, 1 = moderately visible, 2 = obviously present) with high scores reflecting high intensity of a facial action unit (for further details see Langford et al. [5]). An accuracy of 72–97% for humans scorers was reported [5]. Interestingly, the change of facial action units described in the MGS also derive from other etiologies than pain such as post-anesthetic distress, situations associated with fear (i.e., social proximity, cat odor exposure, rat exposure), and aggression or subordination [30, 31, 40]. However, a closer investigation of these facial expressions may reveal slight differences [41]. Moreover, they may be differentiated by considering other behavior and context of the situation [41].

For MGS scoring, one image of high quality showing the mouse face from frontal or lateral view was randomly selected per mouse, and point in time [30, 31, 39] and the face of the mouse were manually cropped from the image so that, if possible, the body posture was not visible. Overall, five persons (three veterinarians, laboratory assistant, and secretary) with different experience levels in laboratory animal science were trained in scoring images for MGS. A minimum of two and a maximum of four individuals (IN: three, KXN: four, C: two) were recruited for MGS scoring. Images were presented in a randomized order on a computer screen to the uninformed individuals, who independently scored the five facial action units of the MGS using the manual provided by Langford et al. [5].

Binary labels. Due to the time and labor intensive process of MGS scoring, only a small amount of images (658 images) in our data set is scored on the MGS. This number does not allow a successful deep learning based regression analysis of the function, which maps from image (pixel) space to MGS score. As we could not predict the MGS score directly, we followed the approach of Tuttle et al. [42] and assigned one of two defined states, “post-anesthetic/surgical effect” and “no post-anesthetic/surgical effect”, to all images in order to train a binary classifier on the whole data set. To do so, time points of image acquisition were either defined as “post-anesthetic/surgical effect” or “no post-anesthetic/surgical effect” based on the MGS scores obtained by humans and the statistical analysis performed in our previous studies [30, 31, 39]: if MGS scores were significantly higher when compared to untreated mice, all images of this point in time were considered to display facial expressions affected by the treatment and were assigned to the label “post-anesthetic/surgical effect” (Fig 2, Table 1). Furthermore, 300 min post-castration and 30 min following inhalation anesthesia with isoflurane in males were also labeled with “post-anesthetic/surgical effect”, though significance versus baseline

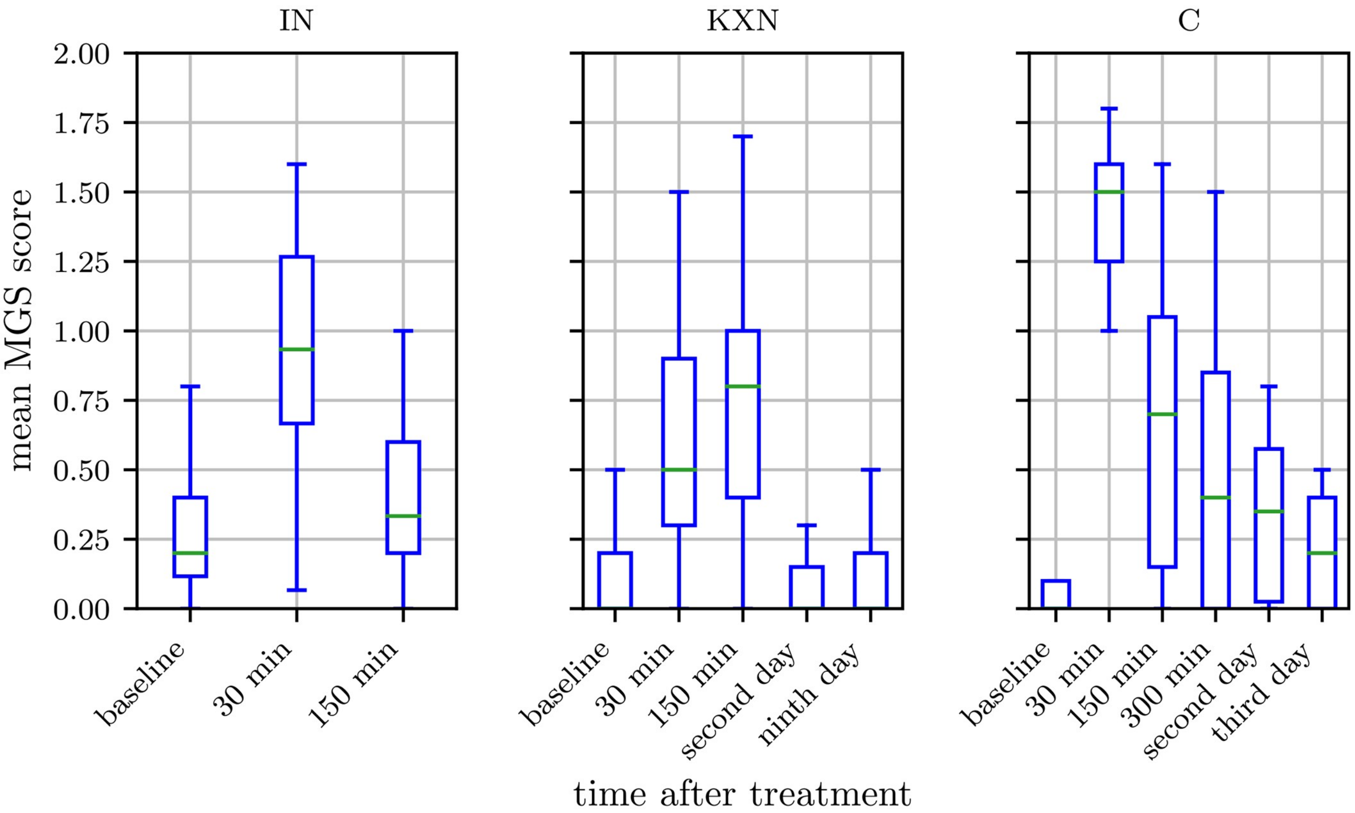


Fig 2. Box plots of the human evaluated Mouse Grimace Scale. Isoflurane anesthesia (IN, left), ketamine/xylazine anesthesia (KXN, middle) and castration (C, right). IN: Scores were obtained from 30 female and 31 male C57BL/6JRj mice. KXN: Scores were obtained from 29 female and 32 male C57BL/6JRj mice. C: Scores were obtained from 19 male C57BL/6JRj mice. Data represent the mean MGS scores averaged over three human scorers. The box represents the interquartile range (IQR), box edges are the 25th and 75th percentile. The whiskers represent values which are no greater than 1.5 × IQR. Outliers were excluded from the figure. This figure contains data from Hohlbaum et al. [30, 31].

https://doi.org/10.1371/journal.pone.0228059.g002

was not reached. Images taken at other points in time as well as images of untreated mice were assigned to the label "no post-anesthetic/surgical effect". Individual MGS scores could not be considered, i.e., an image may have been labeled with "post-anesthetic/surgical effect" even though the image shows a mouse with a low MGS score.

Since mice were repeatedly exposed to the observation cage, they may have habituated to this environment and, subsequently, have experienced less stress, which could have had a slight impact on the facial expression. However, this effect was neglected because these slight changes can not be determined with current facial expression analysis methods and are subject of our future work. Moreover, the binary labels do not allow a finer subdivision.

Face detection. The first step of the evaluation pipeline is to distinguish between valid and invalid images. Valid images are those which contain a mouse face including at least one

Table 1. Points in time labelled with "post-anesthetic/surgical effect" are listed for each procedure. Baseline images and images acquired at a later time were assigned to the label "no post-anesthetic/surgical effect".

Procedure	Time points assigned the label "post-anesthetic/surgical effect"
Isoflurane anesthesia	30 min post-procedure
Ketamine/xylazine anesthesia	30 min, 150 min post-procedure
Castration	30 min, 150 min, 300 min post-procedure

https://doi.org/10.1371/journal.pone.0228059.t001 of the relevant features (eyes, ears, whiskers, nose, cheek) for facial expression analysis. Invalid images do not show a face or are of low quality (i.e., blurry). As not only the mouse but also parts of the observation cage are shown in the images, it was furthermore necessary to crop the relevant area of an image. This facilitates the learning of a mapping between image and label as a greater fraction of the input dimensions (pixels) of the model correlate with the target. Therefore, an automatic face detector was applied.

Sotocinal et al. [6] created the "Rodent Face Finder" for white rats which was used in Tuttle et al. [42]. It combines two detectors, one for ears and one for eyes. Groups of detections are filtered according to heuristic expectations of a typical face (e.g., the ears must be above the eyes). Unfortunately, the trained model, provided by the authors of the "Rodent Face Finder", led to a detection rate of zero, when applied to our data set. This is probably due to the different fur color of the mice in the two data sets. The training of a similar detector on our data set combined with the heuristics from Sotocinal et al. resulted in a detection (face hypothesis) for 40% of all images while about 86% of all images are valid. We suppose that the "Rodent Face Finder" has a similarly low detection rate on white mice, but for the application on video sequences, as in Sotocinal et al. [6] and Tuttle et al. [42], this is not a hindrance as a face needs only to be successfully detected in a fraction of the acquired frames for a correct classification. If a frame of the face is grabbed from a video every few seconds, the classification can be performed accurately. However, to achieve a bigger resulting training set for the facial expression recognition on our data set of still images, we decided to use a face detector with higher sensitivity and lower precision. As in Sotocinal et al., we used a detector based on Boosted Cascades of Simple Features as introduced in Viola and Jones [43]. We used an implementation from the Computer Vision library OpenCV [44]. In contrast to the "Rodent Face Finder", it was trained to detect the whole mouse face. Producing a higher false positive rate, this resulted in face hypotheses for about 80% of the images. An additional application of the eye and ear detector on the remaining 20% of images led to face hypotheses for 92% of the images. We applied these detectors to all 32576 images of 61 female and 65 male C57BL/6JRj mice and automatically cropped the images based on the detected bounding boxes around the mouse faces. Afterwards, we manually sorted out false positives and images of low quality, which resulted in a remaining set of 18273 images (13352 from KXN, 2470 from C, and 2451 from IN; 60 female and 64 male C57BL/6JRj mice) (this is the resulting set we provide under www.scienceofintelligence.de/research/data/black-mice). We discuss a possible replacement of this step for a fully automated system in the Outlook section.

Facial expression recognition

As in Tuttle et al. [42], we trained a binary classifier to distinguish between images labeled with "post-anesthetic/surgical effect" or "no post-anesthetic/surgical effect". The binary classifier was a single layer fully connected neural network with softmax activations on top of a ResNet50 or InceptionV3 architecture. Note that in the study Tuttle et al. [42] only the InceptionV3 architecture was used. The weights of the ResNet50 and InceptionV3 networks were pre-trained on ImageNet [45] and frozen during training of the top layer. While this is common practice, we want to mention that it is also possible to fine tune the already pre-trained layers of the network. However we decided to classify the representations generated with the convolutional base of the architecture without retraining, as at least some of the evaluated subsets are too small and would lead to overfitting of the network parameters to the training data [46, 47]. The models were chosen due to their current success. The InceptionV3 [48] model was tested on the ILSVRC 2012 [49] data set and achieved 21.2% top-1 error rate. The ResNet50 [50] achieved 20.1% top-1 error rate, which corresponds to the current state of the

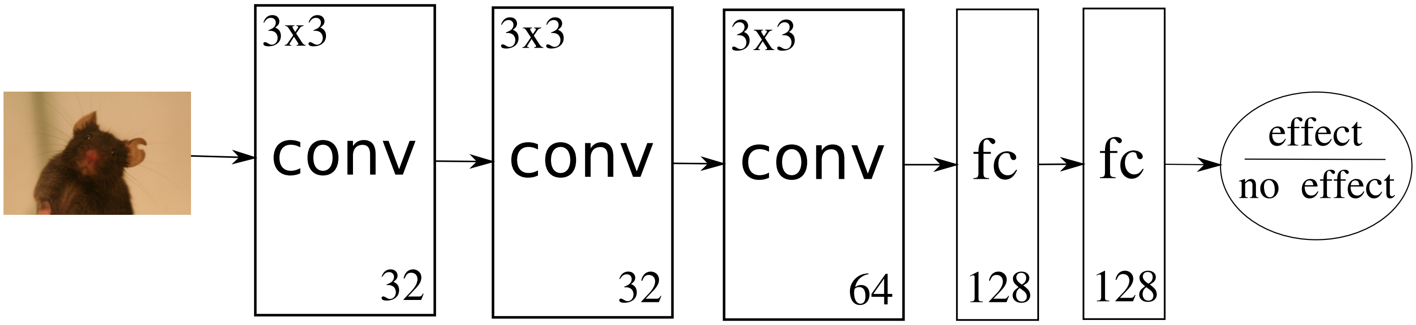


Fig 3. Own network architecture. The image is fed through three convolutional layers with filter size 3x3 and 32, 32 and 64 filters. Two fully connected layers with 128 neurons each follow. The two output neurons give a confidence in either judgment (post-anesthetic/surgical effect or not).

https://doi.org/10.1371/journal.pone.0228059.g003

art in image-based object recognition. Additionally, we compared to a simpler architecture without pre-training (Fig 3). For training we used the Adam optimizer (parameters for Adam optimizer: learning rate 0.001, beta1 0.9, beta2 0.999, epsilon 1e-07, decay 0) to minimize the categorical cross entropy loss. All networks were implemented and trained using TensorFlow [51]. The output of the top layer of the network are the activations of two output neurons, one for each class. We classify an image as "post-anesthetic/surgical effect" if the activation of the "post-anesthetic/surgical effect" neuron is greater than the activation of the "no post-anesthetic/surgical effect" neuron and vice versa. Furthermore, we interpret the resulting activation of the two output neurons as the confidence of the network for the two classes "post-anesthetic/surgical effect" and "no post-anesthetic/surgical effect".

The network was trained with batch size 100 until it converged after 50 epochs using a new random permutation of the training set for each epoch. The data set was split in training and test set with no subject overlap for most experiments. In this context, the term subject stands for mouse. This strict separation ensures that no specific features of an individual are used for classification. It furthermore avoids the possibility of images with high similarity (taken at the same time of the same mouse) in training and test set. As 19 male mice which received ketamine/xylazine anesthesia or no treatment were reused for castration, some experiments have overlapping subjects. However, high similarity of images is ruled out here as images were acquired under different conditions (respective results are marked with superscript SO). To be able to train on the pre-computed representations of the convolutional base of the used pre-trained models and as the top layer do not overfit the data, we applied no data augmentation. This leads to a reduction of the training time to under one minute per experiment on the used hardware (NVIDIA GeForce GTX1080) which allowed us to evaluate a great number of experiments and use cross-validation for evaluation of the results. In future work additional data augmentation could lead to further improvement of the reported results. The data sets were balanced by sub-sampling. All images were resized to 224 × 224 using bilinear transform.

Results and discussion

Binary classification

First we evaluated the performance of the trained classifier, once on a combined set of all images and furthermore on the three subsets KXN, IN and C separately. We performed 10-fold cross-validation for the combined set and on the subsets KXN and IN and leave-one-animal-out-cross-validation for the subset C, which contains only 19 animals in total (Table 2a, 2b and 2c). The split into training and test set resulted in 112 animals for training and 12 animals for testing over all treatments, in 55 animals for training and 6 animals for

Table 2. Cross-validation results. Data are given as mean percent ± standard deviation. IN: isoflurane anesthesia; KXN: ketamine/xylazine anesthesia; C: castration; TPR: true positive rate (sensitivity); TNR: true negative rate (specificity).

(**a**) ResNet.				(**b**) Inception.			
	accuracy	TPR	TNR		accuracy	TPR	TNR
All	88.5 ±2.2	85.8 ±4.4	90.8 ±2.6	All	83.8 ±3.1	82.1 ±4.7	85.1 ±5.7
KXN	93.4 ±1.9	93.4 ±4.2	92.4 ±4.4	KXN	90.0 ±2.4	86.3 ±6.1	92.1 ±3.1
C	89.5 ±3.6	89.6 ±9.0	89.2 ±5.7	C	78.0 ±8.9	74.8 ±17.1	81.2 ±10.0
IN	75.6 ±4.8	74.4 ±7.9	76.6 ±6.3	IN	72.4 ±4.7	67.4 ±11.6	75.4 ±9.1
(**c**) Own architecture.				(**d**) ResNet. Averaging over multiple images.			
	accuracy	TPR	TNR		accuracy	TPR	TNR
All	82.9 ±4.2	81.3 ±7.7	83.3 ±5.0	All	92.5 ±2.4	86.3 ±7.2	95.8 ±2.0
KXN	88.1 ±2.5	87.1 ±4.7	90.4 ±4.5	KXN	98.9 ±1.9	97.2 ±4.7	100.0 ±0.0
C	83.5 ±3.0	82.3 ±5.7	84.3 ±5.1	C	89.8 ±5.5	98.1 ±7.6	85.3 ±9.2
IN	56.5 ±4.7	62.8 ±12.2	55.8 ±13.8	IN	90.1 ±5.7	93.5 ±10.0	89.0 ±9.5

https://doi.org/10.1371/journal.pone.0228059.t002

testing for the KXN subset and in 57 animals for training and 6 animals for testing for the IN subset.

Although the InceptionV3 architecture was used in previous works [42], we achieved highest values for accuracy (= $^{t}p + tn/_{t}p + tn + fp + fn$), sensitivity (= $^{t}p/_{t}p + fn$), and specificity (= $^{t}n/_{t}n + fp$) for the combined data sets on the ResNet architecture. However, if we evaluated the data sets separately, performance differed for KXN, C, and IN. While performance improved for KXN as well as C, it deteriorated for IN. To understand why results were lower for IN, we have to consider the design of the study from which images of the IN subset were obtained. Mice of this study received inhalation anesthesia with isoflurane, which caused statistically significant changes in the facial expression for a relatively short period only [30]. Therefore, images taken 30 min post-anesthesia were considered to display post-anesthetic effects and images generated at baseline or 150 min post-anesthesia were labeled with "no post-anesthetic/surgical effect". While our previous study did not reveal any statistically difference in the facial expressions according to the MGS between treated and untreated mice at 150 min post-anesthesia [30], MGS scores were still slightly increased in some treated animals at 150 min post-anesthesia. As a consequence, the binary classification led to a pool of "no post-anesthetic/surgical effect" images with a high range of intensities (Figs 2 and 4). The "no post-anesthetic/surgical effect" and "post-anesthetic/surgical effect" classes include MGS scores of 0.20 (median; interquartile range: 0.28) and 0.67 (median; interquartile range: 0.73), respectively. The resulting smaller margin in feature space for this treatment would probably require a classifier with higher complexity compared to the other treatments.

As Tuttle et al. [42] showed that results can be improved by using multiple frames from video data for classification, we too averaged the network confidence over all images that were taken at the same point in time (i.e., have same time label) of the same animal and then classified the images based on the average confidence (Table 2d). As the images were taken from different perspectives the additional information here is higher than in the use of video data where the frames potentially contain a lot of redundant information. The results show increased performance for all subsets, especially IN. For the subset KXN, specificity was 100 ±0% (mean ± standard deviation). This high result may be explained by a clear decision boundary between the two classes in the KXN subset. Injection anesthesia is known to intensively impair the general condition of a mouse and to significantly affect its facial expression for a longer period (i.e., up to at least 150 min) [31]. Therefore, images generated 30 min as well as 150 min post-anesthesia were labeled with "post-anesthetic/surgical effect" and images

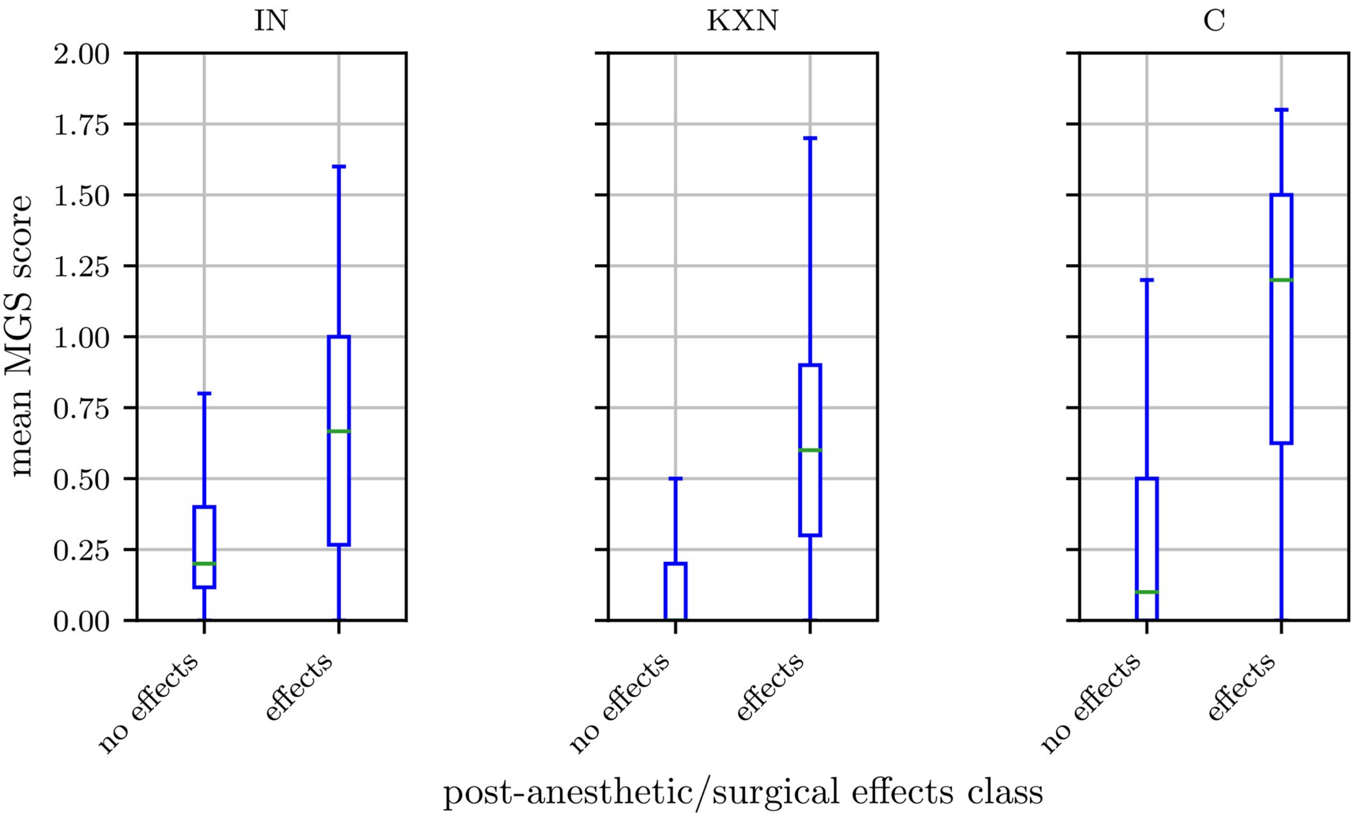


Fig 4. Ground truth distribution of the "post-anesthetic/surgical effect" and "no post-anesthetic/surgical effect" class. Isoflurane anesthesia (IN, left), ketamine/xylazine anesthesia (KXN, middle), and castration (C, right). The box represents the interquartile range (IQR), box edges are the 25th and 75th percentile. The whiskers represent values which are no greater than 1.5 × IQR. Outliers were excluded from the figure.

https://doi.org/10.1371/journal.pone.0228059.g004

of the remaining time points were assigned the label "no post-anesthetic/surgical effect". In contrast to the subset IN, the range of intensities of facial expressions in the "no post-anesthetic/surgical effect" class is smaller in the subset KXN (Figs 2 and 4). MGS scores of the "no post-anesthetic/surgical effect" class (median: 0.00, interquartile range: 0.2) and "post-anesthetic/surgical effect" class (median: 0.60, interquartile range: 0.60) overlap to lesser extent than seen for IN, which may have contributed to a clearer decision boundary between the two classes.

The reported accuracy is based on the assumption, that the ground truth data is accurate in itself. A measurement of the accuracy of the label information is not possible in our case due to the way we infer it (see section Binary Labels). Furthermore, the performance of the algorithms in binary classification of images labeled with "post-anesthetic/surgical effect" versus "no post-anesthetic/surgical effect" cannot be directly compared to the human performance within our data set as we used the MGS scores obtained by humans to generate the ground truth classes. Langford et al. [5] reported an accuracy of 97% and 81% for experienced and inexperienced human scorers, respectively, when high resolution images (1,920 × 1,080 pixels) were used. A lower accuracy of 72% was found for inexperienced humans who scored low resolution images (640 × 480 pixels) [5]. This underlines the importance of both image quality and experience in the use of the MGS for the application of this method by humans. Images of our data set originally had higher resolution (3456 × 2304 pixels) when rated by human scorers, but were resized to 224 × 224 for the present study in order to reduce input dimensionality of the learner. All tested architectures showed higher accuracy on the combined data set of all images (i.e., including IN, KXN, and C) when compared to inexperience humans using low or high resolution images. Similar success to experienced human scorers was achieved by ResNet for the subset KXN when a classification was based on all available images of an animal at a certain point in time (Table 2d).

Network confidence over time

In Figs 5, 6 and 7, we present the network confidence values for the "post-anesthetic/surgical effect" class of ResNet architecture and the human evaluated MGS scores over time for the subset IN, KXN, and C, respectively. In Tuttle et al. [42], the correlation analysis between the confidence of the classifier and the MGS score suggests, that an estimate of the expression intensity can potentially be inferred without the necessity of a regression model trained on MGS score labels. Since the intersection of images scored on the MGS and tested images in our data set is too small, a correlation analysis like in Tuttle et al. could not be carried out.

However, in general, the data suggests that the confidence for the class "post-anesthetic/surgical effect" was higher for images with high MGS scores. This can be clearly seen for 30 min and 150 min post-anesthesia with ketamine/xylazine combination (Fig 6) or 30 min post-castration (Fig 7). The other way around, a very low confidence for the class "post-anesthetic/surgical effect" was found for images with low MGS scores, indicating "no post-anesthetic/

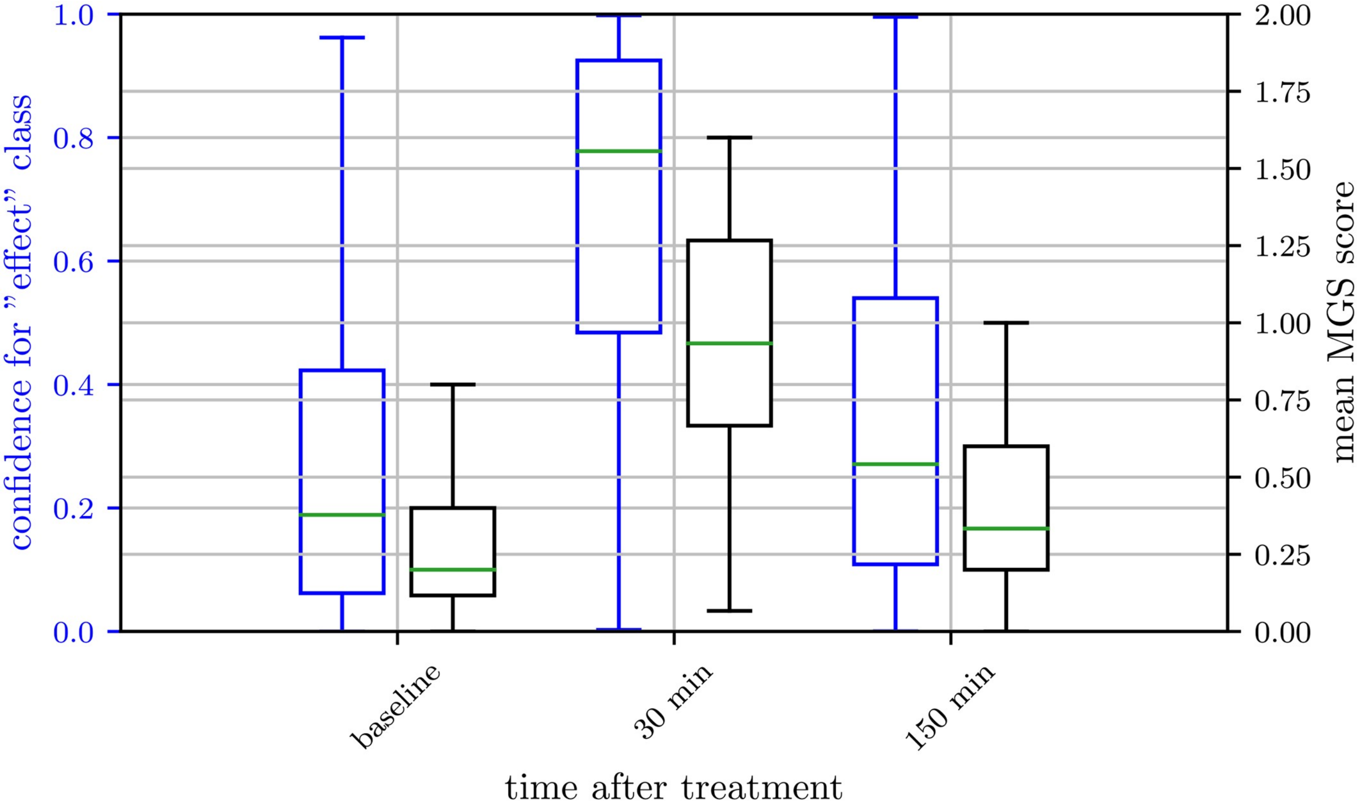


Fig 5. Network confidence over time for isoflurane anesthesia. Box plots of human labeled Mouse Grimace Scale (MGS) scores (grey) and confidence for "post-anesthetic/surgical effect" class of ResNet architecture (blue) for isoflurane anesthesia (IN). Scores were obtained from 33 female and 32 male C57BL/6JRj mice. MGS data represent the mean MGS scores averaged over three human scorers. The box represents the interquartile range (IQR), box edges are the 25th and 75th percentile. The whiskers represent values which are no greater than 1.5 × IQR. Outliers were excluded from the figure. This figure contains data from Hohlbaum et al. [30].

https://doi.org/10.1371/journal.pone.0228059.g005

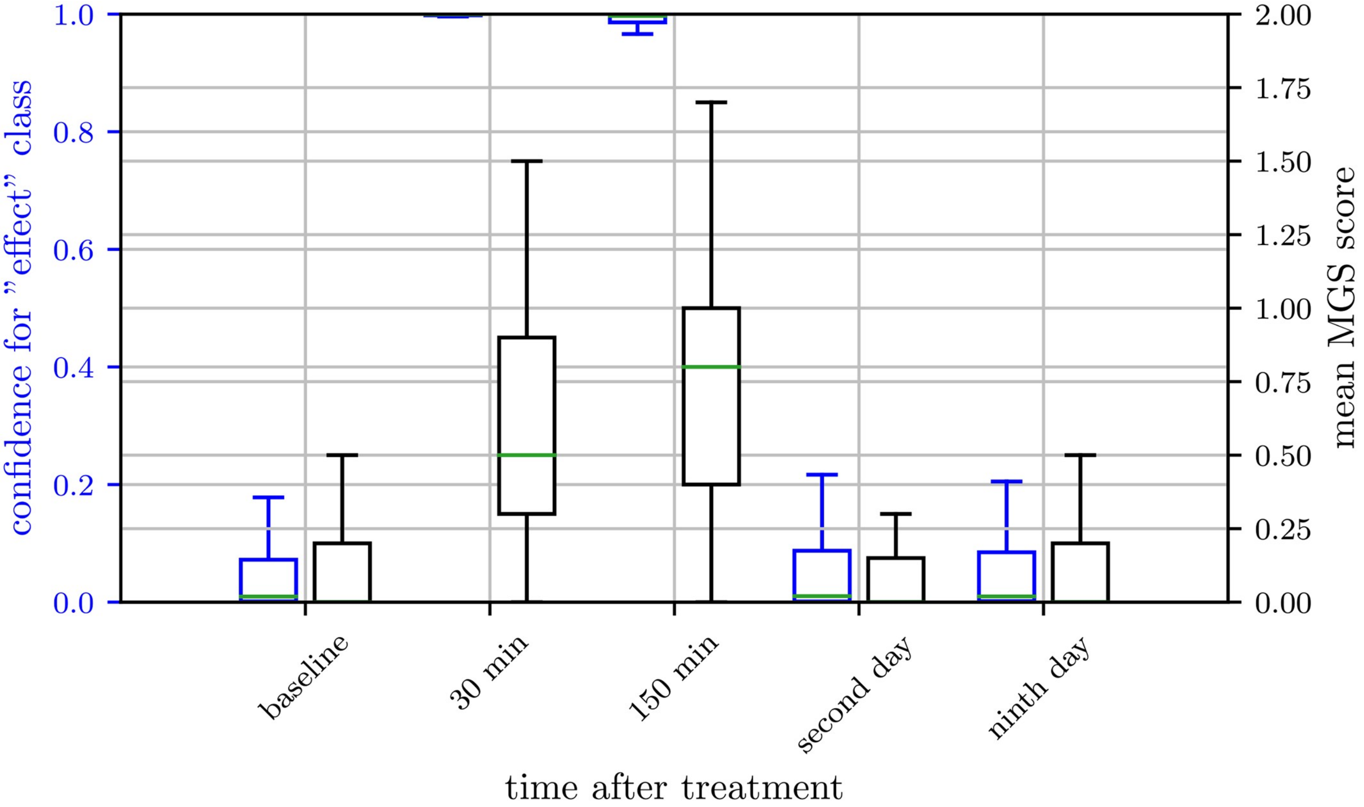


Fig 6. Network confidence over time for ketamine/xylazine anesthesia. Box plots of human labeled Mouse Grimace Scale (MGS) score (grey) and confidence for "post-anesthetic/surgical effect" class of ResNet architecture (blue) for ketamine/xylazine anesthesia (KXN). Scores were obtained from 28 female and 30 male C57BL/6JRj mice. MGS data represent mean MGS scores averaged over four human scorers. The box represents the interquartile range (IQR), box edges are the 25th and 75th percentile. The whiskers represent values which are no greater than 1.5 × IQR. Outliers were excluded from the figure. This figure contains data from Hohlbaum et al. [31].

https://doi.org/10.1371/journal.pone.0228059.g006

surgical effect". In intermediate cases however, while the network output still tends to follow the human evaluated MGS score, the deviation of network confidence and MGS score increases (e.g. 150 min and 300 min post-castration).

Regarding inhalation anesthesia with isoflurane, the confidence for the class "post-anesthetic/surgical effect" reflects the difficult decision boundary between the two classes "post-anesthetic/surgical effect" and "no post-anesthetic/surgical effect" in this data subset (Fig 5), as discussed above. In brief, both classes contain images within a relatively big, shared range of the MGS scores, what makes it more difficult for the algorithm to distinguish between the two classes.

Cross-treatment evaluation

To determine if there is a universal "post-anesthetic/surgical effect" on the facial expression independent of treatment, we performed cross-treatment analysis of our algorithms. In Fig 8 the accuracies of the ResNet50 and InceptionV3 architectures are presented. Although their absolute performance is different, the relative performance for various combinations of training and test data sets seems to be similar. For a fair comparison these results are based on a subset of the data available for KXN with the same size as the C and IN sets. It was noticeable that the performance of both models decreased in nearly all cases when they were trained and tested on mouse images of different treatments. To explain this, we developed two hypotheses

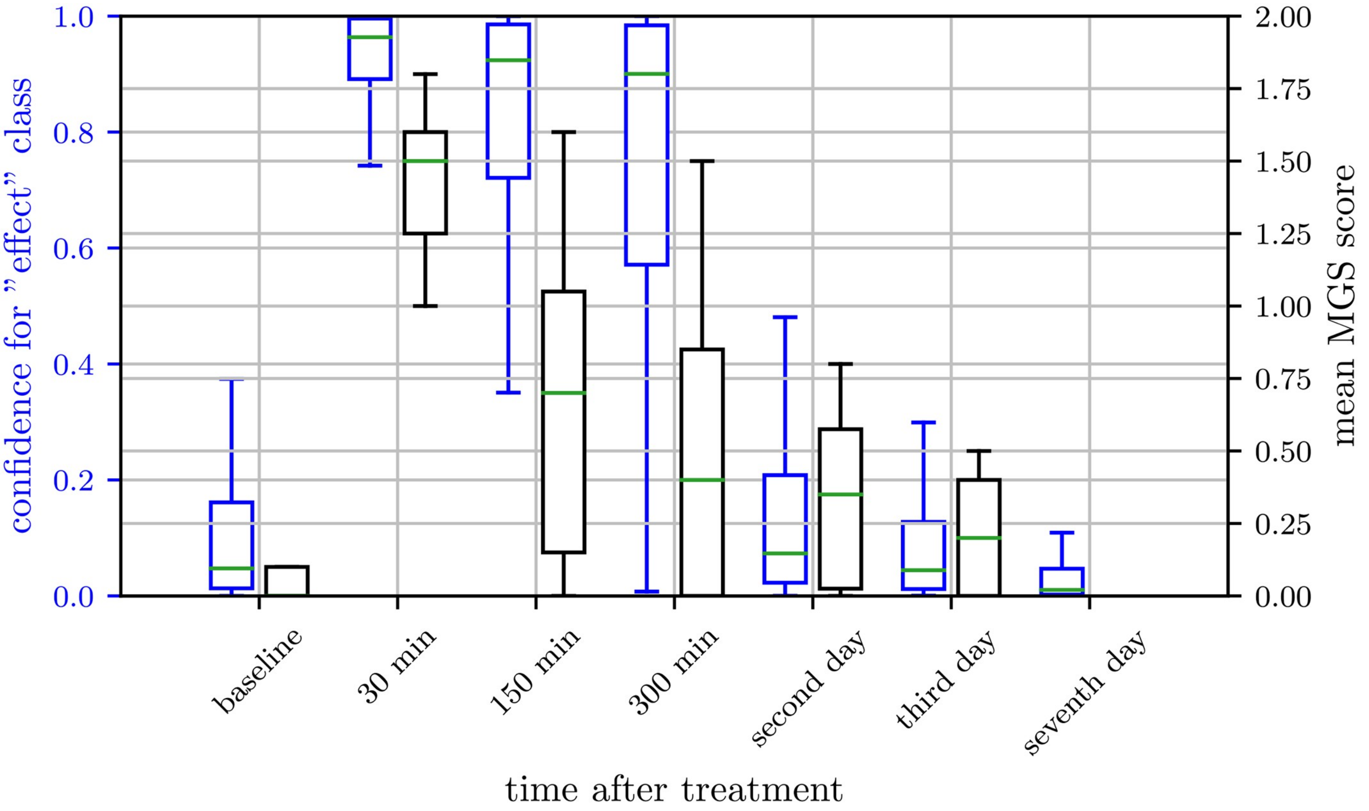
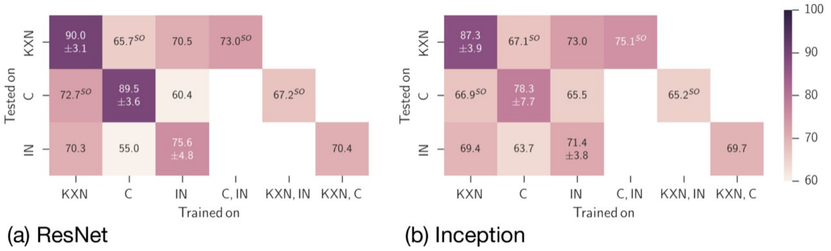


Fig 7. Network confidence over time for castration. Box plots of human labeled Mouse Grimace Scale (MGS) score (grey) and confidence for "post-anesthetic/ surgical effect" class of ResNet architecture (blue) for castration (C). Scores were obtained from 19 male C57BL/6JRj mice. MGS data represent the mean MGS scores averaged over two human scorers. The box represents the interquartile range (IQR), box edges are the 25th and 75th percentile. The whiskers represent values which are no greater than 1.5 × IQR. Outliers were excluded from the figure.

https://doi.org/10.1371/journal.pone.0228059.g007

regarding the dimensionality of the expression space. First, the learner may learn different decision boundaries for different distributions of facial expression intensities in different treatments and therefore performs poorly on the other subsets. This might be counteracted by a better design of the label space. Secondly, the expression space is inherently multidimensional and the learned interpretation cannot be transferred between different treatments. This is harder to overcome but would potentially allow to make higher dimensional estimates for the



(a) ResNet

(b) Inception

Fig 8. Performance for different combination of training and test datasets after 50 epoch of training. These results are based on a subset of the data available for KXN. The subset has the same size as the K and IN sets and allows a fair comparison of the values. Data are given as mean accuracy (± standard deviation) in %. IN: isoflurane anesthesia; KXN: ketamine/xylazine anesthesia; C: castration; SO: subject overlap (the term subject is used as a synonym for mouse).

https://doi.org/10.1371/journal.pone.0228059.g008

state of the animal. Treatments underlying our data set, i.e., inhalation anesthesia, injection anesthesia, and castration, may induce different facial expressions. This hypothesis is supported by the effects of these treatments on the general condition of a mouse. In general, surgery including inhalation anesthesia and analgesic treatment, depending on analgesia management, causes a higher impairment of well-being than inhalation anesthesia only, which can be assessed by behavioral parameters, for instance nest building [52]. Depending on the treatment, a mouse is exposed to different stimuli and experiences different states. Post-surgical pain accompanies castration, whereas anesthesia is unlikely to induce pain. However, the injection with the combination of ketamine and xylazine can damage the tissue at the injection site, which may be associated with a painful sensation [53], but we assume the degree of pain to be much lower when compared to post-surgical pain. Anesthesia induces post-anesthetic distress with a variety of causes. Isoflurane has a pungent odor [54] and induces irritant effects in the airways by activating nociceptive ion channels [55, 56]. In humans, the inhalation of isoflurane causes coughing and subjective sensations of burning as well as irritation [57]. In addition, if 100% oxygen is used as carrier gas, the inhalation gas is very dry and can impair the function of the respiratory mucosa [58]. When anesthesia is induced, distress of a mouse can additionally increase by fixation and injection stress or by exposure to the (irritant) volatile anesthetic. In the latter case, aversion towards this inhalant agent elevates with repeated exposure [59]. Distress a mouse experienced during the induction phase may influence its well-being after anesthesia as well. When mice recover from anesthesia, they can suffer from post-anesthetic nausea [60]. Moreover, in humans emergence delirium can occur during the recovery period and hallucinogenic effects were reported for the use of ketamine [54, 61]. We also have to consider different pharmacological effects of anesthetics on the facial expressions of the mice with longer lasting effects following injection anesthesia due to the pharmacokinetic properties of ketamine and xylazine. In contrast to isoflurane, ketamine and xylazine are subject of an intensive liver metabolism [62–65], which results in longer recovery periods. Ketamine increases the muscle tone, whereas the combination of ketamine and xylazine causes muscle relaxation [54]. With regard to Miller et al. [16] who did not find any changes in MGS scores when CBA mice were treated with meloxicam (5mg/kg s.c.) 24 hours after vasectomy, pharmacological effects of meloxicam on the facial expressions are not expected in the present study, though cannot be excluded due to the lack of a control group treated with meloxicam only.

All in all, the treatments inhalation anesthesia, injection anesthesia, and castration are accompanied by different pharmacological effects and produce different affective states in a mouse. Against the background that the weight of the five facial action units varies between different states like illness and pain [5, 66], our data suggests that the procedures we investigated in the present study may induce different facial expressions. This may explain the reduction in performance when the algorithms were trained and tested on mouse images of different treatments and is a disadvantage for the pure binary classification of "post-anesthetic/surgical effect" and "no post-anesthetic/surgical effect".

Feature importance visualization

To ensure that the predictions of the networks are based on the facial expressions of the mice and not on other high level (e.g., food pellets, ear punches used as markers) or low level (e.g., illumination) features that possibly correlate with the target ("post-anesthetic/surgical effect"/ "no post-anesthetic/surgical effect") and to get insights into which features the neural network uses for classification, we performed a decomposition analysis. Methods as the deep Taylor decomposition [67–69], which we used in this work, propagate the activations of the neurons

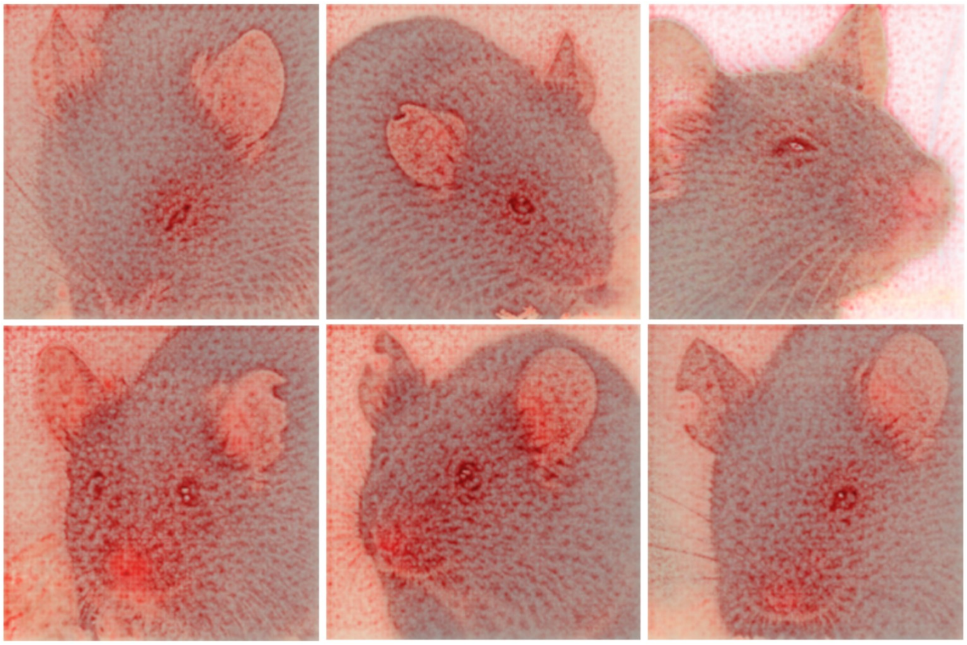


Fig 9. Visualization of the decision finding process using deep Taylor decomposition. Castration (left), ketamine/xylazine anesthesia (middle), and isoflurane anesthesia (right). Mice correctly classified with "post-anesthetic/surgical effect" in top row, mice correctly classified with "no post-anesthetic/surgical effect" in bottom row. Red color indicates that a pixel contributes to the decision.

https://doi.org/10.1371/journal.pone.0228059.g009

backwards through the network and decompose the prediction into contributions of single pixels. These contributions can be visualized as heat maps to explain the decision of a network regarding a certain image. From Fig 9 we clearly see, that the decision is made using mostly features related to the mouse itself and not to the background. The ears and especially the outline of the ears, the eyes and the area around the eyes, the nose tip and the area around the nose (i.e., probably the whisker pad), as well as whiskers or the space between the whiskers are contributing to the decision making by the neural network. Thus, the heatmaps confirm the importance of at least some facial action units of the MGS [5], especially orbital tightening, ear position and whisker change. Additionally, further facial features such as the nose tip and the whisker pad appear to have a significant impact on the decision making process, particularly in mouse images labeled with "no post-anesthetic/surgical effect" (Fig 10). To understand why the nose tip seems to play an important role in the state not affected by the treatment, the color of the nose and its position are of special interest. While the nose points forwards or is slightly elevated in general when no post-anesthetic and/or post-surgical effects are present, the head rather is dropped and the nose tip points downwards after anesthesia or surgery. The nose tip is colored (pale) pink in mice in good general condition. If mice recover from anesthesia or surgery, circulation can be affected in the early post-anesthetic period, hence the color of the nose may turn paler. Overall, the nose tip may play a more important role in the absence of post-anesthetic and/or post-surgical effects because it is clearly visible and the color is very prominent. Figs 10 and 9 (bottom row) reveal that, except from the nose tip, the area around the nose, probably the whisker pad, can also be critical for the decision making process. The importance of this feature may be traced back to the muscles associated with the vibrissal follicles [70], which cause a change in whisker position. A natural downward curve of the whiskers is found in the absence of post-anesthetic and/or post-surgical effects, whereas whiskers stiffen

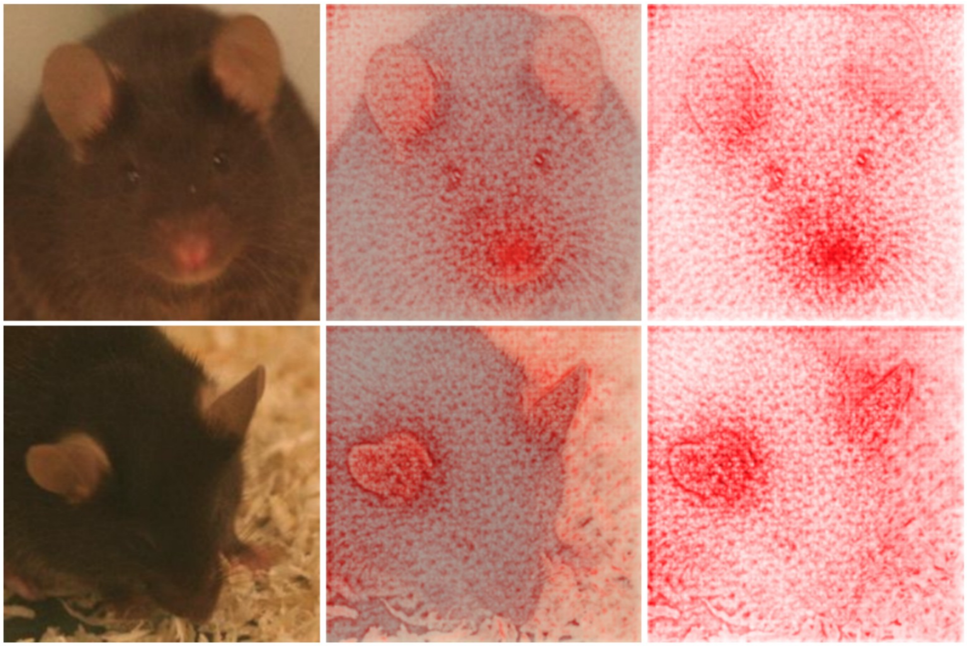


Fig 10. Contribution of nose, whisker pad, and ears to the decision making. Visualization of the decision finding process using deep Taylor decomposition for images generated 30 min (top row) and 2 days (bottom row) after ketamine/xylazine anesthesia. Original image (left), original image combined with heat map (middle), heat map (right). Red color indicates that a pixel contributes to the decision. Top row: The mice was correctly classified as "no post-anesthetic/surgical effect" with a confidence of 96,2%. In particular the nose and the whisker pad seem to contribute to the decision. Bottom row: The mice was correctly classified as "post-anesthetic/surgical effect" with a confidence of 100,0%. The decision appears to be mainly based on the ears.

https://doi.org/10.1371/journal.pone.0228059.g010

and are pulled back or forward, they may also clump together in the presence of those effects. Besides the whisker position, the activity of these muscles may influence the appearance of the whisker pad, which would explain the usefulness of this feature for the decision making. Another relevant feature we detected by visualization was the space between the whiskers, which is obviously influenced by the position of the whiskers and may have the same meaning accordingly (Fig 11). However, the network uses not only pixels of the mouse face but also

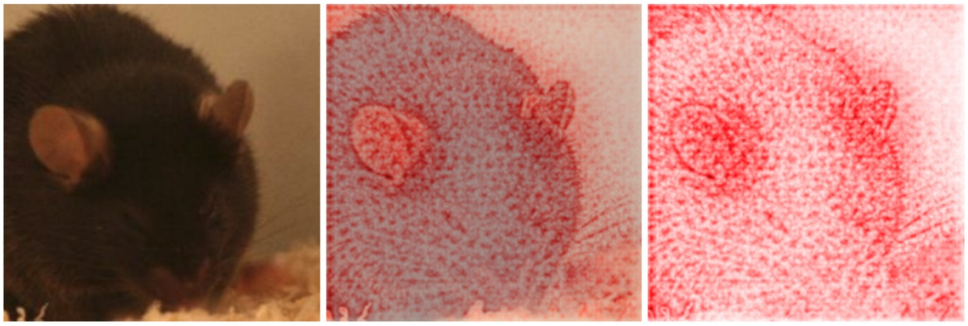


Fig 11. Contribution of piloerection and space between whiskers to the decision making. Visualization of the decision finding process using deep Taylor decomposition for an image generated 150 min after ketamine/xylazine anesthesia. Original image (left), original image combined with heat map (middle), heat map (right). Red color indicates that a pixel contributes to the decision. Piloerection and the space between the whiskers seem to play a role in the decision making progress. The mice was correctly classified as "post-anesthetic/surgical effect" with a confidence of 99,6%.

https://doi.org/10.1371/journal.pone.0228059.g011

pixels of the body if visible, such as the outline of the back or in some cases the cervical/thoracic area (Fig 9). The outline of the back is altered if a mouse shows a hunched posture or piloerection (Fig 11), which can accompany distress and pain [71]. The cervical/thoracic area is only entirely visible if the animal sits upright, which can be used as an additional clue for the state of the mouse, i.e., the head of a mouse lies flat on the ground in most images taken in the early post-anesthetic period following injection anesthesia with ketamine/xylazine combination. Moreover, in images from this period particular facial features are very distinctive, e.g., the ear position. Probably due to pharmacological effects of the injection anesthetics, the ears are dropped and the space between the ears widens. In some cases, decomposition analysis even suggests that the decision for the classification is mainly based on the ears (see bottom row in Fig 10).

However, in order to determine which of the various facial and body features contributed most to the decision making progress of the deep neural network, future work is needed. We hope that we will be able to localize facial elements and body parts in the images. This would enable us to quantitatively analyze the weighting of the used features and allow to compare features of the algorithm with the facial action units of the MGS. It would furthermore enable us to investigate whether the decision for models trained on different procedures (i.e., IN, KXN, C) is based on different features. These information may lead to a better understanding of the high confidence of the network in cases such as 30 min and 150 min post-anesthesia with the combination of ketamine/xylazine.

Conclusion

We developed a semi-automated pipeline to automatically recognize post-anesthetic or post-surgical effects from the face of a mouse using a deep learning neural network framework. For the first time, this approach was pursued involving images of black-furred laboratory mice moving freely in their cages. Depending on the treatment of the mice, an accuracy of up to 99% was reached for the binary classification ("post-anesthetic/surgical effect"/"no post-anesthetic/surgical effect"). A model trained on a particular treatment performs worse if it is tested to another treatment, suggesting that different treatments produce divergent facial expressions. Mainly eyes, ears, and whiskers, but furthermore in some cases additional facial (nose tip, whisker pad) as well as body (outline of the back, cervical/thoracic area) features seem to contribute to binary classification in some cases. The findings of our study promote the development of a prototype tool for monitoring well-being of laboratory mice in experimental settings and their home cages.

Outlook

The long term goal is to devise a smart-surveilled environment for laboratory mice. The proposed approach will be the foundation for a "smart mouse cage", i.e., an integrated system with around-the-clock video monitoring of facial expressions as well as other parameters in laboratory mice and an alert function indicating that the well-being of an animal is impaired. Another possible application will be a computer program or smart phone app, which supports the experimenter in assessing MGS scores in real time. The envisaged app could be used for online analysis of video stream with alert generation in case of a deviance appearance, i.e., changes in facial expressions accompanied by distress and pain.

We believe that our approach for automated surveillance of the well-being state would also be useful for other animal species. However, one of the crucial steps for pipeline automation is a proper face finder. The one currently used would need to be replaced in order to not require manual rejection of false positives (e.g. with Mathis et al. [72]). The use of video data instead of still images could also lead to a fully automated face detection as face hypotheses with low confidence scores could be rejected. Another interesting future direction of research would be an assessment of relative importance of features and its comparison to the MGS scoring scheme, which may reveal new facial features beyond the MGS. Besides negative well-being, facial expressions may also represent positive well-being, as demonstrated in rats [73]. While facial indicators of positive emotions in mice have not been reported yet, it might be achieved with the help of an automated pipeline.

Supporting information

S1 File. Additional information of number of visible action units and sex differences.
(PDF)

S1 Fig.
(TIF)

S2 Fig.
(TIF)

S3 Fig.
(EPS)

Acknowledgments

The authors would like to thank Doris Ciuraj, Sabine Jacobs, Melanie Humpenöder, and Giuliano Corte for assisting in MGS scoring. We are grateful to Oliver Leonhardt who developed our MGS scoring database. The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation: www.dfg.de) under Germany's Excellence Strategy—EXC 2002 "Science of Intelligence"—project number 390523135 (www.scienceofintelligence.de). Mouse Grimace Scale scores were obtained from a project of the Berlin-Brandenburg research platform BB3R (www.bb3r.de) that was funded by the German Federal Ministry of Education and Research (grant number: 031A262A) (www.bmbf.de/en/index.html). We acknowledge support by the German Research Foundation and the Open Access Publication Fund of the Freie Universität Berlin.

Author Contributions

Conceptualization: Lars Lewejohann, Olaf Hellwich, Christa Thöne-Reineke, Vitaly Belik.

Data curation: Niek Andresen, Manuel Wöllhaf, Katharina Hohlbaum.

Funding acquisition: Lars Lewejohann, Olaf Hellwich, Christa Thöne-Reineke.

Investigation: Niek Andresen, Manuel Wöllhaf, Katharina Hohlbaum.

Methodology: Niek Andresen, Manuel Wöllhaf, Vitaly Belik.

Project administration: Katharina Hohlbaum, Olaf Hellwich, Christa Thöne-Reineke, Vitaly Belik.

Resources: Katharina Hohlbaum, Olaf Hellwich.

Software: Niek Andresen, Manuel Wöllhaf, Vitaly Belik.

Supervision: Lars Lewejohann, Olaf Hellwich, Christa Thöne-Reineke, Vitaly Belik.

Visualization: Vitaly Belik.

Writing – original draft: Niek Andresen, Manuel Wöllhaf, Katharina Hohlbaum, Vitaly Belik.

Writing – review & editing: Lars Lewejohann, Olaf Hellwich, Christa Thöne-Reineke, Vitaly Belik.

References

1. Russell WMS, Burch RL. The principles of humane experimental technique. London: Methuen; 1959.
2. Poole T. Happy animals make good science. Lab Anim. 1997; 31:116–124. https://doi.org/10.1258/002367797780600198 PMID: 9175008
3. Hawkins P, Morton D, Burman O, Dennison N, Honess P, Jennings M, et al. A guide to defining and implementing protocols for the welfare assessment of laboratory animals: eleventh report of the BVAAWF/FRAME/RSPCA/UFAW Joint Working Group on Refinement. Lab Anim. 2011; 45(1):1–13. https://doi.org/10.1258/la.2010.010031 PMID: 21123303
4. Ekman P, Friesen WV. Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press. 1978.
5. Langford DJ, Bailey AL, Chanda ML, Clarke SE, Drummond TE, Echols S, et al. Coding of facial expressions of pain in the laboratory mouse. Nat Methods. 2010; 7:447–9. https://doi.org/10.1038/nmeth.1455 PMID: 20453868
6. Sotocinal SG, Sorge RE, Zaloum A, Tuttle AH, Martin LJ, Wieskopf JS, et al. The Rat Grimace Scale: a partially automated method for quantifying pain in the laboratory rat via facial expressions. Mol Pain. 2011; 7:55. https://doi.org/10.1186/1744-8069-7-55 PMID: 21801409
7. Keating SC, Thomas AA, Flecknell PA, Leach MC. Evaluation of EMLA cream for preventing pain during tattooing of rabbits: changes in physiological, behavioural and facial expression responses. PloS One. 2012; 7:e44437. https://doi.org/10.1371/journal.pone.0044437 PMID: 22970216
8. Holden E, Calvo G, Collins M, Bell A, Reid J, Scott E, et al. Evaluation of facial expression in acute pain in cats. J Small Anim Pract. 2014; 55:615–621. https://doi.org/10.1111/jsap.12283 PMID: 25354833
9. Dalla Costa E, Minero M, Lebelt D, Stucke D, Canali E, Leach MC. Development of the Horse Grimace Scale (HGS) as a pain assessment tool in horses undergoing routine castration. PLoS One. 2014; 9: e92281. https://doi.org/10.1371/journal.pone.0092281 PMID: 24647606
10. Gleerup KB, Andersen PH, Munksgaard L, Forkman B. Pain evaluation in dairy cattle. Appl Anim Behav Sci. 2015; 171:25–32. https://doi.org/10.1016/j.applanim.2015.08.023
11. Häger C, Biernot S, Buettner M, Glage S, Keubler L, Held N, et al. The Sheep Grimace Scale as an indicator of post-operative distress and pain in laboratory sheep. PloS One. 2017; 12:e0175839. https://doi.org/10.1371/journal.pone.0175839 PMID: 28422994
12. Lu Y, Mahmoud M, Robinson P. Estimating sheep pain level using facial action unit detection. In: 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Washington, DC: IEEE; 2017. pp. 394–399.
13. Di Giminiani P, Brierley VL, Scollo A, Gottardo F, Malcolm EM, Edwards SA, et al. The assessment of facial expressions in piglets undergoing tail docking and castration: toward the development of the piglet grimace scale. Front Vet Sci. 2016; 3:100. https://doi.org/10.3389/fvets.2016.00100 PMID: 27896270
14. Viscardi AV, Hunniford M, Lawlis P, Leach M, Turner PV. Development of a piglet grimace scale to evaluate piglet pain using facial expressions following castration and tail docking: a pilot study. Front Vet Sci. 2017; 4:51. https://doi.org/10.3389/fvets.2017.00051 PMID: 28459052
15. Miller AL, Leach MC. The mouse grimace scale: a clinically useful tool? PLoS One. 2015; 10:e0136000. https://doi.org/10.1371/journal.pone.0136000 PMID: 26406227
16. Miller AL, Kitson GL, Skalkoyannis B, Flecknell PA, Leach MC. Using the mouse grimace scale and behaviour to assess pain in CBA mice following vasectomy. Appl Anim Behav Sci. 2016; 181:160–165. https://doi.org/10.1016/j.applanim.2016.05.020 PMID: 27499567
17. Leach MC, Klaus K, Miller AL, Di Perrotolo MS, Sotocinal SG, Flecknell PA. The assessment of post-vasectomy pain in mice using behaviour and the Mouse Grimace Scale. PloS One. 2012; 7:e35656. https://doi.org/10.1371/journal.pone.0035656 PMID: 22558191
18. Stasiak K, Maul D, French E, Hellyer P, Vandewoude S. Species-specific assessment of pain in laboratory animals. Contemp Top Lab Anim Sci. 2003; 42:13–20. PMID: 12906396
19. Zhang EQ, Leung VS, Pang DS. Influence of rater training on inter-and intrarater reliability when using the Rat Grimace Scale. Journal of the American Association for Laboratory Animal Science. 2019; 58: 178–183. https://doi.org/10.30802/AALAS-JAALAS-18-000044 PMID: 30755291

20. Martinez B, Valstar MF. Advances, challenges, and opportunities in automatic facial expression recognition. In: Advances in face detection and facial image analysis. Springer; 2016. pp. 63–100.

21. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521:436. https://doi.org/10.1038/nature14539 PMID: 26017442

22. Baldassi C, Borgs C, Chayes JT, Ingrosso A, Lucibello C, Saglietti L, et al. Unreasonable effectiveness of learning neural networks: from accessible states and robust ensembles to basic algorithmic schemes. Proc Natl Acad Sci U S A. 2016; 113:E7655–E7662. https://doi.org/10.1073/pnas.1608103113 PMID: 27856745

23. Dubey A, Naik N, Parikh D, Raskar R, Hidalgo CA. Deep learning the city: quantifying urban perception at a global scale. In: European Conference on Computer Vision. Amsterdam: Springer; 2016. pp. 196–212.

24. Albert A, Kaur J, Gonzalez MC. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax, NS, Canada: ACM; 2017. pp. 1357–1366.

25. Mathis A, Warren RA. On the inference speed and video-compression robustness of DeepLabCut. BioRxiv. 2018; pp. 457242.

26. Bertram CA, Klopfleisch R. The pathologist 2.0: an update on digital pathology in veterinary medicine. Vet Pathol. 2017; 54:756–766. https://doi.org/10.1177/0300985817709888 PMID: 28578626

27. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA; 2012. pp. 1097–1105. Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

28. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus, Ohio; 2014. pp. 580–587.

29. Géron A. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. Sebastopol, CA: O'Reilly Media, Inc.; 2017.

30. Hohlbaum K, Bert B, Dietze S, Palme R, Fink H, Thöne-Reineke C. Severity classification of repeated isoflurane anesthesia in C57BL/6JRj mice—Assessing the degree of distress. PloS One. 2017; 12: e0179588. https://doi.org/10.1371/journal.pone.0179588 PMID: 28617851

31. Hohlbaum K, Bert B, Dietze S, Palme R, Fink H, Thöne-Reineke C. Impact of repeated anesthesia with ketamine and xylazine on the well-being of C57BL/6JRj mice. PloS One. 2018; 13:e0203559. https://doi.org/10.1371/journal.pone.0203559 PMID: 30231081

32. FELASA Working Group on Revision of Guidelines for Health Monitoring of Rodents and Rabbits, Mähler M, Berard M, Feinstein R, Gallagher A, Illgen-Wilcke B, Pritchett-Corning K, Raspa M FELASA recommendations for the health monitoring of mouse, rat, hamster, guinea pig and rabbit colonies in breeding and experimental units. Laboratory animals. 2014; 48: 178–192. https://doi.org/10.1177/0023677213516312

33. Hurst JL, West RS. Taming anxiety in laboratory mice. Nat Methods. 2010; 7:825. https://doi.org/10.1038/nmeth.1500 PMID: 20835246

34. Löscher W, Ungemach FR, Kroker R. [Grundlagen der Pharmakotherapie bei Haus-und Nutztieren]. Berlin Hamburg: Paul Parey; 1991.

35. Behringer R, Gertsenstein M, Vintersten Nagy K, Nagy A. Protocol 11: Castration of mice. In: Manipulating the mouse embryo: a laboratory manual, fourth edition. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2014. pp. 234–235.

36. Wright-Williams SL, Courade JP, Richardson CA, Roughan JV, Flecknell PA Effects of vasectomy surgery and meloxicam treatment on faecal corticosterone levels and behaviour in two strains of laboratory mouse. Pain. 2007; 130: 108–118. https://doi.org/10.1016/j.pain.2006.11.003 PMID: 17196337

37. Matsumiya LC, Sorge RE, Sotocinal SG, Tabaka JM, Wieskopf JS, Zaloum A, et al. Using the Mouse Grimace Scale to reevaluate the efficacy of postoperative analgesics in laboratory mice. J Am Assoc Lab Anim Sci. 2012; 51: 42–49. PMID: 22330867

38. Wright-Williams S, Flecknell PA, Roughan JV Comparative effects of vasectomy surgery and buprenorphine treatment on faecal corticosterone concentrations and behaviour assessed by manual and automated analysis methods in C57 and C3H mice. PloS One. 2013; 8: e75948. https://doi.org/10.1371/journal.pone.0075948 PMID: 24098748

39. Hohlbaum K, Bert B, Dietze S, Palme R, Fink H, Thöne-Reineke C. Systematic assessment of well-being in mice for procedures using general anesthesia. J Vis Exp. 2018; p. e57046.

40. Defensor EB, Corley MJ, Blanchard RJ, Blanchard DC. Facial expressions of mice in aggressive and fearful contexts. Physiol Behav. 2012; 107:680–685. https://doi.org/10.1016/j.physbeh.2012.03.024 PMID: 22484562

41. Descovich KA, Wathan J, Leach MC, Buchanan-Smith HM, Flecknell P, Framingham D, et al. Facial expression: An under-utilised tool for the assessment of welfare in mammals ALTEX. 2017; 34: 409–429. https://doi.org/10.14573/altex.1607161 PMID: 28214916

42. Tuttle AH, Molinaro MJ, Jethwa JF, Sotocinal SG, Prieto JC, Styner MA, et al. A deep neural network to assess spontaneous pain from mouse facial expressions. Mol Pain. 2018; 14:1744806918763658. https://doi.org/10.1177/1744806918763658 PMID: 29546805

43. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Kauai, Hawaii, USA; 2001. pp. 511–518.

44. Bradski G, Kaehler A. Learning OpenCV: Computer vision with the OpenCV library. Sebastopol, California: "O'Reilly Media, Inc."; 2008.

45. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: CVPR09. Miami; 2009. pp. 248–255.

46. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, et al. Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning; 2014. pp. 647–655.

47. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Advances in neural information processing systems; 2014. pp. 3320–3328.

48. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas; 2016. pp. 2818–2826.

49. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015; 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y

50. He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: European conference on computer vision. Amsterdam: Springer; 2016. pp. 630–645.

51. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous systems; 2015. Available from: https://www.tensorflow.org/.

52. Jirkof P, Fleischmann T, Cesarovic N, Rettich A, Vogel J, Arras M. Assessment of postsurgical distress and pain in laboratory mice by nest complexity scoring. Lab Anim. 2013; 47:153–161. https://doi.org/10.1177/0023677213475603 PMID: 23563122

53. Wellington D, Mikaelian I, Singer L. Comparison of ketamine–xylazine and ketamine–dexmedetomidine anesthesia and intraperitoneal tolerance in rats. J Am Assoc Lab Anim Sci. 2013; 52:481–487. PMID: 23849447

54. Flecknell P. 1—Basic principles of anaesthesia. In: Laboratory animal anaesthesia, fourth edition. Boston: Academic press; 2016. pp. 1–75.

55. Matta JA, Cornett PM, Miyares RL, Abe K, Sahibzada N, Ahern GP. General anesthetics activate a nociceptive ion channel to enhance pain and inflammation. Proc Natl Acad Sci U S A. 2008; 105:8784–8789. https://doi.org/10.1073/pnas.0711038105 PMID: 18574153

56. Kichko TI, Niedermirtl F, Leffler A, Reeh PW. Irritant volatile anesthetics induce neurogenic inflammation through TRPA1 and TRPV1 channels in the isolated mouse trachea. Anesth Analg. 2015; 120:467–471. https://doi.org/10.1213/ANE.0000000000000568 PMID: 25517196

57. TerRiet M, DeSouza G, Jacobs J, Young D, Lewis M, Herrington C, et al. Which is most pungent: isoflurane, sevoflurane or desflurane? Br J Anaesth. 2000; 85:305–307. https://doi.org/10.1093/bja/85.2.305 PMID: 10992843

58. McNulty G, Eyre L. Humidification in anaesthesia and critical care. Continuing Education in Anaesthesia Critical Care & Pain. 2014; 15:131–135.

59. Moody CM, Weary DM. Mouse aversion to isoflurane versus carbon dioxide gas. Appl Anim Behav Sci. 2014; 158:95–101. https://doi.org/10.1016/j.applanim.2014.04.011

60. Hayase T, Tachibana S, Yamakage M. Effect of sevoflurane anesthesia on the comprehensive mRNA expression profile of the mouse hippocampus. Med Gas Res. 2016; 6:70–76. https://doi.org/10.4103/2045-9912.184715 PMID: 27867470

61. Viswanath O, Kerner B, Jean YK, Soto R, Rosen G. Emergence delirium: a narrative review. Anesthesiol Clin Sci. 2015; 4:2. https://doi.org/10.7243/2049-9752-4-2

62. Holaday DA, Fiserova-Bergerova V, Latto IP, Zumbiel MA. Resistance of isoflurane to biotransformation in man. Anesthesiology. 1975; 43:325–332. https://doi.org/10.1097/00000542-197509000-00009 PMID: 1163832

63. Hijazi Y, Boulieu R. Contribution of CYP3A4, CYP2B6, and CYP2C9 isoforms toN-demethylation of ketamine in human liver microsomes. Drug Metab Dispos. 2002; 30:853–858. https://doi.org/10.1124/dmd.30.7.853 PMID: 12065445

64. Garcia-Villar R, Toutain P, Alvinerie M, Ruckebusch Y. The pharmacokinetics of xylazine hydrochloride: an interspecific study. J Vet Pharmacol Ther. 1981; 4:87–92. https://doi.org/10.1111/j.1365-2885.1981.tb00715.x PMID: 7349331

65. Choi SO, et al. The metabolism of xylazine in rats. Arch Pharm Res. 1991; 14:346–351. https://doi.org/10.1007/BF02876882

66. Dalla Costa E, Pascuzzo R, Leach MC, Dai F, Lebelt D, Vantini S, et al. Can grimace scales estimate the pain status in horses and mice? A statistical approach to identify a classifier. PloS One. 2018; 13: e0200339. https://doi.org/10.1371/journal.pone.0200339 PMID: 30067759

67. Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition. 2017; 65:211–222. https://doi.org/10.1016/j.patcog.2016.11.008

68. Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018; 73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011

69. Alber M, Lapuschkin S, Seegerer P, Hägele M, Schütt KT, Montavon G, et al. iNNvestigate neural networks! CoRR. 2018;abs/1808.04260.

70. Dörfl J. The musculature of the mystacial vibrissae of the white mouse. J Anat. 1982; 135:147–54. PMID: 7130049

71. Carstens E, Moberg GP. Recognizing pain and distress in laboratory animals. ILAR J. 2000; 41:62–71. https://doi.org/10.1093/ilar.41.2.62 PMID: 11304586

72. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nature Publishing Group; 2018.

73. Finlayson K, Lampe JF, Hintze S, Würbel H, Melotti L. Facial indicators of positive emotions in rats. PloS One. 2016; 11: e0166446. https://doi.org/10.1371/journal.pone.0166446 PMID: 27902721