Using deep learning models to decode emotional states in horses

Romane Phelipon, Lea Lansade, Misbah Razzaq

Using deep learning models to decode emotional states in horses

Romane Phelipon
Inrae
Lea Lansade
Inrae
Misbah Razzaq
misbah.razzaq@inrae.fr

Inrae

Research Article

Keywords: Artificial intelligence, emotion detection, classification, animal welfare, convolutional neural networks.

Posted Date: October 14th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5244800/v1

Additional Declarations: The authors declare no competing interests.

Using deep learning models to decode emotional states in horses

Romane Phelipon[1], **Lea Lansade**[1,+], **and Misbah Razzaq**[1,+,*]

[1]INRAE, CNRS, Université de Tours, PRC, 37380, Nouzilly, France
[*]misbah.razzaq@inrae.fr
[+]These authors share the last authorship.

ABSTRACT

In this work, we present various machine learning models to predict emotional states in horses. We manually label the images to learn the task in a supervised manner. We perform data exploration and use different cropping methods, mainly based on Yolo and Faster R-CNN, to create two new datasets: 1) the cropped body, and 2) the cropped head dataset. We build different models based on convolutional neural networks (CNNs) using (un)cropped datasets and compare their performances to accurately predict emotions. The cropped head dataset yields the best results despite lacking important region of interests like a tail, which experts use to annotate images. Furthermore, we update our models using various techniques, such as transfer learning and fine-tuning to further improve their performance. The best performance is achieved through a model based on stacking principals, which gave a boost to the overall performance with an accuracy of 87%, precision of 79%, and recall of 97%. Finally, we employ three interpretation methods to understand the internal workings of our models. Different interpretation methods seem to highlight different features of the same model. We found that only LIME appears to detect some of the features that are used by experts to annotate emotional states.

Introduction

Animal welfare is becoming an increasing social concern, especially regarding animals utilized by humans. Equestrian sports are no exception, and numerous horse-riding practices are currently criticized for being detrimental to the comfort of the horses (1). In order to identify the horse's state of well-being of ridden horses, certain indicators based on behaviours, posture and facial expressions must be taken into account (2; 3). For example, tail swishing behaviour indicates a state of discomfort in the horse when ridden and is often synchronised with the use of spurs (2). In addition, the head behind the vertical posture is recognised as being a sign of the horse's discomfort (2). Finally, certain facial expressions such as the opening of the eyes or mouth and the position of the ears provide information about the comfort level of the ridden horse (3). However, measuring these indicators is challenging and requires sustained observation by individuals with experience in the field. In fact, it is generally necessary to carefully watch the videos of the horses and manually code each behaviour, posture, or facial expression. This process is highly time-consuming as it is done manually and sometimes even frame by frame (e.g., in the case of facial expressions (4)). Additionally, it involves a subjective element, often necessitating that two individuals independently review the same video to verify the consistency of the measurements. Thus, an automatized deep learning analysis based on the detection of certain regions of interest, such as facial expressions and the horse's posture, would be a great help in taking account of animal welfare.

Artificial intelligence (AI)-based methods have gained popularity in recent years and have been successfully applied in many domains such as image recognition, robotics, speech recognition, life sciences, etc. Regarding deep learning models to predict emotions in horses, in one study (5), authors built classification models using convolutional neural networks (CNNs) based on facial features. These features were annotated using the horse grimace scale method. This study was based on the assessment of the facial expressions of seven horses undergoing castration, which were filmed two days before and four days after the procedure. They selected 3000 images by visual inspection out of 185672 extracted frames of videos. Finally, this dataset was divided into three subsets based on different features: 1) ears, 2) eyes, and 3) chewing muscles, mouths and nostrils. They obtained an accuracy of

75.8% for classifying pain into three categories: not present, moderately present, and obviously present. However, for the binary classification, i.e., presence or absence of the pain, they achieved an accuracy of 88.3%. In another study (6), the authors developed a detector to recognize horses in the image and a classifier to predict the emotion of the detected horse. To train and test the system, a dataset of 440 images was collected from private sources, with each image labeled with one of four emotional markers: alarmed, annoyed, curious, or relaxed. There were 110 images per emotion, and the dataset was divided into 400 training images and 40 test images. Their model achieved an accuracy of 65% on the testing set.

As compared to the aforementioned methods, in this work, we used a combination of labelling methods: HGS (7) and RHpE (8), instead of using a single method. This approach enables us to obtain information not only on facial emotions but also on those coming from the other body parts of horses. Previous results have shown the need to crop the images in the dataset to allow the model to focus on the crucial elements of the image while reducing the influence of the background on prediction accuracy. We used pretrained Yolo and faster-RCNN to identify and crop the horse's body and the horse's head. We used several preprocessing methods to improve quality of the data, such as data or resolution augmentation. Then, for the classification step, we built a model from scratch (called the baseline model). We also employed transfer learning and fine-tuning techniques using different backbone architectures, such as VGG16 and ResNet50. For hyperparameter optimization, we used a Bayesian search algorithm. Finally, we concatenated the best-performing models (VGG16 and Xception) and obtained better results than individual models, recall of 97% and accuracy of 87% on the testing set. Given the well-known black box nature of deep learning models, we employed a variety of interpretation techniques to highlight the salient features of our models. We generated explanations for different predictions of our model. By contrasting the explanations produced by various approaches, we further emphasize the significance of the development of robust interpretation techniques.

Methods

Dataset

Our dataset consists of 1036 images of horses divided into two classes: comfortable (546) and uncomfortable (490), coming from both public and private sources. We divide the dataset into three distinct subsets: the training set (70%), the validation set (15%), and the test set (15%). The training set is used to train the machine learning model, optimizing model parameters to minimize loss. The validation set helps in fine-tuning the model and is typically used for adjusting hyperparameters. The test set is a part of the dataset that the model has never encountered during training or validation, it serves as an impartial assessment of the model's performance.

Annotations

A doctoral student in ethology specialising in the identification of horse emotions and a technician in ethology, selected the images. The images came either from private sources or from the internet on copyright-free image sites. Images in the comfortable and uncomfortable category were selected according to features described in previous studies (2; 3) and depending on the state of these five key points. "Ears forward; erect and parallel with pinnae facing forward" (3) was considered as a comfortable feature as opposed to both ears backward, which is commonly considered to be a negative state of the ridden horse (2; 9). Open, round and tension-free eyes, was considered as a comfortable feature whereas "almond-shaped eyes with tension of musculus levator anguli oculi medialis" (3) (i.e. tension above the eyes) was considered as an uncomfortable feature as suggested in a study on the development of ridden horses ethogram focused on the facial expressions (3). The opening of the horses' mouths was also evaluated. If the mouth was closed, it was considered as comfortable. If the mouth was open, it was considered as uncomfortable, as studies have shown that horses with more constraints when being ridden or those with musculoskeletal pain open their mouths more often (9; 10). Head behind the vertical is known as a practice with negative effects on horses (1; 11). Thus, head behind the vertical was considered as an uncomfortable feature, if not it was considered as a comfortable feature. Tail swishing is a behaviour that is generally expressed by horses when they feel uncomfortable when being ridden or as a conflict behaviour (2) so the feature was therefore considered as part of the uncomfortable category. Basal tail with no swishing was considered as part of the comfortable category. If horses expressed in images at least 2 features belonging to the category uncomfortable, they were placed in this

category. Horses in the category comfortable expressed the 5 comfortable features. We summarize different criteria used in this study for being in comfortable and uncomfortable states in the table 1.

Comfortable	Uncomfortable
Forward ears	Backward ears
Open eyes without tension	Tension above the eyes Closed eyes or Half closed eyes Sclera exposed
Basal tail	Tail swishing
Closed mouth	Open mouth
Head not behind the vertical	Head behind the vertical

Table 1. Key points for annotating different emotional states in horses.

Image resolution

Our dataset contains both low-resolution and high-resolution images. Image resolution plays an important role in the accuracy of pattern recognition in neural networks. We verified the impact of higher resolutions on computational time and accuracy of the model. A smaller image resolution leads to a reduction in training time, while increasing the resolution allows the model to focus on smaller features (e.g., mouth opening). It has been shown that increasing the size of the input image can increase the accuracy of the predictions up to a certain point (12). Thus, a higher resolution does not always lead to better predictions. That is why, we decided to fix the resolution to 256 x 256 which is generally employed in CNN-based methods.

Data augmentation

One of the most common techniques used in machine learning and computer vision for increasing the size and diversity of a training dataset is data augmentation. This technique applies different transformation methods to the images to improve the capacity of the model to generalize, i.e., prevent overfitting, and improve its performance on unseen data (13). There are two main ways to perform data augmentation. One way is to increase the images (by performing certain transformations) within the dataset, resulting in an increase in the dataset itself. This method has a number of disadvantages, in particular because we need to determine the number of images that can be generated from a single image to avoid generating images that are too close semantically and also to separate our dataset into subsets beforehand to avoid data leaking by finding the same or similar images in several subsets, which distorts the results. An alternative is to use a second class of methods that apply a random combination of the transformations to the images during training process. In our study, we employ a second method using the tensorflow image augmentation function (14).

Cropped dataset

Previous results have shown the need to crop the images in the dataset to allow the model to focus on the crucial elements of the image while reducing the influence of the background on prediction accuracy. In this study, we cropped the dataset into two sets: 1) horse body dataset, and 2) horse head dataset. We use Yolo and Faster-RCNN to perform the cropping operation.

Cropped body

To create a horse body cropping dataset, a pipeline was developed using yolov8x (15), to identify and crop only the horse bodies recognized by the model. However, the dataset contains some images with multiple horses per image, so to avoid contamination of the future dataset, a selection is made under the assumption that one image corresponds to one horse. This selection consists of keeping only the image of the horse's body that has the highest resolution, in case multiple horses are recognized. After a manual check, no incorrect cropping was detected, and only two images were not supported by the pipeline (format issues).

Cropped head

In order to perform a cropping of the horse's head, we used a pretrained Faster-RCNN model (6) that can crop the horse's head from horse's body. A manual check was then performed, which revealed an error in six images (incorrect cropping or format errors).

CNN architecture

CNNs are a special type of feed forward neural networks inspired from human vision, and mainly used to perform image classification, object detection, and clustering similar images. They are based on three layer types (16): convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification. A convolutional layer basically first performs element-wise multiplications using different filters or kernels (matrices of numbers) applied to the input data, and then sum the results to generate feature maps. A pooling layer is used to perform the sampling of the feature maps in order to conserve only important information, thereby getting rid of noise and redundancy. Pooling enables CNNs to be invariant to small translations; spatial translation has little effect on the output of the pooling operation (17). Fully connected layer(s) are finally used after pooling to perform classification or regression tasks (18).

We design the architecture by taking into account various factors such as data representation, network topology, activation functions, and hyper-parameter tuning (see figure 1). For the feature extraction architecture, we create four blocks consisting of convolutional layers, batch normalization layer, and max pooling layers. For the classification architecture, we create 3 blocks composed of fully connected (dense) along with a dropout. A flatten layer is used to establish links between feature extraction and classification by reducing the values in a single dimensional vector. To reduce overfitting, we use dropout and L1 or L2 regularization.

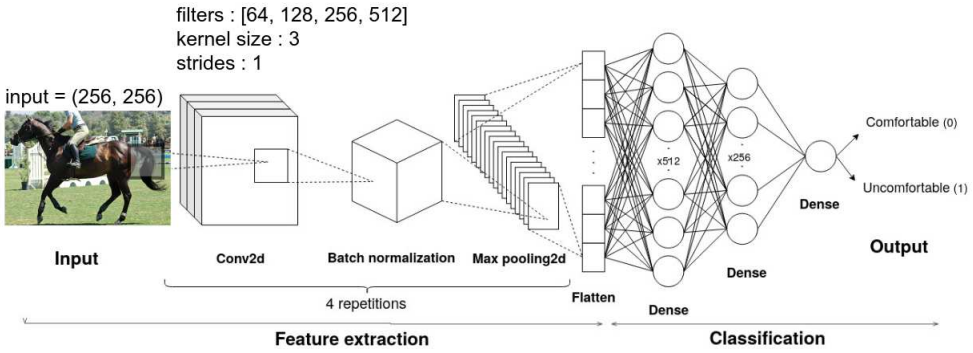


Figure 1. Baseline architecture representation

Hyperparameter tuning

Hyperparameter tuning consists of finding the optimal set of hyperparameters for a learning model. This method works by maximizing the performance of the model according to a desired performance metric. Automated hyperparameter tuning methods use an algorithm to search for optimal values. In this work, we used Bayesian optimization with the Hyperopt Python library (19; 20).

Transfer learning and fine tuning

Transfer learning is a machine learning technique that enables the knowledge acquired from training a model on one classification task, such as classifying one type of class, to be applied to another classification task involving a different class. Using a transfer learning approach, we decided to test and compare widely used pre-trained models: ResNet50, ResNet152, VGG16, VGG19, InceptionV3, Xception, and EfficientNetV2L. In our approach, we removed the classification layers of the different models and added a new classification layer, including a flattening layer or global average pooling layer (depending on the hyperparameter tuning), a fully connected layer, and a dropout. Then we kept only the classification part by freezing the weights of the feature extraction layers of the model, and this allowed us to use the pre-trained weights to learn faster.

By modifying the pre-trained weights and enhancing prediction performance, fine-tuning can unfreeze specific feature extraction layers, enabling the model to adapt and specialize in particular horse feature recognition. This is because training the classification portion of the model alone does not help to create feature maps directly influenced by the features of interest. Since we know that the final convolution layers hold the high-level semantic information, these layers will be unfrozen and retrained at the end of each training session (a learning session that ends when the model is no longer able to learn or is overfitting). On the other hand, because the first few low-level layers have the best weights for identifying low-level characteristics, they will stay frozen.

Stacked model

A stacking model, also known as a stacked ensemble, is a machine learning technique that combines multiple base models to improve overall predictive performance (21). In our case, we combine the models with best performance on our dataset (VGG16 and Xception) to create a stacked model. The first step is to train these models separately, then combine their output to feed it as input to the meta-classifier. Once the stacked model is trained, it can be used to make predictions on new unseen data.

Performance evaluation

Different metrics can be used evaluate a model's ability to generalize and effectiveness in making predictions. In this study, we use accuracy, precision, and recall to gauge model performance. Accuracy describes the overall performance of the model by calculating the ratio of correct predictions (true positives and true negatives) to the total number of predictions. (see equation 1).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP (true positive) is number of correctly predicted as positive by the model (emotional state is comfortable and model's prediction is comfortable), TN (true negative) represent number of correctly predicted as negative by the model (emotional state is uncomfortable and model's prediction is uncomfortable), FP (false positives) is number of incorrectly predicted as positive by the model (emotional state is uncomfortable and model's prediction is comfortable), and FN (false negatives) denotes number of incorrectly predicted as negative by the model (emotional state is comfortable and model's prediction is uncomfortable).

Precision is calculated as the ratio between the number of positive samples correctly classified and the total number of samples classified as positive (in our binary case, it is the correct percentage of comfortable prediction among all comfortable predictions) (equation 2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is calculated the proportion of actual positives samples that was identified correctly (in our case, the percentage of correct prediction of comfortable among all comfortable emotional states)(equation 3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Interpretation methods

Machine learning models have made remarkable progress in recent years, enabling the development of sophisticated models capable of tackling complex tasks with unprecedented accuracy. The downside is that as these models become more complex and opaque, their interpretability becomes a key concern. In this study, we use state-of-the-art methods such as LIME (Local Interpretable Model-Agnostic Explanations) (22), SHAP (SHapley Additive exPlanations) (23), and Grad-CAM (Gradient-weighted Class Activation Mapping) (24) to provide valuable information about the inner workings of the proposed models, explaining how they arrive at their predictions by identifying important features or regions of interest.

Results

Preprocessing

In our dataset, we have images with high and low resolutions. We may lose information while applying rescaling operation on low-resolution images. To avoid this problem, we tested different resolution augmentation tools on our dataset. Manual checking is performed to keep the one that presented the best resolution increase without distorting the horse's feature. In figure 2, we show an example of resolution augmentation on one image.

Figure 2. A horse images from our dataset going through the resolution augmentation.

Furthermore, we transform images using various methods such as rotation, scaling, flipping, zooming, channel shifting, shearing, and filing mode. Since the data augmentation parameters are directly related to the dataset, the optimal configuration of the parameters was tested and evaluated for each new dataset (cropped body and cropped head). In figure 3, we show an example of data augmentation on our dataset.

Figure 3. Examples of some data augmentation.

Data augmentation improves model's performance

We compare the impact of data and resolution augmentation on the baseline model. Results are presented in the table 2. We divide our dataset into training, validation, and testing sets. We evaluate the model using accuracy, precision, and recall on the testing set. To obtain robust and reliable performance measures, we use k-fold cross-validation with k = 10. Our results show a slight increase in performance with the use of data augmentation,

increasing the accuracy by 3%, whereas resolution augmentation only increase the accuracy by 1%. Taken together, these preprocessing methods give a final increase of around 4% for the baseline on the horse dataset. Given the small increase in model performance with the use of resolution augmentation, we apply it only to the uncropped images and the cropped body datasets.

Preprocessing / Metrics	Accuracy (%)	Precision (%)	Recall (%)
Baseline	60.84 ± 4.54	60.34 ± 8.74	60.34 ± 6.54
Baseline + data augmentation	63.12 ± 6.54	**64.34** ± 7.54	59.34 ± 5.56
Baseline + resolution augmentation	61.40 ± 7.25	59.12 ± 8.49	59.59 ± 6.10
Baseline + data and resolution augmentation	**63.92** ± 5.54	62.03 ± 6.42	**62.64** ± 6.98

Table 2. Baseline model results on resolution and data augmentation with 10-fold validation

Models based on cropped head dataset show superior performance

Here, we present the results of a CNN baseline models built using three different datasets: 1) cropped head dataset, cropped body dataset, and uncropped dataset. The goal is to investigate which dataset provides the best performance for the baseline model. From table 3, we can see that the baseline CNN model achieved the highest accuracy and precision on the cropped head dataset, followed by the cropped body and the uncropped dataset. The superior performance of the model on the cropped head dataset can be attributed to its ability to focus on key features related to facial recognition or head pose estimation. Similarly, the cropped body dataset provides additional features, such as the tail, that are not present in the cropped head dataset, contributing to the improved performance compared to the horses dataset, which contains a lot of unhelpful information that can hinder learning.

Dataset / Metrics	Accuracy (%)	Precision (%)	Recall (%)
Uncropped	63.59 ± 4.57	62.44 ± 8.54	58.70 ± 5.95
Cropped body	66.65 ± 8.26	63.85 ± 8.77	**62.19** ± 7.40
Cropped head	**70.48** ± 4.26	**74.35** ± 4.67	61.84 ±4.75

Table 3. Baseline model results on horse, cropped body and cropped head dataset with 10-fold validation

Transfer learning improves the model's performance

We further improve the accuracy of the model by applying transfer learning and fine-tuning. We employed several pre-trained models ResNet50, ResNet152, VGG16, VGG19, InceptionV3, Xception and EfficientNetV2L as backbone architecture. The best models are obtained using VGG16 and Xception as pre-trained models.

Image resolution has an important impact on model performance: We tested two models to compare the impact of image resolution, i.e., Xception and VGG16, on 7 different resolution sizes in table 4. These models are built using horse dataset without any cropping operation. The results show that below 128x128 and above 320x320 the 2 models no longer increase their accuracy. Peak of the performance is reached at 128x128 for VFF16 and at 256x256 for Xception. Hence, we set the resolution for the models at 256x256. Note that, we are increasing the resolution mainly on images below this value.

Resolution / Models	(48*48)	(64*64)	(128*128)	(224*224)	(256*256)	(320*320)	(360*360)
VGG16	65.71	66.99	**70.83**	66.98	69.55	70.51	68.65
Xception	57.67	61.99	66.96	71.19	**76.97**	76.92	73.43

Table 4. Impact of image resolution in terms of accuracy on testing dataset using VGG16 and Xception.

Fine-tuning improves the model's performance: We compare the performance of the fine-tuning models on the cropped body dataset in table 5. The evaluation parameters used for the comparison are accuracy, precision, and recall (on the test dataset). A total of eight models were evaluated, and the results show that the Xception model achieved the best overall performance. The Xception model achieved an accuracy of 78%, which indicates that the Xception model succeeded in classifying the horse's body with a high level of accuracy. On the other hand, the Xception model achieved a precision rate of 72%, which is lower than the VGG16 result. Furthermore, the model achieved a recall rate of 70%.

Model / Metrics	Acc (%)	Precision (%)	Recall (%)
Baseline	66.65	63.85	62.19
VGG16	72.44	**76.81**	66.25
VGG19	71.79	76.47	65.00
ResNet50	61.54	63.89	57.50
ResNet152	62.82	64.86	60.00
InceptionV3	60.90	64.62	52.50
Xception	**78.010**	72.13	**69.84**
EfficientNetV2L	61.544	66.137	51.25

Table 5. Fine-tuning models on the body crop dataset (testing set).

Stacked models outperform individual models: Here, we create a stacked model by combining the best models obtained through the previous analysis as discussed above, i.e., VGG16 and Xception. The accuracy curves for the stacked model are shown in figure 4. We show confusion matrix in figure 5a and ROC curve in figure 5b. We obtain AUC score of 0.99 on the training set, AUC of 0.85 on the validation set, and AUC of 0.93 on the testing set.

Furthermore, we compare the performance of different model, i.e., baseline, fine-tuning on models, and staked model. The results are shown in table 6. These models are built using cropped head and body dataset. The stacked model achieved excellent results, with an accuracy of 87%. In addition, it achieved a precision of 79% and a recall rate of 97%, suggesting that the model is confident in predicting comfortable horses. These results highlight the superior performance of the stacked model, which outperforms individual models.

Different interpretation methods highlight different parts of the image

In this section, we discuss different region of interest in relation to what the model finds relevant utilizing the LIME, SHAP, and GRAD-CAM approaches, as well as the main features used by experts for deciphering emotional states. In tables 7 and 8, we compare explanations based on accurate predictions of the model on both comfortable and uncomfortable emotional states. Features that correspond to comfortable and uncomfortable emotional states are highlighted in the green and red colors. We employ a model constructed with a cropped head dataset to derive these explanations. We display explanations when the original and predicted labels are comfortable in table 7. We

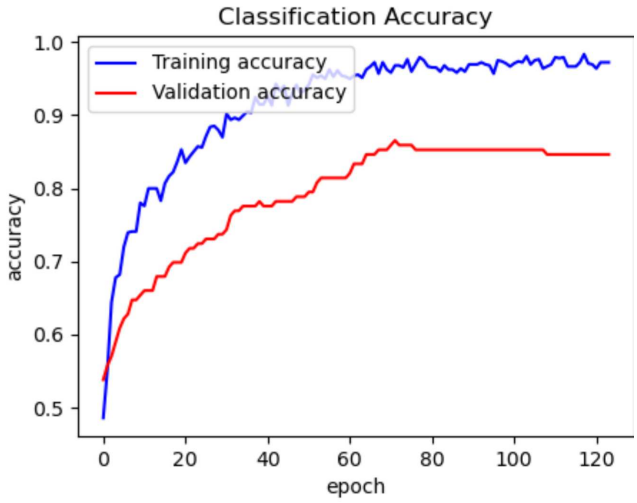
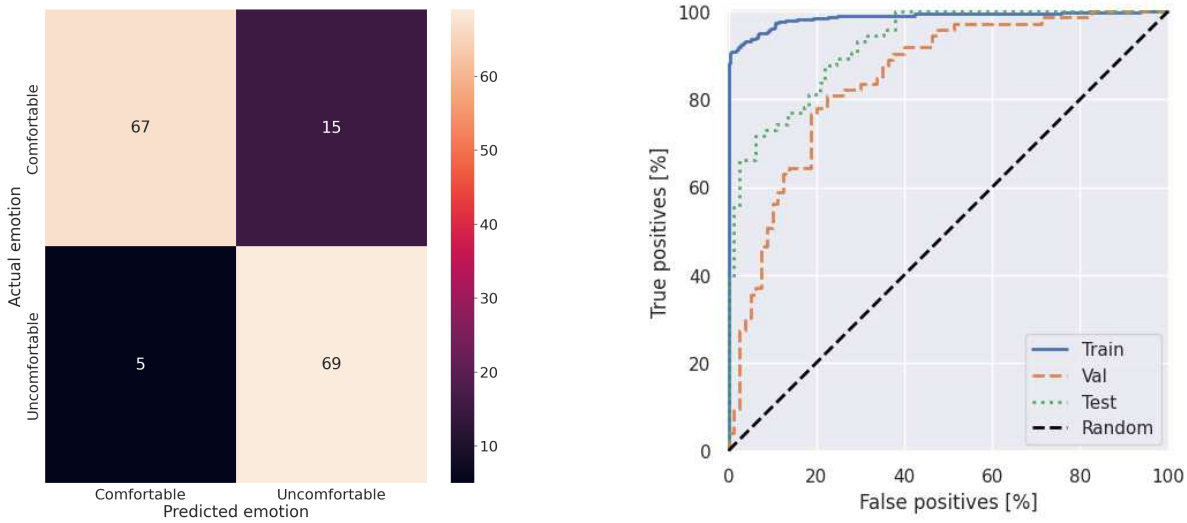


Figure 4. Training and validation accuracy on the cropped head dataset.



(a) Confusion matrix.

(b) ROC curve.

Figure 5. Performance on cropped head dataset

can observe that the mouth and nostril are the main focus of the LIME explanation in both images. In the first image, the SHAP highlights the chamfer and ears, whereas in the second image, it emphasizes the neck. Grad-CAM emphasizes the neck in the first picture, while in the second, it concentrates on the chamfer and background. The features that the expert has underlined in the images are highlighted in the column True features. We note that while each method reveals distinct regions of interest, these regions are consistent with some of the expected features. It's interesting to see that for the explanation for comfortable prediction with true comfortable label, the only negative component is the horse's chamfer behind the vertical. However, it should be noted that there are different degrees of hyperflexion, and the further back the chamfer is, the greater the discomfort, creating breathing problems (11). Here, the prediction still put him in a state of comfort, and you can see on the image that the hyperflexion is very slight. The degree of hyperflexion may have been taken into account when generating the prediction, which would resemble the reasoning used to assess a horse's state of comfort by a human expert.

We present explanations in table 8 where the horse is predicted to be in an uncomfortable state by our model

Model / Metrics	Acc (%)	Precision (%)	Recall (%)
Baseline (Body)	66.65	63.85	62.19
Baseline (Head)	70.59	74.14	58.90
VGG16 (Body)	72.44	76.81	66.25
VGG16 (Head)	83.33	79.27	87.84
Xception (Body)	78.01	72.13	69.84
Xception (Head)	84.23	**84.66**	83.91
Stacking(Xc+VG) (Head)	**86.54**	79.12	**97.30**

Table 6. Comparison of different model on testing dataset.

and it is actually in an uncomfortable state. In the first row, we can observe that LIME emphasizes the mouth and nostrils, while SHAP emphasizes the background and Grad-CAM emphasizes the neck. The second image shows that Grad-CAM concentrates on the neck, SHAP emphasizes the mouth, edge, and ears, while LIME concentrates on the mouth and ears. While there are some interesting aspects in the interpretation methods, most of the important points are in the background.

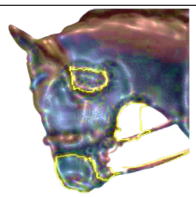
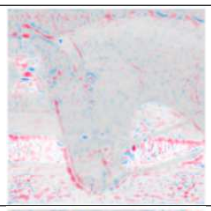
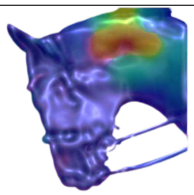
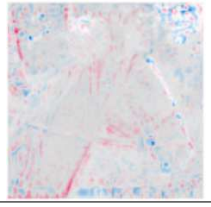
LIME	SHAP	Grad-CAM	Explanation	True features
			LIME: Mouth and nostril SHAP: Chamfer and ears Grad-CAM: Neck	Forward ears Open eyes Closed mouth Head behind vertical
			LIME: Mouth and nostril SHAP: Neck Grad-CAM: Background and chamfer	Forward ears Open eyes Closed mouth Head not behind vertical

Table 7. Explanation for comfortable prediction with true comfortable label.

Our findings show how the explanation approaches successfully detected some keypoints, such as the mouth, nostrils, and chamfer, even when certain background elements are also present. It's interesting to note that some images (first row in both tables 7 and 8) include contradictory keypoints, like a closed mouth signifying a comfortable condition or a head behind vertical indicating an uncomfortable state. Although these opposing keypoints are included in both tables, it is unclear how this element affects the final prediction. We also see that the key points of each method differ, even if they are applied to the same images and model. Furthermore, when compared to experts' annotation methods, LIME identified more accurate regions of interest than Grad-cam and SHAP.

Discussion

In this work, we present an original strategy based on various updates to the baseline architecture to predict emotional states in horses, including data augmentation, different training strategies, annotation strategies, and cropping datasets, with the aim of improving the overall performance and generalizability of the model. First,

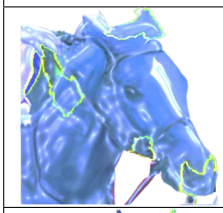
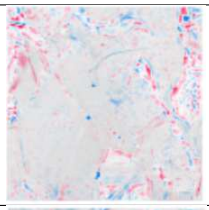
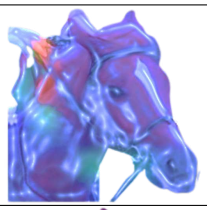
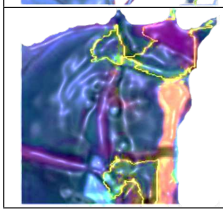
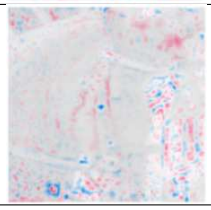
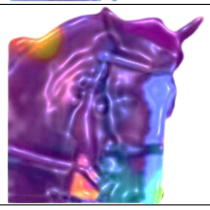
LIME	SHAP	Grad-CAM	Explanation	True features
			LIME : Nostril and mouth SHAP : Background Grad-CAM : Neck	Backward ears Tension above the eyes Half closed eyes Head not behind vertical Closed mouth
			LIME : Mouth and ears SHAP : Mouth, ears, and edges Grad-CAM : Neck	Backward ears Tension above the eyes Half closed eyes Head behind the vertical

Table 8. Explanation for uncomfortable prediction with true uncomfortable label.

our annotation strategy to construct the datasets combines the knowledge from two annotation methods. The incorporation of several pre-processing and model-building techniques, such as resolution augmentation, horse body cropping, and stacking model construction, gave the model an overall boost in performance. Finally, we implemented various interpretation methods to explain our models by highlighting the important features and comparing the explanations proposed by these methods.

With an accuracy of 87% in classifying an image of a ridden horse in the right category (Comfortable/Comfortable), our model performs well. It's worth further discussing two additional points: firstly, the discovery of a new component by the model for classification, and secondly, the model's lack of consideration for the position of the horse's muzzle, which is a key factor in correctly categorizing ridden horses. First, it seems that the model also takes into account a key point that was not initially included. We can see that the area around the rider's hands and reins is an area of interest for the model in order to predict that the horse is in a state of discomfort. This finding seems perfectly logical, given that the rider's hand actions can have an impact on the horse. In particular, one study has shown that the rider has an effect on the tension in the reins, which can lead to a change in heart rate and an increase in the horse's cortisol levels (25). In addition, it has been shown that short reins generate more conflict behaviour from the horse as tail swishing and more backward ears (9), features that were taken into account when annotating the different images in the uncomfortable category. The model not only classified the images according to the expected key points, but also found a new component to take into account. Second, one of the features that our model underestimated to classify ridden horses in a comfortable state was the chamfer behind the vertical. The chamfer behind the vertical is one of the head neck positions widely known to have detrimental physiological effects (26; 27; 28; 29) such as vascularization issues (30). However, it should be noted that there are different degrees of head and neck position when the chamfer is behind the vertical and these parameters were not taken into account when selecting the images. A study showed that the more the head neck position was in hyperflexion, the higher the cortisol rate was (1). Also, it is known that the head and the neck position can affect independently the pharyngeal diameter (11), creating different degrees of discomfort in the ridden horse. The model classified the images as comfortable despite a chamfer slightly behind the vertical. The degree of hyperflexion may have been taken into account when generating the prediction. As it's complicated for the human eye to perceive these nuances of chamfer behind the vertical, our model could therefore be improved and used to detect in ambiguous head neck positions.

Nonetheless, there are a few limitations to this study: 1) The dataset is obtained from private as well as public sources, which hinders our ability to share it publicly; 2) the size of the dataset is small, around 1000 images to build a machine learning model; and 3) the models are non-interpretable. In the future, we plan to work on resolving some of these issues. First, we want to create a larger and more diverse dataset by automatically exploiting the video dataset to generate horse images. Even after this, we will still need to manually annotate these images in order to enrich the training dataset. Furthermore, we plan to work more on interpreting the models, as currently we only use a

few images to generate explanations, which hinders our ability to draw general conclusions regarding the regions of interest in the images. In our current pipeline, we have identified that different methods generate diverse explanations. In the future, we want to identify a global set of explanations per category for different explanation methods. These explanations will be discussed with the experts to compare the key regions of interest on the images identified by the model, which may lead to the identification of new regions of interest or the validation of explanation methods. Finally, we plan to develop a user-friendly interface that, after due validation, can then be used by experts in their work to avoid the manual interpretation of emotions.

Conclusion

In this work, we have demonstrated how deep learning models can be used to predict emotional states in horses. We show how model built on cropped dataset has higher performance thanks to being able to focus on crucial features. Different updates in architecture or training strategies improve the overall performance of the proposed model. Further, we demonstrated how different interpretation methods can be applied to identify the important features of the model. Lastly, we demonstrated the requirement for a reliable interpretation method by highlighting differences in explanations across various interpretation approaches.

Acknowledgment

This research was funded with the support of the Institut National de la Recherche pour I'Agriculture, I'Alimmentation et I'Environnement (INRAE) and Phase innovative project. We thank Céline Parias for her contribution in selecting the images used in this study. Additionally, we extend our appreciation to Matis Alias Bagarre for his efforts in coding the model. We are also grateful to the Institut Français du Cheval et de l'Équitation (IFCE) for their financial support.

References

1. Christensen, J. W., Beekmans, M., Van Dalum, M. & VanDierendonck, M. Effects of hyperflexion on acute stress responses in ridden dressage horses. *Physiol. & behavior* **128**, 39–45 (2014).

2. Dyson, S. & Pollard, D. Application of the ridden horse pain ethogram to horses competing at the hickstead-rotterdam grand prix challenge and the british dressage grand prix national championship 2020 and comparison with world cup grand prix competitions. *Animals* **11**, 1820 (2021).

3. Mullard, J., Berger, J. M., Ellis, A. D. & Dyson, S. Development of an ethogram to describe facial expressions in ridden horses (fereq). *J. veterinary behavior* **18**, 7–12 (2017).

4. Lansade, L. *et al.* Facial expression and oxytocin as possible markers of positive emotions in horses. *Sci. reports* **8**, 14680 (2018).

5. Lencioni, G. C., de Sousa, R. V., de Souza Sardinha, E. J., Corrêa, R. R. & Zanella, A. J. Pain assessment in horses using automatic facial expression recognition through deep learning-based modeling. *PLOS ONE* **16**, 1–12, DOI: 10.1371/journal.pone.0258672 (2021).

6. Corujo, L. A., Gloor, P. A., Kieson, E. & Schloesser, T. Emotion recognition in horses with convolutional neural networks (2022). 2105.11953.

7. Dalla Costa, E. *et al.* Development of the horse grimace scale (hgs) as a pain assessment tool in horses undergoing routine castration. *PloS one* **9**, e92281, DOI: 10.1371/journal.pone.0092281 (2014).

8. Dyson, S. The ridden horse pain ethogram. *Equine Vet. Educ.* **34**, 372–380 (2022).

9. Ludewig, A., Gauly, M. & von Borstel, U. K. Effect of shortened reins on rein tension, stress and discomfort behavior in dressage horses. *J. Vet. Behav. Clin. Appl. Res.* **2**, e15–e16 (2013).

10. Dyson, S., Berger, J., Ellis, A. D. & Mullard, J. Development of an ethogram for a pain scoring system in ridden horses and its application to determine the presence of musculoskeletal pain. *J. Vet. Behav.* **23**, 47–57 (2018).

11. Cehak, A., Rohn, K., BARTON, A.-K., Stadler, P. & Ohnesorge, B. Effect of head and neck position on pharyngeal diameter in horses. *Vet. Radiol. & Ultrasound* **51**, 491–497 (2010).

12. Sabottke, C. & Spieler, B. The effect of image resolution on deep learning in radiography. *Radiol. Artif. Intell.* **2**, e190015, DOI: 10.1148/ryai.2019190015 (2020).

13. Yang, S. *et al.* Image data augmentation for deep learning: A survey (2022). 2204.08610.

14. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning (2016). 1605.08695.

15. Reis, D., Kupec, J., Hong, J. & Daoudi, A. Real-time flying object detection with yolov8 (2023). 2305.09972.

16. Yamashita, R., Nishio, M., Do, R. K. G. & Togashi, K. Convolutional neural networks: an overview and application in radiology. *Insights into imaging* **9**, 611–629 (2018).

17. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org.

18. Parhi, K. K. & Unnikrishnan, N. K. Brain-inspired computing: Models and architectures. *IEEE Open J. Circuits Syst.* **1**, 185–204 (2020).

19. Snoek, J., Larochelle, H. & Adams, R. P. Practical bayesian optimization of machine learning algorithms (2012). 1206.2944.

20. James Bergstra, Dan Yamins & David D. Cox. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. In Stéfan van der Walt, Jarrod Millman & Katy Huff (eds.) *Proceedings of the 12th Python in Science Conference*, 13 – 19, DOI: 10.25080/Majora-8b375195-003 (2013).

21. Proscura, P. & Zaytsev, A. Effective training-time stacking for ensembling of deep neural networks (2022). 2206.13491.

22. Ribeiro, M. T., Singh, S. & Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier (2016). 1602.04938.

23. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions (2017). 1705.07874.

24. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **128**, 336–359, DOI: 10.1007/s11263-019-01228-7 (2019).

25. Christensen, J. W. *et al.* Rider effects on horses' conflict behaviour, rein tension, physiological measures and rideability scores. *Appl. Animal Behav. Sci.* **234**, 105184 (2021).

26. Kienapfel, K., Link, Y. & König v. Borstel, U. Prevalence of different head-neck positions in horses shown at dressage competitions and their relation to conflict behaviour and performance marks. *PloS one* **9**, e103140 (2014).

27. Smiet, E. *et al.* Effect of different head and neck positions on behaviour, heart rate variability and cortisol levels in lunged royal dutch sport horses. *The Vet. J.* **202**, 26–32 (2014).

28. Wijnberg, I., Sleutjens, J., Van Der Kolk, J. & Back, W. Effect of head and neck position on outcome of quantitative neuromuscular diagnostic techniques in warmblood riding horses directly following moderate exercise. *Equine veterinary journal* **42**, 261–267 (2010).

29. Zebisch, A., May, A., Reese, S. & Gehlen, H. Effect of different head–neck positions on physical and psychological stress parameters in the ridden horse. *J. animal physiology animal nutrition* **98**, 901–907 (2014).

30. Sleutjens, J. *et al.* Effect of head and neck position on intrathoracic pressure and arterial blood gas values in dutch warmblood riding horses during moderate exercise. *Am. journal veterinary research* **73**, 522–528 (2012).