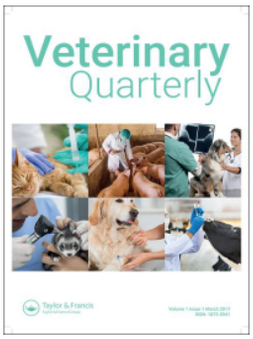

Veterinary Quarterly

ISSN: 0165-2176 (Print) 1875-5941 (Online) Journal homepage: www.tandfonline.com/journals/tveq20

Monitoring horse behaviour with deep learning models

Claudia Giannone , Chiara Maccario , Emanuela Dalla Costa , Elie Atallah & Marco Bovo

To cite this article: Claudia Giannone , Chiara Maccario , Emanuela Dalla Costa , Elie Atallah & Marco Bovo (2026) Monitoring horse behaviour with deep learning models, Veterinary Quarterly, 46:1, 2665442, DOI: 10.1080/01652176.2026.2665442

To link to this article: https://doi.org/10.1080/01652176.2026.2665442

Published online: 28 Apr 2026.

VETERINARY QUARTERLY
2026, VOL. 46, NO. 1, 2665442
https://doi.org/10.1080/01652176.2026.2665442

Taylor & Francis
by informa

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Monitoring horse behaviour with deep learning models

Claudia Giannone[a], Chiara Maccario[b], Emanuela Dalla Costa[b], Elie Atallah[b] and Marco Bovo[a]

[a]Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy; [b]Department of Veterinary Medicine and Animal Science, University of Milan, Lodi, Italy

ABSTRACT

Detailed analysis of stabled horse behaviour can reveal accurate information about its well-being. Advances in deep learning now allow these behaviours to be tracked without being invasive through the use of video data. This study evaluated a convolutional neural network for recognising standing, lying, and drinking behaviours in a horse housed in a wooden stall and recorded continuously over 29 consecutive days. Model predictions were compared with manually annotated ground truth data. Standing was detected with high precision (97.5%) and high recall (89.2%). Lying behaviour was classified with high precision (92.8%) but lower recall (63.1%). Activity patterns showed that standing dominated daily time budgets (>85%), lying accounted for 5-10%, and drinking occurred most often between 04:00 pm and 10:00 pm. These results demonstrate that deep learning can classify common equine behaviours from video, supporting its use in automated welfare monitoring. Future evaluations will explore the recognition of less frequent behaviours.

ARTICLE HISTORY
Received 18 September 2025
Accepted 20 April 2026

KEYWORDS
Equine welfare; computer vision; time budget; pose estimation; object detection; deep learning; equine behaviour; activity recognition

1. Introduction

Time budget is defined as an objective and quantifiable measure that represents the amount of time an individual dedicates to different activities (Auer et al. 2021; Lamanna et al. 2025). Assessing time budgets provides valuable insights into the welfare status of animals as 'animal-based indicators' (Auer et al. 2021), since it considers the frequency and duration of these behaviours and the ability of the animal to fully express the species-typical repertoire, known as the ethogram (Auer et al. 2021). Analysing behavioural routines over time can help compare different environments and investigate how horses of different ages and health statuses are influenced by management conditions. However, to achieve reliable results, horse activity must be monitored continuously, accounting for circadian variations (Bertolucci et al. 2008; Auer et al. 2021; Aragona et al. 2025). Therefore, measuring time budgets requires detailed surveillance over several days by trained observers, a process that is time-consuming, costly, and subject to observer bias (Auer et al. 2021).

To address these limitations, research is advancing in the use of biotelemetry, employing sensors and video analysis with algorithms to record behaviour, extend observation periods, and quantify behavioural patterns with greater accuracy and objectivity than human observers (Auer et al. 2021; Yigit et al. 2022). Healthy horses living in natural environments allocate a large part of their time to behaviours essential for survival and health, collectively referred to as 'maintenance behaviours'. This category includes ingestive behaviour (feeding and drinking), elimination, locomotion, rest, and shelter- or comfort-seeking behaviours (McDonnell 2003). Monitoring these behaviours provides a valuable starting point for assessing natural routines and, consequently, for inferring positive welfare (Gobbo et al. 2025). Among these, sleep, particularly paradoxical sleep, is especially important as it is essential for both physiological (e.g. high caloric ingestion without weight gain, reduction in anabolic hormones) and cognitive functions (e.g. regulation of neuronal functioning) during memory storage and consolidation and cerebral metabolism within the prefrontal cortex, which is responsible for judgement and decision making (Gobbo et al. 2025).

CONTACT Claudia Giannone claudia.giannone2@unibo.it Department of Agricultural and Food Sciences, University of Bologna, Bologna, Italy

Paradoxical sleep, also known as the Rapid Eye Movement (REM) phase, is fundamental to prevent sleep deprivation, a major stressor that can alter both behaviour and emotional state. Moreover, the REM phase can be achieved only in a recumbent position, which horses, as prey animals, adopt only when they feel sufficiently comfortable and safe. Additionally, observing other key behaviours such as drinking and feeding can provide valuable insights into welfare, as sudden interruptions or the inability to perform normal foraging and resting sequences can be associated with increased levels of discomfort (Torcivia and McDonnell 2021).

Recognising discomfort in horses is essential for maintaining welfare, supporting daily husbandry and veterinary care, ensuring the safety of both animals and caretakers (Dalla Costa et al. 2014; Torcivia and McDonnell 2021). However, directly observing these behaviours, whether in person or via video recordings, can be time-consuming, particularly because horses spend only a small portion of the day lying down or drinking. Furthermore, the manifestations of pain and discomfort tend to be subtle in horses due to their prey nature, which encourages them to conceal vulnerability from predators, including humans (Torcivia and McDonnell 2020; 2021).

To address these challenges, computer vision technology offers automated methods to interpret and analyse visual data in animal environments (Lawin et al. 2023; Yang et al. 2025; Giannone et al. 2025). By applying techniques from image processing and machine learning, computer vision can extract detailed behavioural information, providing more comprehensive data than direct observation. This approach facilitates the detection of variations in time budgets, as well as repetitive behaviours and postures that are difficult to identify during short or manual observations (Torcivia and McDonnell 2021; Zuerl et al. 2024). Computer vision, sensors, and automated analysis applied to locomotion, stereotypies, and animal identification have been increasingly utilised to assess activity and behavioural patterns relevant to health and welfare. These technologies have already shown usefulness and potential for continuous and objective monitoring, supporting the early detection of welfare impairment and health problems (Yigit et al. 2022; Zuerl et al. 2024; Aragona et al. 2025; Yang et al. 2025). Therefore, the aim of the study was to investigate the application of a deep learning-based computer vision system to identify the behaviours of individual stabled horses.

2. Materials and methods

2.1. Horse ethology

Table 1 presents some data available in literature related to time budgets, in percentages over the 24 hours, comparing feral (i.e. free-range) and domesticated horses (Boyd et al. 1988; Auer et al. 2021). The high variability in time budget percentages between different studies can be attributed to several factors related to biological and environmental conditions (Auer et al. 2021). Age and gender play a major role in the distribution of time budgets. Foals until weaning spend significantly more time in a recumbent position, while after weaning, the time spent foraging increases, whilst stallions tend to spend more time standing alert and moving rapidly but less time foraging than mares (Boyd et al. 1988). Behavioural analysis revealed a complex diurnal and ultradian rhythmicity and seasonal variations of activity patterns (Auer et al. 2021) with an increase in the hours spent drinking and standing rather than grazing as temperature rose during the daylight hours (Boyd et al. 1988). Furthermore, the distribution of time budgets is strongly influenced by management conditions, such as, for instance, feeding regimes and availability, stocking density, and paddock and box sizes (Auer et al. 2021). It is worth noting that, even if very informative and useful from a veterinary point of view, the creation of the daily budget time is highly time-consuming if made by hand by human personnel, and furthermore, some behaviours occur for a limited time percentage during the 24 hours, making the human procedure not viable for long-period monitoring. These pave the way for the development of alternative solutions based on non-wearable sensors and human-independent procedures.

2.2. Description of the experimental setup

The study was conducted at an equine centre located in the Province of Lodi, close to Milan, Italy. The horse object of the study was stabled in a precast wooden box measuring 3×3 square metres and structured as

Table 1. Information on horse time budget available in literature. Values in the third column ('Average daily time budget') represent the overall min-max interval reported across studies considering all horse populations. The last two columns ('Time budget over 24 h observation') report, when available, the mean ± standard error of the mean (SEM) and the minimum and maximum values calculated separately for studies on feral horses and domesticated horses.

Behaviour	Description	Average daily time budget	Time budget over 24 h observation (feral)	Time budget over 24 h observation (domesticated)
Movement	Horse taking at least three steps. McDonnell (2003)	Ranging from 0.015% to 19.1% Auer et al.(2021)	Mean: 7.4% ± 1.0% Boyd et al. (1988) Ranging from 4.3% to 13.4% Auer et al. (2021)	Ranging from 0.015% to 19.3% Auer et al. (2021) Ranging from 2.5 to 19.3% (horses not confined in a stable) Auer et al. (2021)
Lying	Horse lies flat on the ground. Sternal: with sternum in contact with the ground, legs folded beneath the body, no or limited ear movement, eyes open or closed, and muzzle in contact or not with the floor. Lateral: either lateral thoracic area parallel to and in contact with the ground, head immobile and in contact with the ground, or legs extended. The head, chest, and abdomen in contact with the floor; and eyes closed. McDonnell (2003)	Ranging from 2.7% to 27.3% Auer et al. (2021) (Subject needs at least 30 minutes of lying down per day, corresponding to 2.1% of the time over 24 hours, to achieve 3.5-4.5 minutes of REM) Kelemen et al. (2021)	Mean: 1.2% ± 0.5% (laterally) Mean: 4.1% ± 3.0% (sternally) Ranging from 2.7% to 15.5% Auer et al. (2021)	Ranging from 3% to 7.3% Auer et al. (2021)
Standing	Horse maintains a stable quadrupedal posture. Standing rest: immobile display, no or limited ear movement, relaxed tail, eyes closed or half shut, head close to level with the withers or lower. McDonnell (2003)	Ranging from 8.1% to 66% (resting) Auer et al. (2021)	Mean: 20.6% ± 5.4% Boyd et al. (1988) Mean: 15.7% ± 3.2% (resting) Ranging from 8.1% to 29.3% (resting) Auer et al. (2021)	Ranging from 15.6 to 68% (resting) Auer et al. (2021)
Drinking	Horse approach the water surface with the mouth, and the muzzle intersects with the water surface. Ingest water, typically by using lips at or slightly below the surface of water, and drawing water with sucking action through slightly parted lips and teeth and swallowing. McDonnell (2003)		Mean: 0.5% ± 0.1% Boyd et al. (1988)	
Grazing/eating	Muzzle is lowered to ground/ within bucket; lips grasp hay/ bedding; masticating, prehending or swallowing food. McDonnell (2003)	Ranging from 10% to 66.6% Auer et al. (2021)	Mean: 46.4% ± 5.9% Boyd et al. (1988) Ranging from 13% to 67% Auer et al. (2021)	Ranging from 10% to 64% Auer et al. (2021)

illustrated in Figure 1. It was enclosed on three sides with a double-leaf door 150 cm large on the fourth side, which gave the horse the possibility to face outside. The box had a wood shavings litter box. Droppings were removed once a day, and clean shavings were added to replace those removed. A metal pressure drinker was installed in the far-right corner, opposite the door, and a plastic corner feeder was positioned in the left corner, next to the door, for the administration of concentrate feed. The camera

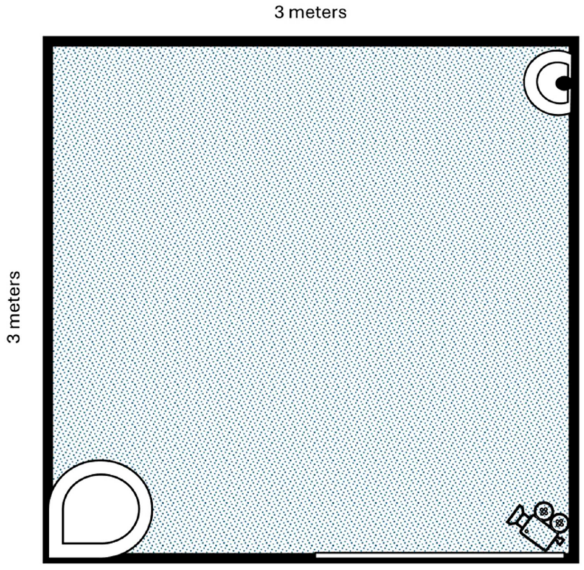

Figure 1. Layout of the box and experimental setup.

was fixed above the door on the right in order to frame the box completely. The horse was kept in the box during both day and night hours, except for work time, i.e. about two hours per week, and for the hours approximately between 8:00 am and 4:00 pm that were normally spent by the horses in the paddock under favourable weather conditions. On Mondays, the facility is closed to the public, and the horses are left in their box throughout the day. The horse was given concentrated feed and hay twice a day, as was customary. Hay was administered at 8:00 am in quantities of 5 kg in the paddock, or in the stall on rainy days, and at 5:00 pm in quantities of 6.5 kg in the stall every day. It was always given on the ground and, when given indoors, in front of the stall door. Concentrated feed was provided at 8:00 am in quantities of 1.5 kg and at 5:00 pm in quantities of 1.5 kg for a total of 3 kg/day inside the plastic feeder or in a bucket when fed in the box.

2.3. Recording system and data collection

Video data were collected over a continuous period of nearly one month, from 28/05/2024 to 25/06/2024, corresponding to 29 calendar days, including 28 complete 24-hour recording periods and 8 additional hours. A PoE-powered (Power over Ethernet) HIK DS-2CD2T87G2H-LI4 camera (HIKVISION, Hangzhou, China), which provided high-resolution imaging (3840 × 2160 pixels, 8 MP). The camera was mounted in a fixed corner position (see Figure 1), at a height that allowed for a full view of the horse without causing any disturbance to the animal. The camera angle allowed a complete top-down overview of both the horse and its surrounding environment of the box. The camera was connected to a dedicated workstation for continuous 24-hour video recording. Video acquisition was performed using the FFmpeg open-source software (Tomar 2006), and the files were automatically uploaded to a shared OneDrive folder to allow remote access by collaborators. Continuous footage recorded at 2 fps during the first four complete days (May 28th–31st) was used to extract 6,186 random frames, which were used to build the dataset for training, validation, and testing of the neural network to recognise different horse behaviours. The remaining 24 complete days were used for horse monitoring. During the first four days, the horse's behaviour was observed and manually annotated to identify relevant patterns. A subsequent data cleaning step was performed to discard frames where the horse was not clearly visible. The remaining frames were uploaded to Roboflow (Dwyer et al.), an online annotation platform, where they were manually labelled. For each image, an operator drew a bounding box around the horse and classified its posture, as illustrated in Figure 2. Roboflow was also used to generate YOLO-compatible annotation files for training and evaluation purposes.

Figure 2. Frame example of behaviour annotation. This frame shows the horse annotated with the label "lying", using Roboflow interface.

2.4. Data labelling

For the model development, a total of 6186 images were manually labelled to construct two distinct datasets, each designed to train a different deep learning model: one for behaviour classification (object detection) and the other for pose estimation. The first dataset focused on identifying the horse's posture, distinguishing between lying and standing behaviours. To this end, 2980 frames showing the horse in a lying position and 3206 frames with the horse standing were manually annotated. Behaviours were only considered valid and eligible for labelling if they were sustained for at least 2 seconds, allowing a temporal consistency. Otherwise, the behaviour was classified the same as the one in the previous frame.

A second dataset was built for pose estimation purposes and involved the annotation of eight anatomical keypoints: muzzle, head, withers, tail, right front point, left front point, right back hind and left back hind. These keypoints were selected for their relevance in describing the horse's body orientation and movement. Throughout these landmarks, the main goal of our study was to identify and estimate how much time and how frequently the horse spent drinking. To increase annotation accuracy, each keypoint was placed at the outermost visible edge of the corresponding anatomical region (Martvel et al. 2024). All keypoints were connected to form a structured representation of the horse's posture, as illustrated in Figure 3. Once labelled, each subset of data was randomly divided into training (70%), validation (20%), and test (10%) sets using a Python script.

2.5. Model training

To train the two models, we adopted a transfer learning approach, initialising the networks with pre-trained weights and subsequently fine-tuning them on our custom datasets. This strategy allowed us to get general features learnt from large-scale datasets while adapting the models to the specific context of equine behaviour and posture recognition.

2.5.1. Behaviour classification purpose

For the purpose of behaviour classification (standing or lying), a separate YOLOv8 model was trained using the object detection framework (Diwan et al. 2023). The initial dataset consisted of 6186 labelled frames, with 2980 images of lying behaviour and 3206 of standing. The model was trained on this dataset using YOLOv8 with pre-trained COCO weights, then fine-tuned over 70 epochs. Training was conducted on images resized to 640×640 with standard YOLO augmentations applied, which included cropping, rotation, saturation and definition of exposure (Giannone et al. 2025). Model training was performed using the previously defined training, validation, and test subsets. Table 2 summarises the behaviour instances of distributing training, validation and testing sets. To further validate the model performance in the operative

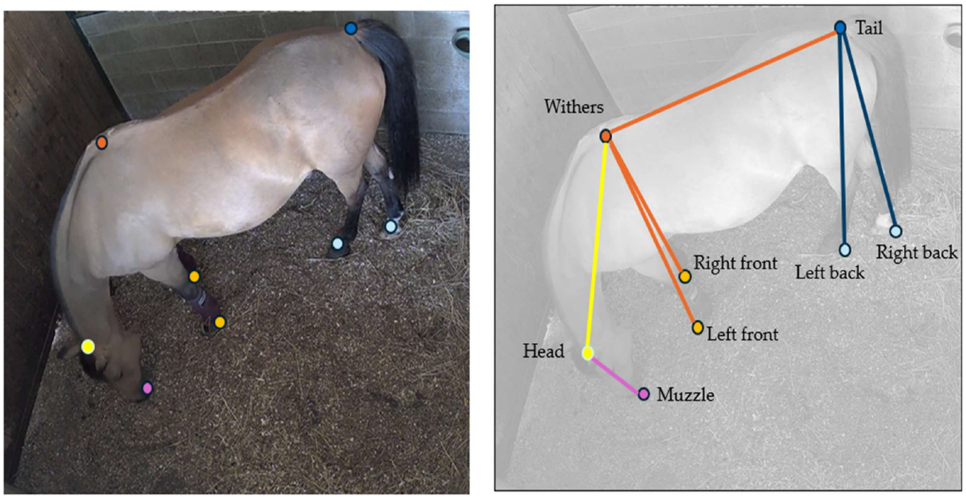

Figure 3. Overview of the keypoints used for pose estimation.

Table 2. Behaviour instance distribution of training, validation and testing sets.

Behaviour	Training set	Validation set	Test set	Total
Standing	2086	596	298	2980
Lying	2244	641	321	3206

environment, we conducted a parallel annotation and a comparison process: a human operator manually labelled video frames from the first four days and the trained YOLOv8 model was used to predict the horse behaviour on the same subset.

2.5.2. Keypoint detection purpose

The pose estimation model was trained using YOLOv8's keypoint detection module. Eight keypoints were labelled per image: muzzle, head, withers, tail, right front point, left front point, right back hind and left back hind. These points were manually annotated by a trained operator also in this case for the four days focusing on consistency and anatomical precision. Annotations were placed at the outermost visible parts of each body region. The model was fine-tuned on the annotated dataset using a batch size of 16, a learning rate of 0.01, and training for 70 epochs. Standard image augmentation techniques, including flipping, rotation, and brightness adjustments, were applied to increase generalisation capability. Table 3 summarises the parameters used to fine-tune the two models.

2.6. Model computing configuration

The network models used in the experiment were implemented and run in Python. The experiments were conducted using the hardware and software configuration summarised in Table 4. The data pre-processing, training and testing phases were performed using Python 3.11.5 with the following libraries: *Ultralytics YOLOv8* (version 8.4.7), *PyTorch* (version 2.9.1), *OpenCV* (version 4.12.0), *pandas* (version 2.3.1), *SciPy* (version 1.17.0), *NumPy* (version 2.4.1).

2.7. Behavioural metrics and analysis

To translate raw model outputs into behavioural indicators, we defined a set of metrics aimed at quantifying the daily and hourly activity patterns of the monitored horse. For each frame processed by the system, a behaviour label was assigned based on object detection or pose estimation output. During the initial four-day period, annotated data were used to train, validate, and test the neural network, as well as to define the threshold for activity detection based on agreement with human observations. Model

Table 3. Training parameters used to fine-tune YOLOv8 models.

Parameter	Value
Framework	YOLOv8
Pre-trained weights	YOLO8n (trained on COCO)
Fine-tuning dataset	Custom horse dataset
Annotation format	YOLO (.txt)
Image Size	640 × 640
Batch Size	16
Epochs	70
Patience	10
Learning Rate	0.01
Optimiser	AdamW
Preprocessing	Auto-Orient
Confidence Threshold	0.25
IoU Threshold	0.5
Momentum	0.937
Weight Decay	0.0005
Augmentations	Crop: 0% Min Zoom, 20% Max Zoom Rotation: −15° to + 15° Saturation: −25% to + 25% Exposure: −10% to + 10%

Table 4. Description of the hardware/software used for the analysis.

Hardware/Software	Value
CPU	Intel Core i7-9700K @ 3.60 GHz × 8
Memory	32 GB
GPU	NVIDIA GeForce RTX 2080 Ti (11 GB)
Hard Disk	9.1 TB (primary disk)
Operating System	Ubuntu 22.04
Driver	NVIDIA Driver 535.183.01
CUDA Version	12.2
Python Version	3.11.5

performance metrics were calculated exclusively on the predefined internal test subset (10%) from this period. Accuracy (A), precision (*P*), recall (R), and F1-score were calculated according to Equation (1).

$$\begin{aligned} A &= \frac{TP + TN}{TP + TN + FN + FP} \\ P &= \frac{TP}{TP + FP} \\ R &= \frac{TP}{TP + FN} \\ F1 &= 2 \cdot \frac{P \cdot R}{P + R} \end{aligned} \tag{1}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively (Giannone et al. 2025).

Subsequently, the trained model was applied to the remaining twenty-four days of recordings. In the case of lying behaviour, the system recorded the total time spent lying per hour. These hourly values were aggregated daily and converted into percentages relative to the total observation time. It is important to note that the system occasionally labelled frames as "unknown' either due to poor detection confidence or because the horse was physically absent from the monitored area. Behavioural frequency was also computed by identifying discrete episodes of lying behaviour. A new episode was defined as a continuous lying event lasting at least 2 seconds, followed by a transition to a standing position. Each such instance was counted as a single event, allowing the construction of both hourly and daily frequency profiles. To better interpret rest-related activity, we calculated the ratio between lying and standing time per day, which served as a simplified index of resting behaviour. This ratio was used to identify shifts in the horse's rest allocation and to detect possible deviations from typical patterns. For general activity classification, a motion-based approach was adopted. In terms of frames, comparisons between model predictions and

manual annotations were used to generate a confusion matrix and calculate classification performance metrics. These evaluation were conducted on the internal test subset derived from the initial four day annotated dataset. These annotated days were also used to define and refine the threshold used for activity detection, based on agreement with human interpretation. Table 5 shows the definition of the behaviours analysed.

3. Results and discussion

3.1. Evaluation of the model performance

In order to evaluate the performance of the behavioural classification model also in an operative environment, a comparison was conducted between the model predictions and manually annotated ground truth labels. The dataset used for the comparison consisted of 293680 labelled frames, corresponding to the whole dataset of approximately 87 hours of video footage collected in the first four days of experimental tests. Each frame was annotated by a human expert and categorised into one of the three behavioural classes: standing, lying or unknown, when the model does not detect the horse. It is important to note that the camera continuously recorded only the interior of the box. When the horse was temporarily outside, the video frames showed an empty box. Manual annotations were performed on the same video frames used for model evaluation. Therefore, the 'unknown' label included both cases of physical absence of the horse from the box and instances of low detection confidence. The evaluation involved calculating a confusion matrix to visualise the distribution of correct and incorrect predictions across all classes, along with standard classification metrics (precision, recall, F1-score, and overall accuracy), as defined in the methods section. Confusion matrix is shown in Figure 4, while the corresponding row and column totals are reported in Table 6. The overall accuracy of the behaviour classification model was equal to 88%.

The model achieved high performance in classifying the standing behaviour, correctly identifying over 240000 frames. Approximately 10% of standing frames were classified as 'unknown.' Although this proportion may appear notable, most misclassifications occurred toward the conservative 'unknown'

Table 5. Definition of the behaviours analysed.

Behaviour	Description	Instance
Standing	Horse maintain a stable quadrupedal posture	
Lying	Horse lies flat on the ground	
Drinking	Horse approach the water surface with his mouth, and his muzzle intersect with the water surface	

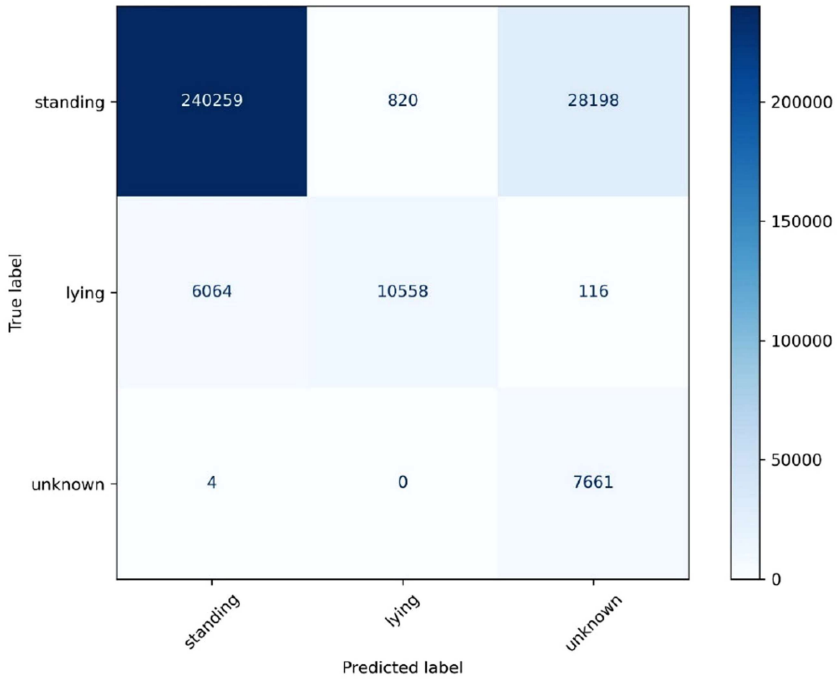

Figure 4. Confusion matrix of the automatic classification of horse behaviours. The values represent the number of frames for each combination between the true label (Y axis) and the predicted label (X axis).

Table 6. Behaviour classification performance including row and column totals. Percentages in parentheses represent class prevalence relative to the total number of evaluated frames ($n = 293680$).

True label	Predicted label			
	Standing	Lying	Unknown	Row total (*n*; %)
Standing	240259	820	28198	269277; 91.7%
Lying	6064	10558	116	16738; 5.7%
Unknown	4	0	7661	7665; 2.6%
Column total (*n*)	246327	11378	35975	293680; 100%

category rather than between standing and lying behaviours. From a behavioural monitoring perspective, this pattern reduces the risk of systematic misestimation of biologically distinct states such as lying. The high precision (97.5%) and recall (89.2%) for standing indicate strong classification performance, with recall corresponding to sensitivity for this behavioural state. For the lying class, the model showed high precision (92.8%), suggesting that when it predicted lying, it was usually correct. However, the lower recall (63.1%) indicates that many actual lying frames were misclassified, most commonly as standing. This suggests an opportunity for improvement in the model's ability to detect lying postures consistently.

The unknown class exhibited a very high recall (99.9%), meaning the model almost never failed to identify an actual unknown state. Probably the horse was out of the frame and out of the box. However, the very low precision (21.3%) suggests that the model tended to overpredict the unknown class, labelling many frames as such when the ground truth indicated a known behaviour. This likely reflects a conservative prediction strategy or difficulty in distinguishing ambiguous or transitional behaviours, which have not been separated into a specific category, but, during manual annotation, the period in which the horse was in the process of lying down was labelled as 'standing' until the horse had fully assumed the lying posture (i.e. until the belly was in contact with the ground), at which point the behaviour was labelled as 'lying'. On the contrary, when the horse began to push with its front legs to get up, the behaviour was labelled as 'standing', interrupting the annotation as 'lying'. It should be noted that the dataset exhibited marked class imbalance, with standing behaviour representing the majority of frames. While this imbalance influences absolute confusion matrix counts, the confusion between standing and lying remained limited. The overestimation of the 'unknown' category may

lead to a modest underestimation of standing duration. However, the estimation of biologically distinct states such as lying remained almost unaffected, as misclassification mostly happened between standing and unknown rather than between standing and lying. The general interpretation of resting behaviour patterns is thus not significantly changed, even though the amount of time spent standing may be slightly underestimated. Table 7 summarises the evaluation metrics.

A random selection of different frames extracted by the dataset used for the test of the model in operative configuration is displayed in Figure 5 to demonstrate the prediction results. Some environmental factors may have influenced classification performance. Detection confidence may have been impacted by fixed camera location and light variations, which may have contributed to the occurrence of 'unknown' classification. Further tests and development will involve the collection of additional data related to other horses to improve the robustness of the model. Moreover, further improvements and analyses will involve the possibility to monitor the animals in different boxes and with different litter types in order to explore the possible applications of the networks also in other environments and contexts.

3.2. Long term behaviour monitoring

A first application of the classification network developed involved the possibility to adopt the model for a long-term monitoring of the main behaviours of the horse. In fact, the computer vision system was maintained for a 29-day period. During this time, the horse was housed in the box but was periodically taken outside, while the camera remained active. The model developed and tested on the first four days of observation was then applied for monitoring the horse, classifying the three main behaviours: lying, standing, and unknown, where the

Table 7. Evaluation metrics for each class.

Class	Precision	Recall	F1-score	Frames (*n*; %)
Standing	0.975	0.892	0.932	269277; 91.7%
Lying	0.928	0.631	0.751	16738; 5.7%
Unknown Total	0.213	0.999	0.351	7665; 2.6% 293680; 100%
Overall accuracy				0.88

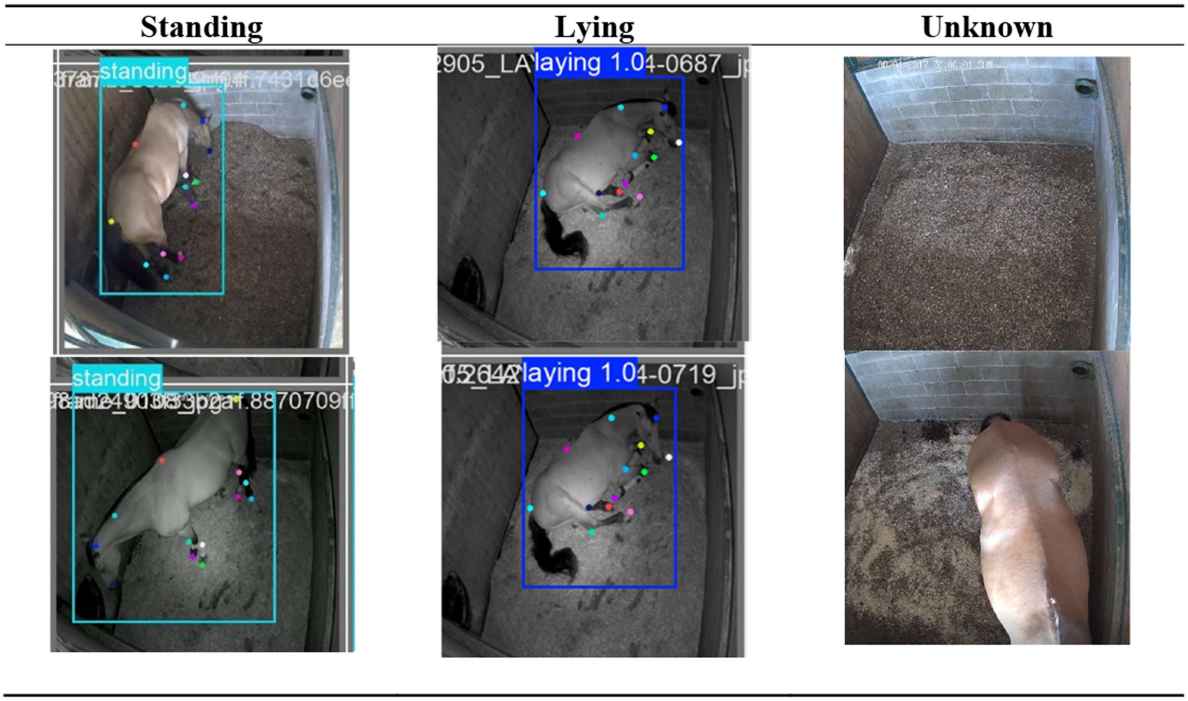

Figure 5. Representative prediction results on the test set for the three behavioural classes (standing, lying, and unknown). Two illustrative frames are shown per category.

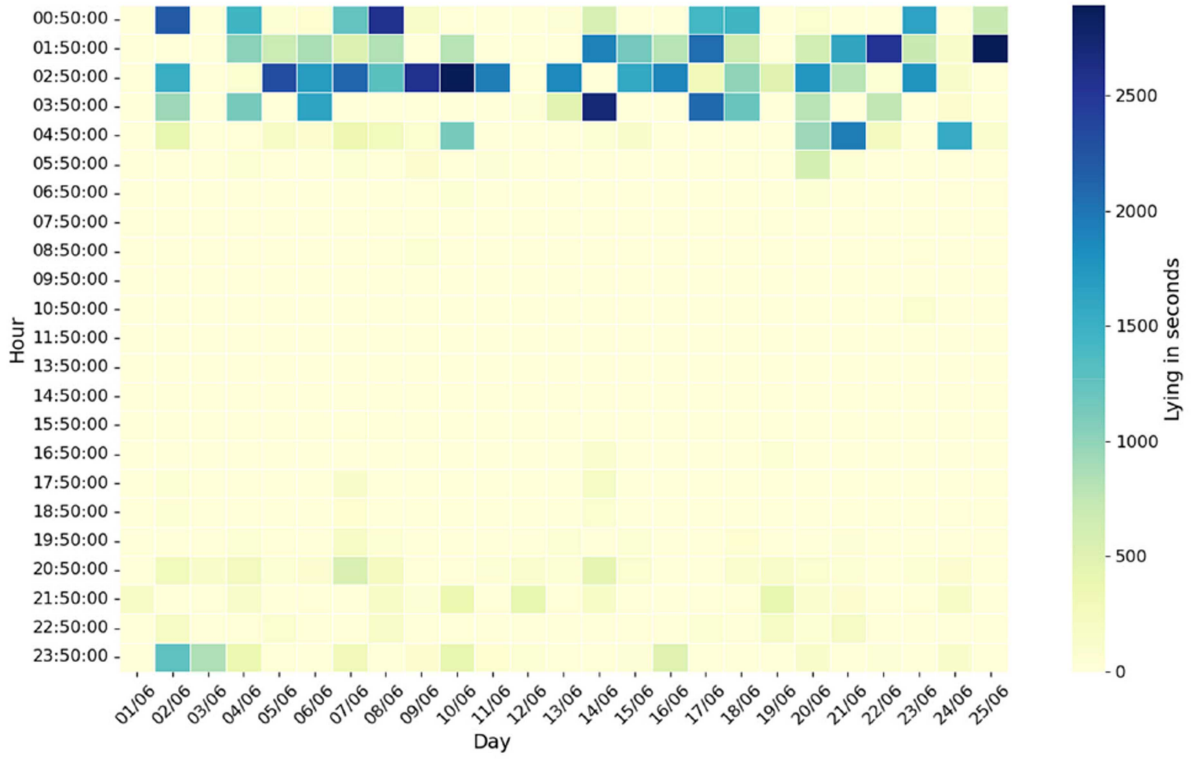

Figure 6. Heatmap of the distribution of lying behaviour along the monitored period. The different colour intensity indicates the different time spent lying.

behaviour could not be confidently determined, often due to visual limitations. To visualise the temporal patterns of lying behaviour, a heatmap was generated (Figure 6) showing the total lying time per hour for each day. The data clearly indicate a marked circadian rhythm, with the horse predominantly lying down during the night. More specifically, the highest concentration of lying occurred between 12:00 am and 4:00 am, with the peak typically observed between 01:00 am and 03:00 am. In some hours, lying time exceeded 2500 seconds (42 minutes), suggesting long, uninterrupted rest periods. Conversely, during daytime hours, lying behaviour was rare or entirely absent, which aligns with known equine behavioural patterns.

Further insight was provided through two additional visualisations focusing on the distribution of behaviours across time. In Figure 7, daily percentages of lying, standing, and unknown behaviours are presented. As expected, standing behaviour dominated the daily activity profile, generally accounting for over 85% of the total time. Lying was consistently present but limited, typically ranging from 5% to 10% of daily time, depending on the day. The proportion of unknown behaviour varied more substantially and appeared higher on certain days, which may be attributed to changes in lighting conditions affecting detection confidence. While Figure 7 highlights day variability in behavioural proportions, Figure 8 illustrates average circadian distribution of behaviours across the entire monitoring period.

In Figure 8 the average behavioural distribution per hour across the month is shown. This graph reinforces the nocturnal tendency of lying: between 12:00 am and 04:00 am, the percentage of lying behaviour increased, peaking at around 34% at 02:50. During daylight hours, especially between 06:00 am and 10:00 pm, standing behaviour accounted for over 90% of the time. Interestingly, the proportion of unknown detections increased notably during mid-morning to early afternoon (08:00 am to 02:00 pm). This pattern may reflect periods when the horse was more frequently outside the monitored area or when environmental conditions affected detection confidence.

3.3. Long term drinking behaviour monitoring

A further application of the developed model was the possibility to monitor for a long period some specific behaviours like drinking. In this work the drinking behaviour was detected using a pose estimation pipeline

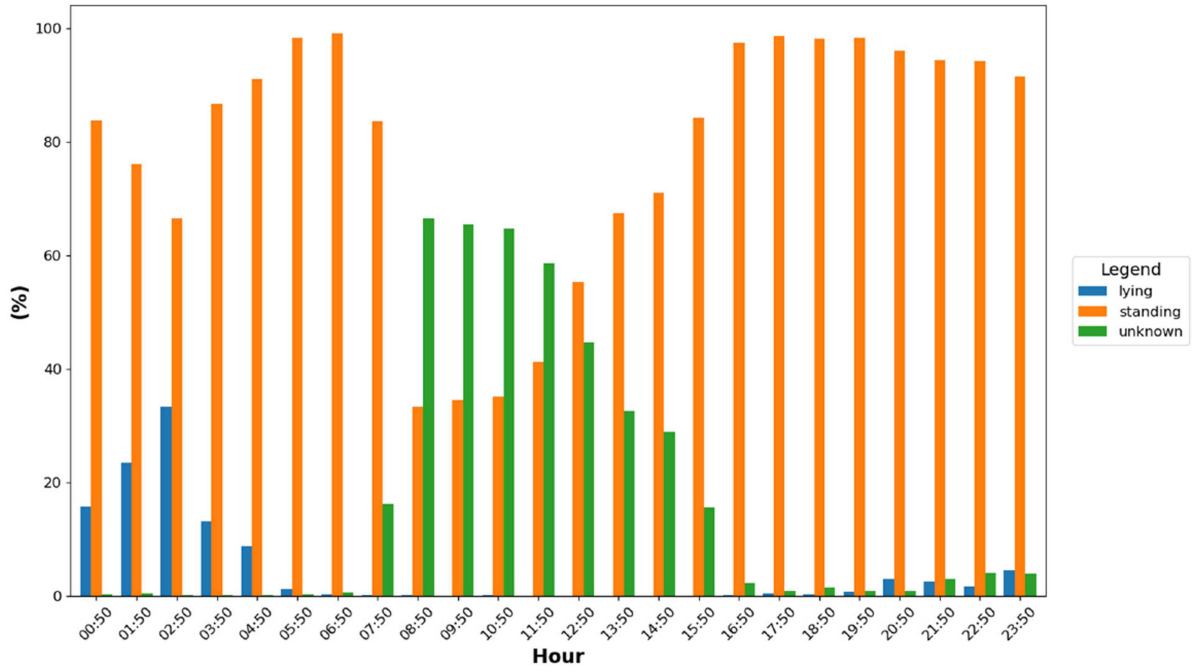

Figure 7. Daily percentage of lying, standing and unknown behaviours.

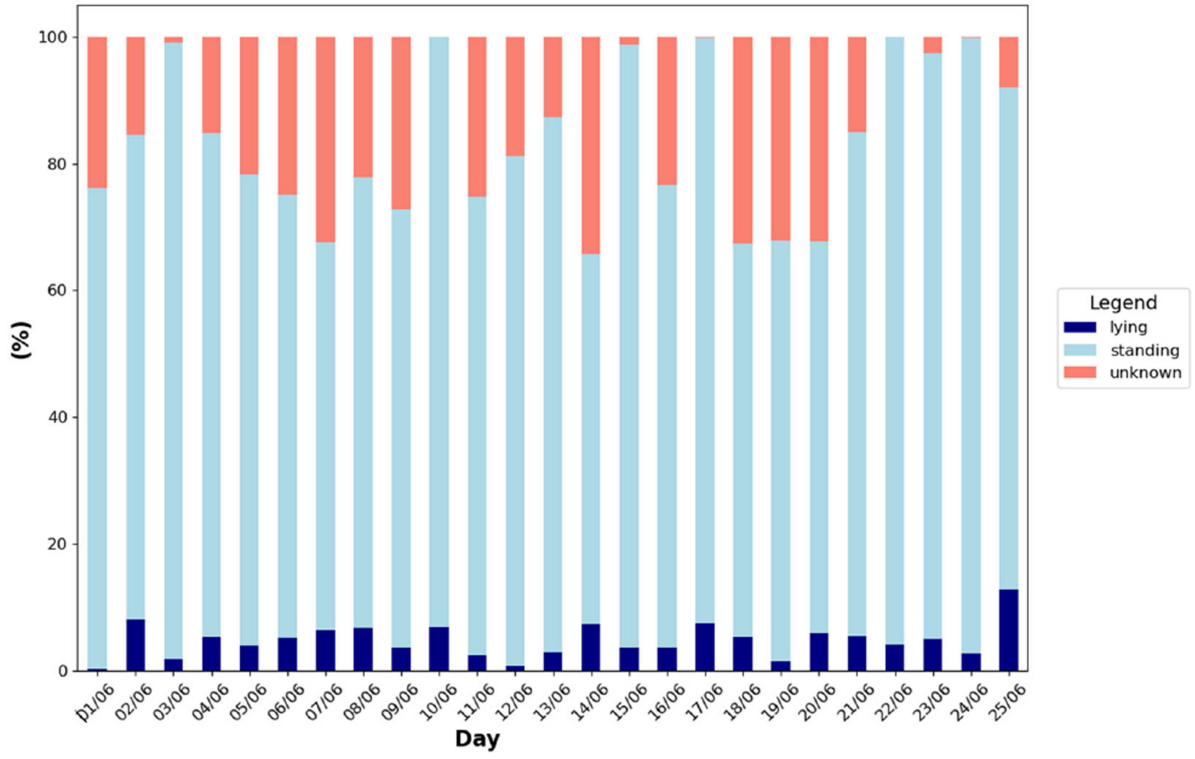

Figure 8. Average behavioural distribution per hour.

based on keypoint detection, allowing the system to identify specific postural configurations associated with the act of drinking. Unlike lying and standing, which were recognised through object detection, drinking required spatial information and was therefore inferred using a deep learning model capable of tracking body landmarks with temporal continuity. To assess that the horse was drinking, a Region Of

Interest (ROI) was drawn around the drinker. When the muzzle keypoint was detected in the ROI for more than 3 seconds, it was annotated as a drinking event by our model. The performance of drinking model was evaluated on the internal test subsets. The pose estimation pipeline obtained an F1-score of 0.71, recall of 0.57 and precision of 0.94. These findings show high reliability of positive detections, even though a proportion of drinking events were not identified, as reflected by the moderate recall value. Partial occlusions or changes in head posture with respect to the predefined ROI could be the cause of the decreased sensitivity. The temporal distribution of drinking behaviour is illustrated in Figure 9, where a heatmap displays the duration (in seconds) of drinking events across each hour and day. The behaviour appeared more sporadic, and patterns appear less clearly than in the previous case compared to lying (Giannone et al. 2025). Drinking occurred at various times throughout the day, with increased frequency during afternoon and evening hours, particularly between 04:00 pm and 10:00 pm. In some instances, drinking durations peaked above 160 seconds per hour, indicating extended drinking bouts. Unlike the consistent circadian pattern observed for lying, drinking events were less predictable and exhibited a higher degree of daily variability. The frequency data further supported this interpretation, revealing that on certain days the horse approached the water source repeatedly within short intervals, whereas on others, drinking activity was more limited or clustered within narrow time windows. The identification of drinking using pose-based deep learning methods highlights the potential of combining object detection and keypoint tracking for comprehensive behavioural monitoring. This multimodal approach allows the detection of complex or short behaviours that may otherwise be difficult to identify through bounding-box tracking alone.

4. Conclusions

Our findings show that computer vision can classify key behaviours of a stabled horse from continuous video data. Since many veterinary clinics are already equipped with surveillance cameras, computer vision could be used to automatically detect patterns that indicate behavioural anomalies. The frequency with

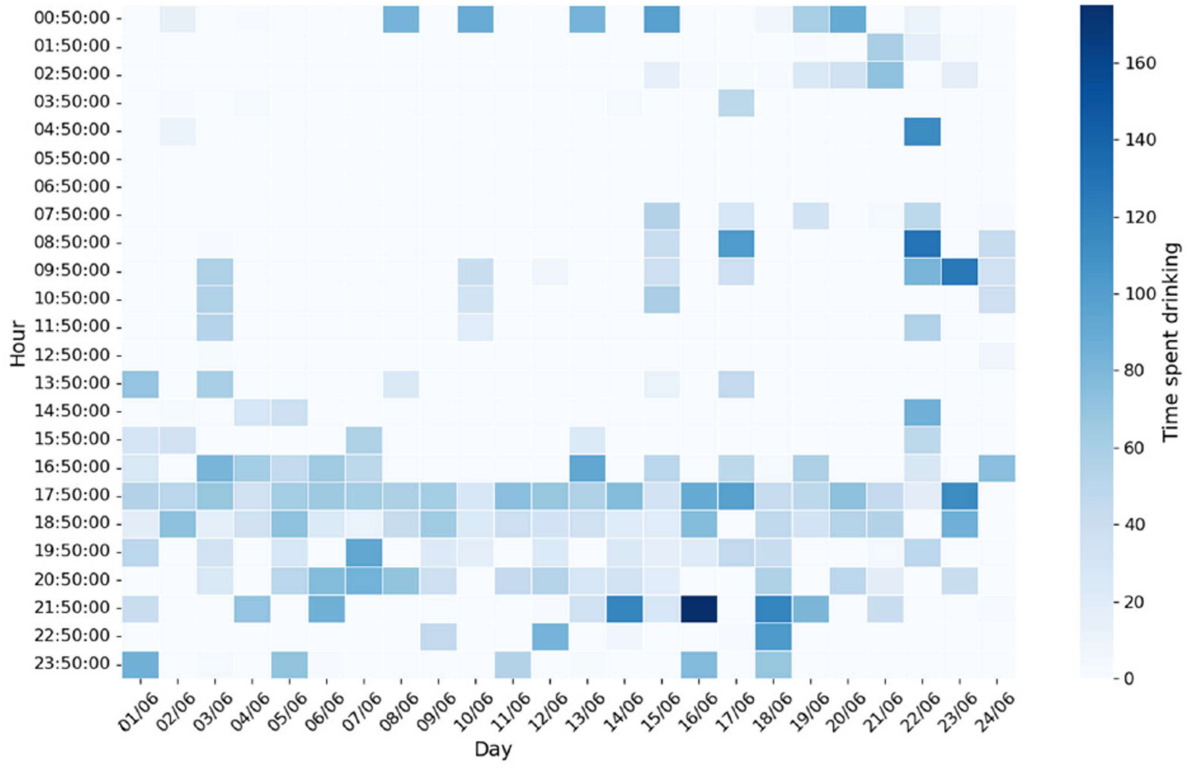

Figure 9. Heatmap showing the time spent drinking per hour over the monitored period.

which different behaviours alternate, defined behavioural variability, represents an adaptive strategy indicating overall good health, with a reduction in behaviour switching linked thanks to severe pain (Nowak et al. 2024). Since evaluating discomfort is indispensable aspect of effective clinical decision-making, the analysis of time budget and early identification of abnormal behavioural patterns could act as an early warning system, enabling veterinarians and other professionals to detect animals at risk before the appearance of clinical symptoms supporting intervention and helping to improve treatment. These behavioural indicators could help detect potential health problems. Moreover, this approach is a non-invasive alternative to traditional monitoring, and it does not require direct contact with the animal. The behaviour detection showed strong performance for standing and lying, revealing that lying typically peaked during the nocturnal hours, while drinking events detected were more common in the late afternoon and evening. The unknown class reached high recall (99.9%) but poor precision (21.3%), indicating a tendency to overpredict when the horse remained in view, a limitation to address in future work. The main limitation of this study is its dataset built on a single horse, which reduces the level of generalisation of the results. However, this study was designed as a pilot investigation to evaluate the practicality and efficiency of the suggested computer vision pipeline in realistic stable conditions. Future research will extend the dataset to more horses and different stable environments, allowing the model to better generalise. The dataset will be progressively expanded and made publicly available to allow further validation. Such improvements will strengthen the value of automated behaviour recognition systems for monitoring horse welfare.

Author contributions

CRediT: **Claudia Giannone:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft; **Chiara Maccario:** Conceptualization, Investigation, Visualization, Writing – original draft; **Emanuela Dalla Costa:** Conceptualization, Investigation, Project administration, Writing – review & editing; **Elie Atallah:** Writing – review & editing; **Marco Bovo:** Conceptualization, Supervision, Visualization, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Informed consent statement

Not applicable.

Funding

No funding has been received for the preparation of the current manuscript.

References

Aragona F et al. 2025. Pilot study: simultaneous daily recording of total locomotor activity and heart rate in horses for application in precision livestock farming. Animals (Basel). 15(9):1189. https://doi.org/10.3390/ani15091189

Auer U, Kelemen Z, Engl V, Jenner F. 2021. Activity time Budgets—A potential tool to monitor equine welfare? Animals (Basel). 11(3):850. https://doi.org/10.3390/ani11030850

Bertolucci C, Giannetto C, Fazio F, Piccione G. 2008. Seasonal variation in daily rhythms of activity in athletic horses. Animal: an international journal of Animal bioscience. 2:1055–1060. https://doi.org/10.1017/S1751731108002267

Boyd LE, Carbonaro DA, Houpt KA. 1988. The 24-hour time budget of przewalski horses. Applied animal behaviour science. Behavior of Przewalski horses. 21(1):5–17. https://doi.org/10.1016/0168-1591(88)90098-6

Dalla Costa E, et al. 2014. Development of the horse grimace scale (HGS) as a pain assessment tool in horses undergoing routine castration. PLoS One. 9(3):e92281. https://doi.org/10.1371/journal.pone.0092281

Diwan T, Anirudh G, Tembhurne JV. 2023. Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tools Appl. 82(6):9243–9275. https://doi.org/10.1007/s11042-022-13644-y

Dwyer B., Nelson J., Hansen T., et al. Roboflow (Version 1.0). [cited 2025 May 23]. Available from: https://roboflow.com/

Giannone C et al. 2025. Impact of the technology to monitor horse behaviour and health: a scoping review. J Equine Vet Sci. 155:105734. https://doi.org/10.1016/j.jevs.2025.105734

Giannone C et al. 2025. Automated dairy cow identification and feeding behaviour analysis using a computer vision model based on YOLOv8. Smart Agricultural Technology. 12:101304. https://doi.org/10.1016/j.atech.2025.101304
Gobbo E et al. 2025. Exploring the impact of housing routine on lying behavior in horses measured with triaxial accelerometer. Front Vet Sci. 12:12. https://doi.org/10.3389/fvets.2025.1572051
Kelemen Z, Grimm H, Long M, Auer U, Jenner F. 2021. Recumbency as an equine welfare indicator in geriatric horses and horses with chronic orthopaedic disease. Animals. 11(11):3189. https://doi.org/10.3390/ani11113189
Lamanna M, et al. 2025. Time-activity budget in horses and ponies: a systematic review and meta-analysis on feeding dynamics and management implications. Journal of Equine Veterinary Science. 154:105684. https://doi.org/10.1016/j.jevs.2025.105684
Lawin F et al. 2023. Is markerless more or less? Comparing a smartphone computer vision method for equine lameness assessment to multi-camera motion capture. Animals. 13:390. https://doi.org/10.3390/ani13030390
Martvel G, Shimshoni I, Zamansky A. 2024. Automated detection of cat facial landmarks. Int J Comput Vis. 132(8):3103–3118. https://doi.org/10.1007/s11263-024-02006-w
McDonnell SM. 2003. The Equid Ethogram: A Practical Field Guide to Horse Behavior. Eclipse Press. p 398.
Nowak M, Martin-Cirera A, Jenner F, Auer U. 2024. Time budgets and weight shifting as indicators of pain in hospitalized horses. Front Pain Res. 5, https://doi.org/10.3389/fpain.2024.1410302
Tomar S. 2006. Converting video formats with FFmpeg. Linux Journal. 2006(146):10. https://ffmpeg.org/.
Torcivia C, McDonnell S. 2020. In-person caretaker visits disrupt ongoing discomfort behavior in hospitalized equine orthopedic surgical patients. Animals (Basel). 10(2):210. https://doi.org/10.3390/ani10020210. PubMed PMID: 32012670;. PubMed Central PMCID: PMC7070845.
Torcivia C, McDonnell S. 2021. Equine discomfort ethogram. Animals. 11(2):2. https://doi.org/10.3390/ani11020580
Yang SX et al. 2025. Review of computer vision for livestock body conformation assessment. Agriculture Communications. 3(3):100099. https://doi.org/10.1016/j.agrcom.2025.100099
Yigit T, et al. 2022 Jul. Wearable inertial sensor-based limb lameness detection and pose estimation for horses. IEEE Trans Autom Sci Eng. 19(3):1365–1379. https://doi.org/10.1109/TASE.2022.3157793
Zuerl M et al. 2024. Automated long-term monitoring of stereotypical movement in polar bears under human care using machine learning. Ecological Informatics. 83:102840. https://doi.org/10.1016/j.ecoinf.2024.102840