Biosystems Engineering 242 (2024) 154–168

Contents lists available at ScienceDirect

Biosystems Engineering

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Comparison of Transformers with LSTM for classification of the behavioural time budget in horses based on video data

Albert Martin-Cirera [a,b,d,*], Magdelena Nowak [c], Tomas Norton [d], Ulrike Auer [c], Maciej Oczak [a,b]

[a] Precision Livestock Farming Hub, University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, 1210, Vienna, Austria
[b] Institute of Animal Husbandry and Animal Welfare, University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, 1210, Vienna, Austria
[c] Division of Anesthesiology, Institute of Animal Husbandry and Animal Welfare, University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, 1210, Vienna, Austria
[d] M3-BIORES: Measure, Model, Manage Bioresponses, Katholieke Universiteit Leuven, Kasteelpark Arenberg 30, B-3001, Leuven, Belgium

ARTICLE INFO

Keywords:
Equine welfare
Time budget
Computer vision
Multi-input classification
Multi-output classification

ABSTRACT

This study compares the performance of Transformers with LSTM for the classification of the behavioural time budget in horses based on video data. The behavioural time budget of a horse consists of amount of time of the activities such as feeding, resting, lying, and moving, which are important indicators of welfare and can be a basis of pain detection. Video technology offers a non-invasive and continuous monitoring approach for automated detection of horse behaviours. Computer vision and deep learning methods have been used for automated monitoring of animal behaviours, but accurate behaviour recognition remains a challenge. Previous studies have employed Convolutional LSTM models for behaviour classification, and more recently, Transformer-based models have shown superior performance in various tasks. This study proposes a multi-input, multi-output classification methodology to address the challenges of accurately detecting and classifying horse behaviours. The results demonstrate that the multi-input and multi-output Transformer model achieves the best performance in behaviour classification compared with single input and single output strategy. The proposed methodology provides a basis for detecting changes in behaviour time budgets related to pain and discomfort in horses, which can be valuable for monitoring and treating horse health problems.

1. Introduction

Horses divide their time between activities that allow them to satisfy their basic requirements i.e., feeding, resting, lying, and moving (Feist & McCullough, 1976; Mayes & Duncan, 1986; Sweeting, Houpt, & Houpt, 1985). These behaviours constitute the behaviour time budget of a horse (Auer, Kelemen, Engl, & Jenner, 2021). Moreover, horses exhibit a highly repetitive daily routine with almost identical time patterns of behaviours from day-to-day (Berger, Scheibe, Michaelis, & Streich, 2003; Boy & Duncan, 1979; Duncan, 1980; Yarnell, Hall, Royle, & Walker, 2015). The amount of time an animal engages in specific behavioural activities is considered an informative welfare indicator (Flannigan & Stookey, 2002; Goodwin, 1999; Hausberger et al., 2020; Sarrafchi & Blokhuis, 2013). Horses with pathogenesis of pain sensation show significantly different behaviour in terms of time spent eating, sleeping, and moving compared to horses without pain. In addition, changes in posture and specific body behaviour e.g., weight shifting, unbalanced posture and head position are indicators of pain in horses. Activity time budgets for specific behaviour in horses have been identified as sensitive indicators of equine discomfort, and thus could facilitate rapid detection of painful conditions and monitoring the success of therapeutic interventions (Clothier, Small, Hinch, Barwick, & Brown, 2019; Hausberger, Fureix, & Lesimple, 2016; Price, Catriona, Welsh, & Waran, 2003). In addition, horses with moderate or low pain may show unstable behaviours i.e., movements or swing their weight to reduce pain or put one front leg further forward than the other (Torcivia & McDonnell, 2020). Thus, subtle changes of the time budget can potentially be a basis for detection of mild, acute, and chronic pain, which would not otherwise be detectable with current available pain scales (Auer et al., 2021).

Subtle changes in the time budget should be analysed for each individual horse. Moreover, it is crucial to consider recording the baseline state in the equine hospital before treatment, as this allows the animal to serve as its own reference. As a result, the variation in time budgets during and after the treatment could potentially be an indicator of a

* Corresponding author. Precision Livestock Farming Hub, University of Veterinary Medicine Vienna (Vetmeduni Vienna), Veterinärplatz 1, 1210, Vienna, Austria.
E-mail address: 12135451@students.vetmeduni.ac.at (A. Martin-Cirera).

https://doi.org/10.1016/j.biosystemseng.2024.04.014
Received 11 August 2023; Received in revised form 18 April 2024; Accepted 23 April 2024
Available online 5 May 2024

Nomenclature

$\overrightarrow{v}$	Vector or link between two key body points
f_x	Horizontal pixel position
f_y	Vertical pixel position
$\overrightarrow{v_o}$	Origin vector
θ	Vector angle
d	Vector direction
fr_x	Horizontal pixel position rescaled
fr_y	Vertical pixel position released
$\lvert\overrightarrow{v}\rvert n$	Magnitude of the vector normalised
θr	Vector angle rescaled
w_l	Long window
w_s	Short window
$\overrightarrow{x}$	Feature vector
X_l	Matrix feature vector with long window
X_s	Matrix feature vector with short window
L	Total loss function
L_{SM}	Loss function for short window and multi-classification tasks
L_{LM}	Loss function for long window and multi-classification tasks
L_{SB}	Loss function for short window and binary-classification tasks
L_{MM}	Loss function for multi-input and multi-classification tasks
b	Behaviour
B	number of behaviours
tp	True positive
fn	False Negative
% *Error*	Difference between the true and the predicted tant per cent time
% *true time*	True tant per cent time
% *predicted time*	Predicted tant per cent time

Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
TLM	Time-lapse mode
HRNet	High-Resolution neural network model
PCK	Percentage of Correct Keypoints

higher or lower level of pain.

Real-time analysis of equine behaviour time budgets may also facilitate early detection of other health problems, such as colic, lameness or other painful conditions (Ashley, Waterman-Pearson, & Whay, 2005; Lesimple, 2020). Using video technology for automated detection of horse behaviours mitigates certain drawbacks associated with direct human. This is crucial since horses may hide certain behaviours, such as signs of pain, in the presence of humans (Torcivia & McDonnell, 2020). Camera technology enables continuous 24/7 behaviour recording (Frost et al., 1997), while computer vision can detect brief, interspersed behaviours occurring over extended periods (Martin, Prescott, & Zhu, 1992).

Cameras, in conjunction with computer vision, have the potential to positively impact the well-being of animals. Neethirajan (2020) employed cameras and computer vision for the automated monitoring of animals and behaviours closely associated with growth status. Cangar et al. (2008), posture behaviour of pregnant cows nearing calving was monitored to assess the necessity for human intervention. Nilsson et al. (2015) outlined a surveillance method for analysing pig behaviours in pens through image analysis, offering the prospect of enhancing animal welfare by automatically detecting signs of abnormal behaviours. Oczak et al. (2014) used activity index in image and neural networks to recognise aggressive behaviours. Furthermore, Liu et al. (2020) located instances of pig tail-biting, while Chen, Zhu, Oczak, et al. (2020) developed a system to recognise pig interactions with various objects, aiming to reduce occurrences of tail-biting and aggression. The physical damage resulting from a pig bite, whether to the tail or other areas of the body, can induce severe pain and secondary infections that may spread throughout the victim pig's body. Therefore, the integration of cameras and computer vision holds the potential to proactively prevent unhealthy behaviours or respond more promptly and effectively, ultimately enhancing animal welfare.

When animal motions should be quantified using computer vision with high-fidelity, a promising approach to process the video data by first detecting the key body points to estimate the skeleton or spatial features followed by modelling the temporal features of the skeleton (Si, Chen, Wang, Wang, & Tan, 2019; Song, Yu, Yuan, & Liu, 2021; B. Yu, Yin, & Zhu, 2017). On the contrary, approaches based on RGB video primarily focus on the development of spatial and temporal representations extracted from RGB frames and temporal optical flow (Chen, Zhu, Steibel, et al., 2020; Wang et al., 2021). These methods may encounter certain limitations, including issues related to background clutter, illumination variations, and appearance diversity, among others. Pose estimation data delineates the body structure through a collection of 2D coordinate positions corresponding to key joints. As skeletal sequences lack colour information, they remain unaffected by the constraints associated with RGB video. However, a potential drawback of this approach is that inaccurate detection of keypoints may lead to erroneous behaviour recognition (Duan, Zhao, Chen, Lin, & Dai, 2022, pp. 2969–2978). Therefore, it is important to start with a powerful keypoint estimation model such as the High-Resolution neural network (HRNet) model designed by Sun, Xiao, Liu, and Wang (2019). The effectiveness of the HRNet network has substantiated through superior keypoints detection outcomes over three benchmark datasets: the COCO keypoint detection dataset Lin et al. (2014), the MPII Human (Andriluka, Pishchulin, Gehler, & Schiele, 2014) and AP-10 (H. Yu et al., 2021) a large-scale benchmark for general animal pose estimation.

Once the keypoint model is chosen a key challenge is classifying animal behaviour using the detected key points. Models available for this task include the specific Recurrent Neural Networks (RNN) called Long Short-Term Memory (LSTM) designed by Hochreiter and Schmidhuber (1997), Bidirectional LSTMs (BiLSTM), and Transformer-based models (Vaswani et al., 2017). LSTM has had a transformative impact on machine learning and neural computing, addressing the challenge of the exploding or vanishing gradient problem (Van Houdt, Mosquera, & Nápoles, 2020). Utilizing three gates, LSTM modulates information flow to prevent gradient vanishing and explosion. BiLSTMs process input in both directions, leveraging past and present information for real-world time-series analysis. Research by Siami-Namini, Tavakoli, and Namin (2019) suggests that BiLSTM-based models yield more accurate predictions in time series problems but take longer to reach equilibrium than LSTM-based models. In the field of animal monitoring LSTM's have been used for behaviour recognition (Chen, Zhu, & Norton, 2021; D. Liu et al., 2020; Yin, Wu, Shang, Jiang, & Song, 2020) to detect aggressive behaviours such as tail biting in pigs or basic behaviours such as drinking, feeding, walking, laying, and standing using an input window size between 30 and 60 frames.

More recently, Transformers-based models, where the key

component is the self-attention architecture, have shown success in various artificial intelligence (AI) domains, attracting interest from different domains (T. Lin, Wang, Liu, & Qiu, 2021). Transformers have achieved superior performances in many tasks in natural language processing and computer vision, which also triggered great interest among scientists working on the time series analysis. Among the multiple advantages of Transformers, is the ability to capture long-range dependencies and interactions, which is especially useful for time series modelling, leading to progress in time series analysis (Ahmed et al., 2022; Horn, Moor, Bock, Rieck, & Borgwardt, 2020, pp. 4353–4363; Wen et al., 2022) and improved performance of models used in applications such as human activity recognition (Ijaz, Diaz, & Chen, 2022; Shavit & Klein, 2021).

Automated detection of pain is especially important in a horse hospital, where most animals are in some level of pain and receive medication like fluids, analgesics, or antibiotics regarding their underlying health problems. Automated detection of pain in these horses might be useful when deciding on the type and amount of analgesics. The main objective of this study was to automatically classify behavioural time budgets of horses housed in a horse hospital. Any equine behaviour that involves movement, such as walking, eating, or unstable, can be exhibited at various speeds. Additionally, multiple behaviours can be observed simultaneously, for instance, moving and feeding. This presents a challenge in accurately detecting and classifying behavioural time budgets. To address these issues, we propose a novel multi-input, multi-output classification methodology. The purpose of this approach was to detect behaviours that may be masked by other behaviours with excessively long duration, as well as behaviours that may be misclassified due to occurrence of multiple behaviours at the same time. The one proposed in this study is a complex structure designed to better discern various behaviours and thereby achieve improved results. Our hypothesis is that this proposed method increases the F-Score of behaviour classification, as it takes into consideration the variability in behaviour duration and the fact that behaviours overlap with each other. We aimed to identify the optimal state-of-the-art computer vision and time series analysis for behaviour classification by comparing the performance of Transformers, LSTMs, and Bidirectional LSTMs on our dataset. The developed technique should be applicable for relating changes in behaviour time budgets to pain and discomfort behaviour in subsequent research.

2. Materials and methods

2.1. Experimental setup

2.1.1. Animals and housing

Studies were conducted at the clinic of the University of Veterinary Medicine Vienna. In total, 10 videos with 10 horses were recorded for behaviour classification. The horses had different pain levels. Additionally, 36 videos with 27 horses, were recorded for keypoint detection. Thus, yielding two datasets: a primary dataset of videos for keypoint detection and a secondary dataset of videos for behaviour classification. Video recordings started during the morning feeding time, between 6 and 8 a clock. Horses were fed 4 times per day with hay, with higher caloric density. Feed rations for horses with colic were reduced. Some horses received medication like fluids, analgesics, antibiotics regarding their underlying health problem. However, for the sake of the study, at this stage, medication remained unknown. Each horse was kept individually in a square pen with dimensions of 4 m × 4 m. Additionally, 36 other videos were used with 27 different horses to train a model for extracting keypoints.

2.1.2. Video recording

The videos for post hoc analysis were recorded with Gopro Hero 4 action camera (San Mateo, USA) in a time-lapse mode (TLM) of two frames per second (fps). The wide-angle lens allowed an overview of the whole box and a horse (Fig. 1).

The advantage of using the GoPro Hero 4 camera was that it could be moved between pens easily. The camera was powered with battery packs that provided power sufficient for more than 24 h of active recording. We decided to use TLM with 2 fps to limit the memory size of the videos and compress 3 h of real duration in a video length of approximately 12 min. Real-time of 15 min had a duration of 1 min in the video. The quality of the video (1440 × 1080 pixels resolution) was estimated as good enough to visually identify horse key body points and behaviour even with quick movements of the horse.

2.2. Data labelling

2.2.1. Frames selection for keypoint labelling

To train a keypoint detection algorithm for automated detection of the keypoints of the horse's body, it was necessary to select a set of frames that had high variance and the maximum relevant information about the horse's poses. This approach can reduce the size of the dataset, decrease the use of computational resources needed to train the algorithm for keypoint detection, and reduce the workload related to manual labelling of keypoints on images by human labellers.

In the first step, we selected 36 videos with 27 different horses, which constitute the dataset for keypoint detection. Each video had a duration of 3–4 h in real-time. The videos were recorded at different times of the day. We selected videos with horses of different sizes, colours, and shapes, with and without the presence of humans in the box and even without the presence of the horses. We selected empty boxes to imply that a horse is not always present inside. The above steps were crucial to obtain a good diversity of data. In total, around 780,000 frames (3 h * 3600 s per hour * 2 frames per second * 36 videos) were recorded. In a second step, we applied the k-means clustering algorithm described in Pereira et al. (2019),. Application of the K-means algorithm allowed reduction of the similarity between the sampled images. A total of 6000 frames with 5899 frames with a horse and 101 frames without a horse were selected with the K-means algorithm. It was possible to label the key body points on this number of images within 2 months by one vet labeller (20 h/week).

2.2.2. Labelling horse key body points

The following key body points were labelled on 6000 images selected with K-means algorithm: nose, withers, tail, ears, and limbs (Table 1). Each labelled key body point was connected to at least one other key body point with a link connection. All link connections are listed in Table 1 and their visualisation is presented in Fig. 6.

To increase the accuracy of annotations of the key body points, we assigned the location of keypoints to the edges of horse body parts. E.g., an ear key body point was defined as located on the contour of the ear or a limb was defined as located at the end of the front of the hoof.

2.2.3. Labelling horse behaviour

The behaviour of 10 horses (Table 2), distinct from those used for keypoints detection, was recorded using a frame rate of 2fps, which constitute the dataset for behaviour classification. As horses tend to hide some behaviours such as pain in the presence of humans, we decided to exclude the frames with human presence from the dataset used for labelling of horse behaviour.

Four-time budget behaviours i.e., lying, resting, feeding/foraging, moving and one discomfort behaviour i.e., unstable were labelled on the videos (Table 3). Furthermore, some of these behaviours were expressed by horses simultaneously i.e., moving during foraging or rolling during lying. These behaviours were included in Table 3. It was possible to label this data set within 2 months by a labeller, who was a veterinarian (5 h/week). Behaviours had to be performed for at least 10 s to be labelled.

Fig. 1. Horses in individual pens. a) Feed provided on the ground. b) Feed provided in a sack and on the ground.

Table 1
Key body points and link connections.

Key body points	Definition and annotation	Link connection
Nose	Area including the nose bridge and mouth.	Withers
Withers	The point in the shoulder area where the neck merges with the back and the mane ends.	Ears, nose, front limbs, and tail.
Tail	Where the tail base is located.	Withers and hind limbs.
Ears	Edge of the ear.	Withers
Front and hind limbs	The cranial point of the hoof of each limb.	Front limbs with withers and hind limbs with tail.

Table 2
Dataset with horse behaviour.

Horse	Duration of video (HH:MM:SS)	Total number of frames	Number of frames with human presence	Number of frames without human presence
#1	01:11:49	8619	0	8619
#2	02:52:59	20,759	521	20,238
#3	02:57:44	21,329	17,617	3712
#4	02:51:22	20,564	410	20,154
#5	02:42:44	19,529	230	19,299
#6	02:50:29	20,459	32	20,427
#7	02:41:22	19,364	1001	18,363
#8	02:57:44	21,329	36	21,293
#9	02:50:44	20,489	908	19,581
#10	02:52:29	20,699	1199	19,500
Total	26:49:26	193,140	21,954	171,186

2.3. Datasets

Datasets for the training of two types of models used in our study i.e., for key body point detection and behaviour classification were split into 3 subsets i.e., training, used to fit the model, validation, used to provide an unbiased evaluation of a model fit during training and test, used only to assess the final model performance.

The dataset used to train the model for keypoints detection was composed of 6000 images. From this dataset, 900 (15%) images were randomly assigned to the validation set, 900 (15%) to the test set and 4200 (70%) images were used to train the model.

Automatically detected key body points of a horse were used as the input data for behaviour classification representing the spatial modality. In total, we used 18 feature variables extracted from key body points location as each of 9 labelled key body points had 2 feature variables i.e., first on the horizontal axis f_x and second on the vertical axis f_y of the pixel position in the frame. The keypoints themselves contained relevant information that was used to gain insights into the posture of the animal.

Table 3
Definitions of horse behaviours.

Behaviour*	Definition
Feeding/foraging	The horse was eating, chewing, biting, sniffing the feed located on the ground or in the feeder or was searching for the food (Goodwin, Davidson, & Harris, 2002; Thorne, Goodwin, Kennedy, Davidson, & Harris, 2005).
Lying	The horse was on the ground with the head and legs extended or the horse was in sternal recumbence with the head in upright position and legs folded under the body (Duncan, 1980).
Moving	The horse was moving around more than 3 steps in the pen.
Resting	The horse was standing but not moving at all. Ears and tail can move (Duncan, 1980).
Unstable	The horse with moderate or low pain may show unstable behaviours, swinging their weight to reduce pain or put one front leg further forward than the other.
Behaviours overlapped	
Moving during foraging	Moving during feeding or searching for the food spread on the ground.
Moving during unstable	Moving less than 3 steps in any direction caused by unstable behaviour.
Lying during rolling	Horse was lying and rolling on the ground.

The distance and angle between keypoints were identified as significant factors in assessing a horse's posture. There were 8 link connections (vectors) between key body points (Fig. 2). Out of 8 link connections 5 were defined in the origin of coordinates of withers and 3 in the tail. Each of the 8 vectors had 2 feature variables i.e., magnitude, which expressed the distance between key body points and angle. Magnitude was calculated according to Eq. (1),

$$|\vec{v}| = \sqrt{\left(f_{x,b} - f_{x,a}\right)^2 + \left(f_{y,b} - f_{y,a}\right)^2} \quad (1)$$

where, $|\vec{v}|$ represents the magnitude of the vector, m represents the index vector, f_x and f_y denote the relative horizontal and vertical positions, respectively, of the pixel position within the frame, a and b represents the origin and the end of the vector, respectively. Angle was calculated according to Eq. (2),

$$\theta = d \bullet cos^{-1}\left(\frac{\vec{v} \bullet \vec{v_o}}{|\vec{v}| \bullet |\vec{v}|}\right) \quad (2)$$

where, θ represents the angle of the vector, d represents the direction of the vector, $\vec{v}$ represents the vector, $|\vec{v}|$ represents the magnitude of the vector and $\vec{v_o}$ represents the origin vector declared as a horizontal vector (1,0) where the angles were calculated.

Because of the position of the camera utilised for video recording in our study (Figs. 1 and 2), some key body points on some of the frames in our video dataset were occluded by the parts of the horse body or were

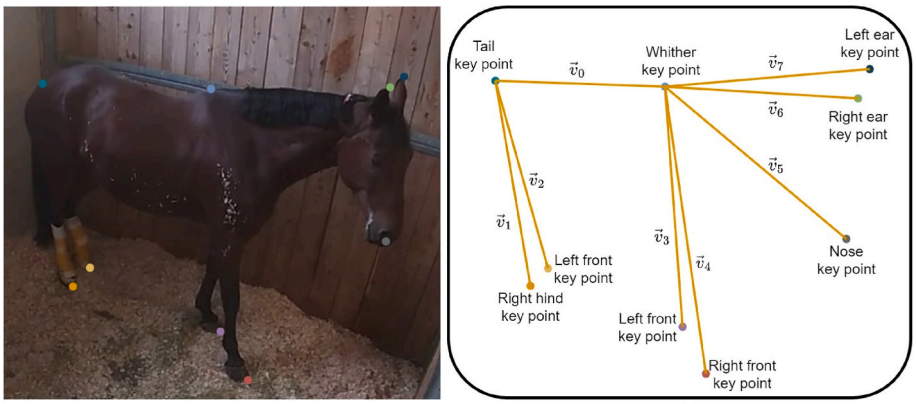

Fig. 2. Detection of the skeleton of a horse. a) Recorded image with key body points b) two-dimensional vectors.

simply not visible within the frame. To avoid variability in the number of features in consecutive frames, a new feature variable was added to each of the keypoints, indicating whether the keypoint was visible (value equal to 1) or not (value equal to 0), following the schema of COCO (Lin et al., 2014) annotation format for key body points. One more variable was introduced to indicate whether a connection existed between key body points. If the key body point was not detected in the image, the connection from this key body point to the other key body point could not exist. In total, the number of feature variable was 51 (9 *keypoint* × 3 *features variable* + 8 *vectors* × 3 *features variable*), where this feature variables representing spatial modality denoted as a vector $\overrightarrow{x}$ for each video frame.

Given that f_x and f_y are features variables that indicate the position of pixels within the frame, it is noteworthy that f_x could range from 0 to 1440, while values of f_y could range from 0 to 1080. The angle could range between −1 and 1, and the magnitude could have a range of $\pm\sqrt{width^2 \bullet height^2}$. All the features' variables were rescaled using a mean normalisation. It was done to unify a scale of all feature variables. Mean normalisation was calculated according to Eq. (3),

$$fr_x = \frac{f_x}{width}, fr_y = \frac{f_y}{height}, |\overrightarrow{v}|n = \frac{|\overrightarrow{v}|}{2\sqrt{width^2 \bullet height^2}}, \theta r = \frac{\theta}{2} \quad (3)$$

where fr_x represents the rescaled value of f_x obtained by dividing it by the width resolution of the camera (1440). Similarly, fr_y represents the rescaled value of f_y obtained by dividing it by the height resolution of the camera (1080) and θr represents the angle value rescaled to a common scale between 0 and 1.

To convert the data $\overrightarrow{x}$ with all the spatial feature variables associated with each frame into temporary data format, we added a second dimension to the dataset represented by "window" (*w*). The window defines the number of frames in a time unit. Addition of *w* resulted in transformation of our dataset into matrix $X \in R^{w \times 51}$ consisting of vectors $\overrightarrow{x}$.

2.4. Model

Fig. 3 illustrates how the models have been implemented, with the first step in automated detection of time budgets was to detect key body points (spatial domain) with HRNet model on each frame of the video with horses. In the second step, the dataset with key body points was divided into batches with a length equal to the window or number of frames resulted in the creation of a sample denoted as *X*. Subsequently, a sliding window of one-step was used and LSTMs and Transformers to extract spatial-temporal features applied to classify behaviours. Finally, based on the results of behaviour classification for each horse, the corresponding time budgets were estimated.

2.4.1. Keypoint detection

HRNet-32 was used to detect the key body points in horses. Output data from the automated detection of HRNet was trained and fine-tuned using MMPose (v.0.29.0) and PyTorch framework (Contributors, 2020).

HRNet was pre-trained on the COCO keypoint detection dataset using 1333 × 800 image resolution, with input size of 512 × 512 and fine-tuned for the task of keypoints detection in horses on our dataset. A single GPU (24 GB NVIDIA GeForce RTX 3090) was used to fine-tune the HRNet for a maximum of 100 epochs, using the learning rate as recommended in the framework with adjusted regularization parameter to avoid model overfitting.

We used the Percentage of Correct Keypoints (PCK) to quantify HRNet's performance (Yang & Ramanan, 2013). This metric evaluates how many automatically detected keypoints are located within a bounding box, which has a centre at the labelled key body point. PCK defined the width and height of the bounding box as 20% of the width of the human's torso (PCK@0.2). We followed the same methodology using 20% of the width of a horse's torso. We analysed for which value

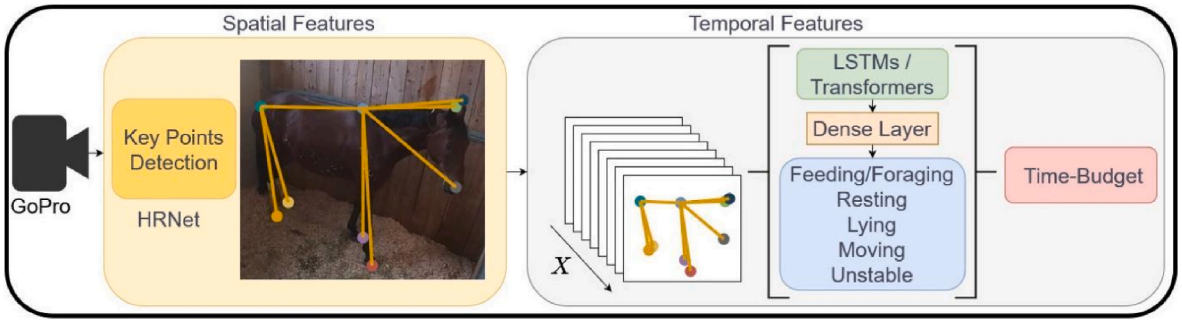

Fig. 3. Two steps in time budget estimation a) key body point detection b) behaviour classification.

of HRNet's confidence the PCK was the highest and selected this model confidence as the optimal.

2.4.2. Behaviour classification

We compared the performance of three state-of-the-art multivariate time series classification methods i.e., LSTM's (Fig. 4a), Bidirectional LSTM (Fig. 4b) and Transformers (Fig. 4c) to recognise behaviours that constitute horse time budgets. In our Transformer application, we used the encoder block, comprising a multi-head self-attention module and a position-wise feed-forward network (FFN).

Models used to extract temporal features i.e., LSTMs and Transformers and dense layers were structured with two outputs i.e., multi-class, and binary and multi-input. The proposed model introduces a novel structure incorporating both multi-output and multi-input components for behaviour classification, representing the first instance, to the best of our knowledge. The primary objective is to alleviate inaccuracies in classifying behaviours in horses within continuous recordings.

The application of multi-output methodology, as illustrated in Fig. 5, was implemented to reduce misclassifications between closely related behaviours e.g., moving behaviour and foraging or moving and moving during unstable. The first output consisted of a multi-classification task, which was applied to the basic and discomfort behaviours such as feeding, lying, moving, resting, and unstable. These behaviours were used to compute the time budget. Binary-classification was used to distinguish between movement behaviours and non-movement behaviours. Resulting in two distinct loss functions: one for multi-classification employing a softmax activation function, and another for binary-classification employing a sigmoid activation function. To optimise the model, both losses are combined. Therefore, the loss is minimum when it is capable of correctly predicting both outputs correctly. As a result, moving during foraging was appropriately classified as feeding. Moving during unstable behaviour was classified as unstable. Lying during rolling was classified as 'moving'.

The implementation of a multi-input approach was designed to address the fact that equine behaviours may occur at varying speeds and different time durations. This makes the use of one unique window for optimal recognition ineffective. In cases where a behaviour occurs in a short time, the application of a window that is too long may result in a misclassification of the behaviour. Conversely, the adoption of a window that is too long may hinder the identification of shorter behaviours, which may become occluded by the longer behaviours. To address this issue, two models with different window sizes were implemented, each centred at the same time (as shown in Fig. 5). The first model used a long window or 'w_l' (60 frames or 30 s) as input, resulting in 'X_l' sample while the second model used a shorter window or 'w_s' (20 frames or 10 s), resulting in 'X_s' sample. The used window sizes were decided based on manual inspection of several videos, wherein two values were selected that were deemed adequately disparate to enable the recognition of both short and long duration behaviours in horses. Subsequently, a loss function was assigned to each model. The dense layers in both models were concatenated to create a new output layer, which was used to predict the behaviour that occurred in the midpoint of both windows (frame highlighted in red in Fig. 5). It is worth noting that since the moving behaviour mostly occurred in short periods, the multi-output classification was applied only using the short window.

The total loss function in multi-class classification was calculated through the summation of all four-loss functions, which guaranteed that

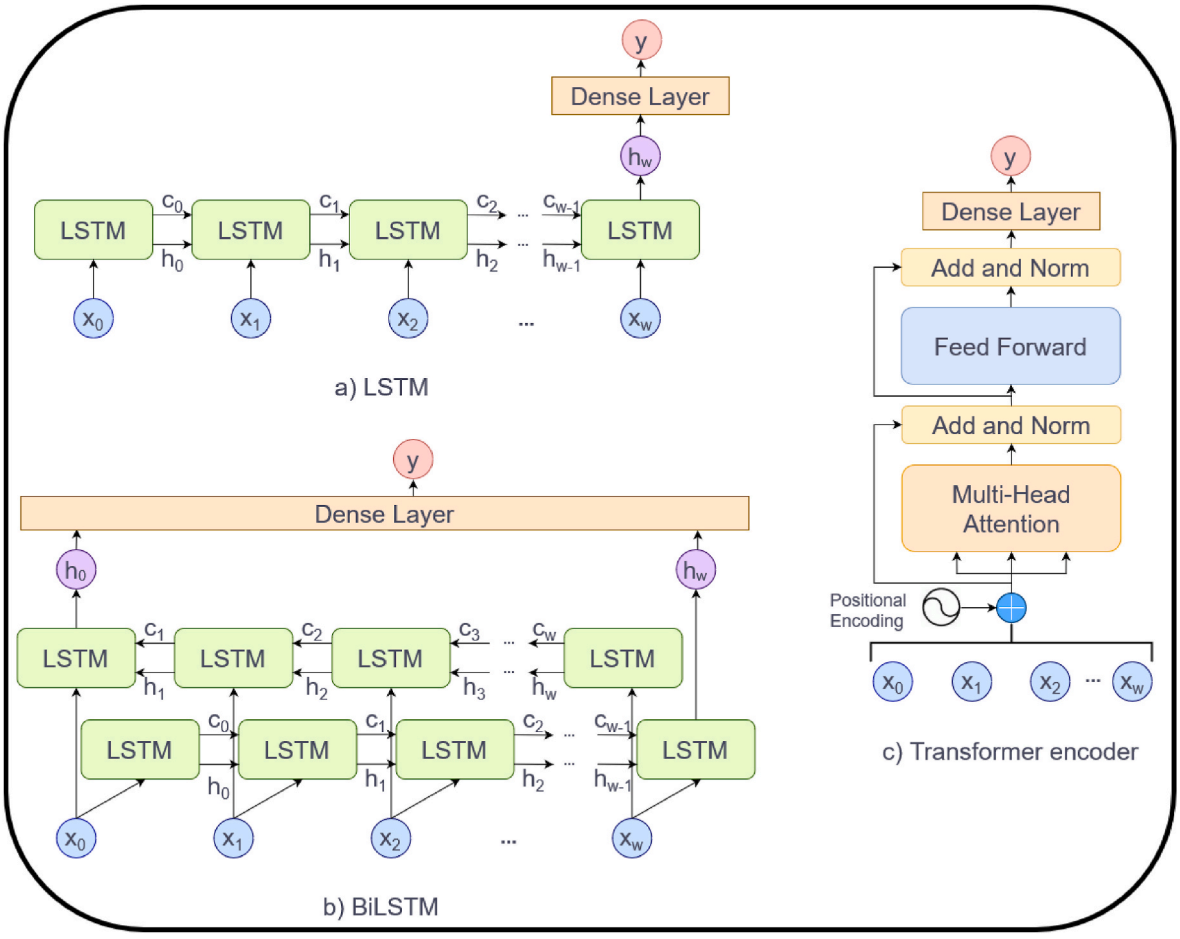

Fig. 4. Models' architectures a) LSTM. b) BiLSTM. c) Transformers.

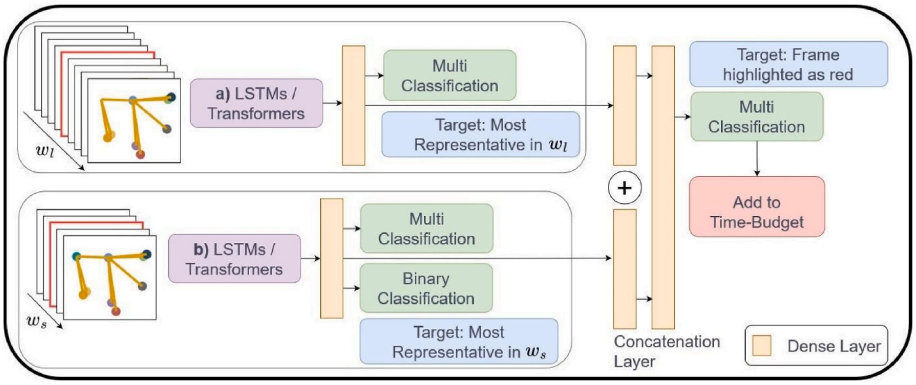

Fig. 5. Behaviour classification using multi-input and multi-output.

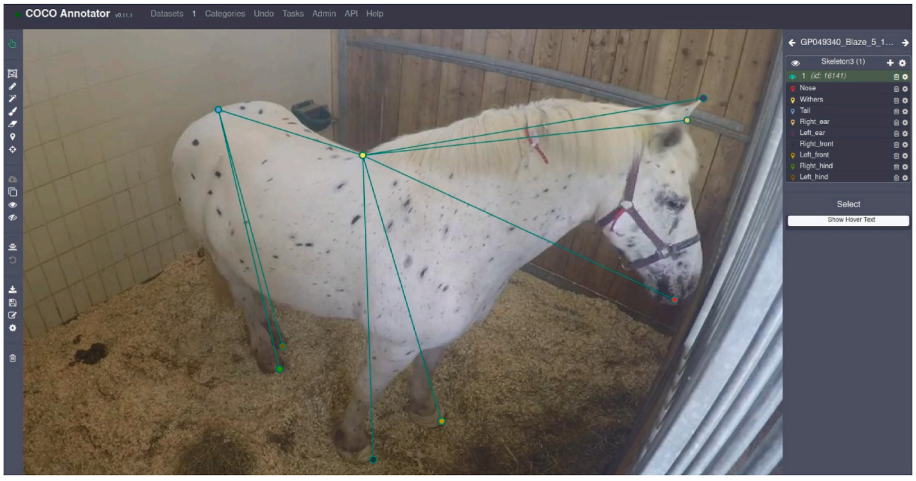

Fig. 6. Keypoints and skeleton of the horse annotated using COCO Annotator tool. Annotated keypoints: ears, nose, withers, tail, and limbs.

the model was subject to penalty for any misclassification during the training process. It was calculated according to Eq. (4),

$$L = L_{SM} + L_{LM} + L_{SB} + L_{MM} \tag{4}$$

where L is the total loss function, L_{SM} represents the loss function for short window and multi-classification tasks, L_{LM} represents the loss function for long window and multi-classification tasks, L_{SB} represents the loss function for short window and binary-classification tasks and L_{MM} represents the loss function for multi-input and multi-classification tasks.

Both methodologies, multi-output, and multi-input, constitute a complex structure, which, once trained and optimised, can better discern horse behaviours, whether occurring simultaneously or at different speeds.

To analyse the proposed methods, three experiments were conducted. The first experiment involved comparing the various proposed models (LSTM, BiLSTM and Transformers) with a single output and multi-output. The second experiment involved comparing the different models with a single input and multi-inputs. Finally, a third experiment was conducted with the test dataset subjected to leave-one-out cross-validation (Berrar, 2018), to analyse the performance of the models outside of the common data environment.

To sample data for the training dataset a sliding window was used with the length w and steps equal to one time point. This meant that between consecutive samples, there was an overlap equal to the window size w minus 1. As training, validation and test sets were randomly sampled, there was an overlap of part of data windows in some of the data samples between 3 data subsets i.e., training, validation, and test. To avoid it we removed data samples overlapping between 3 data subsets, to mitigate any possible bias towards the training dataset.

As our dataset was imbalanced, we used the algorithm performance metrics appropriate for this type of datasets. F-Score (Sokolova, Japkowicz, & Szpakowicz, 2006) was calculated according to Eq. (5) and balanced accuracy (Grandini, Bagli, & Visani, 2020) was calculated according to Eq. (6),

$$F - Score = \frac{2tp_b}{2tp_b + fp_b + fn_b} \tag{5}$$

where b represents each class or behaviour, tp_b represents the true positive to the corresponding class, fp_b represents the false positive to the corresponding class and fn_b represents the false negative to the corresponding class.

$$Balanced\ accuracy = \frac{1}{B}\sum_{b=1}^{B}\frac{tp_b}{tp_b + fn_b} \tag{6}$$

where B is the number of classes (5 behaviours), b represents each class or behaviour, tp_b represents the true positive to the corresponding class and fn_b represents the false negative to the corresponding class.

TensorFlow version 2.11 framework was used to modify the model's architecture. A single GPU (24 GB NVIDIA GeForce RTX 3090) was used to train the models during a maximum of 50 epochs, using a batch size of 64. Early stop and restore best weights were used to save the best model and Keras Tuner was used to find the best hyper-parameters as learning rate, batch size, number of units of the hidden states, number of units in the dense layers and dropout values.

2.4.3. Time-budgets

To determine the time budget for individual behavioural categories, we utilised a sliding window technique. This approach involved sequentially processing video frames, each iteration progressing by a single frame. Cumulative aggregation of results for each step was performed for predicted behavioural categories. The overall time budget was then calculated by summing predicted durations for each behavioural category across all iterations.

The predicted time budgets for each behaviour and horse were compared to the previous manually annotated values. The time budget error was calculated as the sum of the differences between the predicted and true values for each behaviour, according to Eq. (7),

$$\%\ Error = \frac{1}{2}\sum_{b=1}^{B} \%\ true\ time_b - \%\ predicted\ time_b \quad (7)$$

Where % *Error* the difference between the true tant per cent time and the predicted tant per cent time for each behaviour, B is the number of classes (5 behaviours), b represents each class or behaviour, % $true\ time_b$ represents the true tant per cent of the corresponding behaviour and % $predicted\ time_b$ represents the predicted tant per cent of the corresponding behaviour.

3. Results

3.1. Dataset results

3.1.1. Keypoints

COCO Annotator tool V0.11.1 (Brooks, 2019) was used to label the key body points of horses and their skeleton (Fig. 6).

The dataset for key body point labelling with 6000 images was recorded on 27 horses. A total of 6000 images that were annotated, where each image had one horse. It was not possible to label all key body points on every image in the dataset due to occlusions (Table 4).

3.1.2. Behaviours time budget

Loopy tool (Loopbio, Vienna, Austria), was used for labelling behaviours recorded in videos to create a reference dataset based on which further data analysis could be performed (Fig. 7).

A total of 10 videos were manually annotated, distinct from those used for key point detection and each featuring one horse per video, resulting in a cumulative duration of 26 h and 49 min in real time. As shown in Table 5, not all horses exhibited all five behaviours, with some horses lacking the behaviours of lying, feeding, or moving. It is noteworthy that one of the horses (horse number 3) demonstrated protracted periods of moving behaviours, in contrast to the other horses, where movement was predominantly of a brief duration and occurred infrequently.

As observed in Table 5, the overlapping behaviours have been grouped under a single label, given their relatively infrequent occurrence compared to other behaviours, and their significant similarities among themselves.

Table 4
Annotated key body points in a dataset with 27 horses.

Horse	No. Images labelled
Nose	5055
Withers	5819
Tail	5883
Right ear	4746
Left ear	4845
Right front limb	4793
Left front limb	4796
Right hind limb	4416
Left hind limb	4169

3.2. Model results

3.2.1. Keypoint detection

HRNet model was evaluated in Fig. 8, after training the model with the train and validation datasets labelled in section 3.1.1. The evaluation involved metrics such as PCK and confidence, to identify the optimal value that minimised false positives and false negatives on the test dataset. The PCK-confidence curve (Fig. 8a) proved useful in determining the optimal confidence value, which was found to be 0.25, resulting in the highest level of PCK.

The results revealed that the PCK values for the nose and ears were lower than the other key body points. The PCK@0.2 was determined to be 94.17%. It is worth noting that a decrease in PCK@0.2 was observed in cases where the horse was in a lying down position or in environments with very much light (Fig. 9).

3.2.2. Behaviour classification

In this study, three experiments were carried out to analyse the results of classification of horse behaviours. The first experiment aimed to examine whether the addition of multi-output to the model could enhance the recognition of moving behaviours when the horse is solely moving, as opposed to moving during foraging or moving during unstable (Table 6). F-Score values for all three models LSTM, BiLSTM and Transformers indicated superior outcomes when multi-output was incorporated into the model.

To assess the performance of three tested models in the context of short windows, long windows, and multi-input, a second experiment was conducted. The primary objective of this experiment was to compare the results obtained when a single behaviour was present in the same window to those obtained in a real scenario where a sample may contain more than one behaviour in the same window. Table 7 provides a comparison of the various models employing a long window, short window, and multi-input with both windows. All the models used were multi-output. The outcomes revealed that models with multi-input offered superior results when compared to those with a single input since the model could learn from both types of inputs, better discerning the correct behaviour at a given time. Furthermore, it was observed that unstable behaviour yielded lower F-Scores compared to other behaviours. In terms of model performance, the Transformer model had slightly better than the LSTM-based models. However, a reduction in balanced accuracy was observed when comparing the balanced accuracy of models trained on samples containing only one behaviour to those containing one or more behaviours in a window of time, which is a realistic scenario.

To investigate the ability of the models to generalise to new data, a third and final experiment was conducted using the leave-one-out cross-validation methodology. Multiple models were trained on distinct horses from the dataset, and each model was tested on a horse that was not utilised for training. The outcomes of this experiment are detailed in Table 8, which compares the performance of various models using a short window, long window, and multi-input. All models utilised multi-output, and a comparison was made using different types of samples, including those with a single behaviour in a window of time, as well as those featuring one or more behaviours in a window of time, simulating a real-world scenario. It is important to note that F-Score values were not available for samples featuring the long moving behaviour due to the limited number of samples featuring this behaviour. Only one horse expressed this behaviour in our dataset. The outcomes of this experiment indicated a notable decline in F-Score values when compared to those presented in Table 8. Nevertheless, the multi-input Transformer model once again showed the best performance among the tested models.

To offer a more comprehensive analysis, Table 9 presents the balanced accuracy values obtained for each horse using the best performing models, i.e., LSTM, BiLSTM, and Transformers, with both multi-input and multi-output configurations. It is notable that horse number

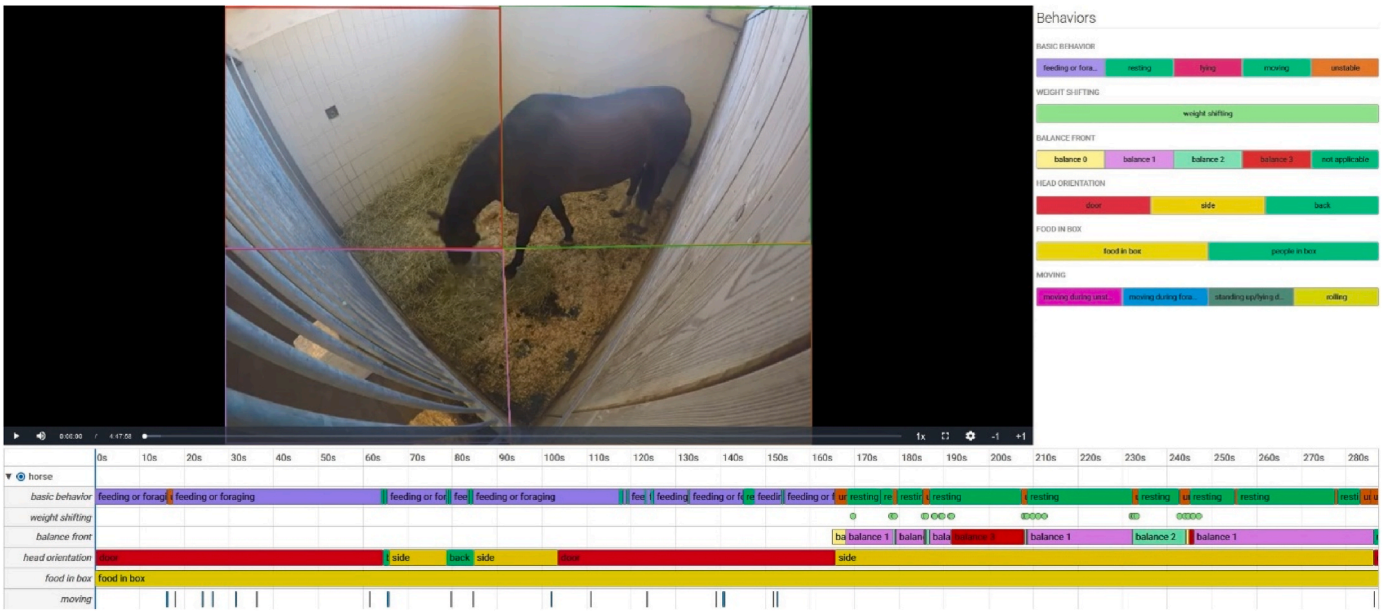

Fig. 7. The Loopy tool was used to label the horse's behaviour. The bottom part of the figure presents results of labelling over time.

Table 5
Behaviour time budget.

Horse	Duration (HH:MM:SS)	No. Horses
Lying	02:00:03	6
Resting	07:45:28	10
Feeding/Foraging	09:03:05	9
Unstable	04:18:39	10
Moving	00:39:18	9
Total	23:46:33	10
Overlapped behaviours	00:42:36	10

nine presented lower balanced accuracy values when using LSTM models, whereas horse number one demonstrated higher balanced accuracy values. However, the Transformer-based model showed greater stability compared to the LSTM models, as evidenced by the lower variance in balanced accuracy values between horses.

The performance of behaviour classification was compared using the confusion matrix. Fig. 10 illustrates the confusion matrix for horses numbered 1 and 9, as well as the overall matrix for all horses. These horses represented the best and worst performers, respectively, using the model with the highest total balanced accuracy, which was the Transformer model with multi-input and multi-output. The results revealed a high confusion between unstable and resting behaviours, with some confusion between feeding and both resting and unstable behaviours.

3.2.3. *Time budgets*

To calculate the time budget, the sum of predictions for each behaviour at each time instance was computed. Table 10 presents a comparison of the error between the true time budgets and predicted time budgets using the three proposed models. On a horse-by-horse basis, the LSTM models with multi-input and multi-output configurations generally exhibited lower error. However, akin to the behaviour classification task, the multi-input and multi-output Transformer models demonstrated more stable results and achieved better overall performance.

To enhance the visual representation of the time budgets, a bar plot was used. Fig. 11 presents the bar plots for horses' number 5 and 2. These results were obtained using the multi-input and multi-output Transformer model, which achieved the highest accuracy in comparison to LSTM's and BiLSTM's. Horse number 5 and 2 represent the best and worst performers, respectively.

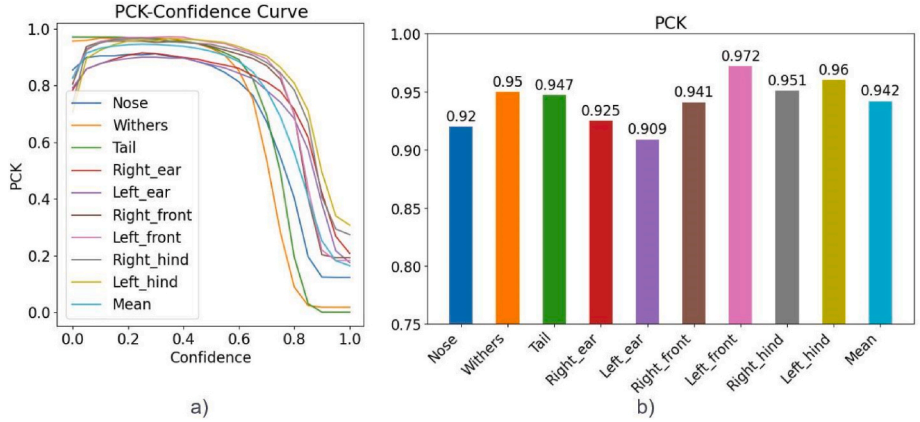

Fig. 8. Metrics used to select the best model for key body point detection a) PCK-Confidence curve. b) PCK Torso diameter.

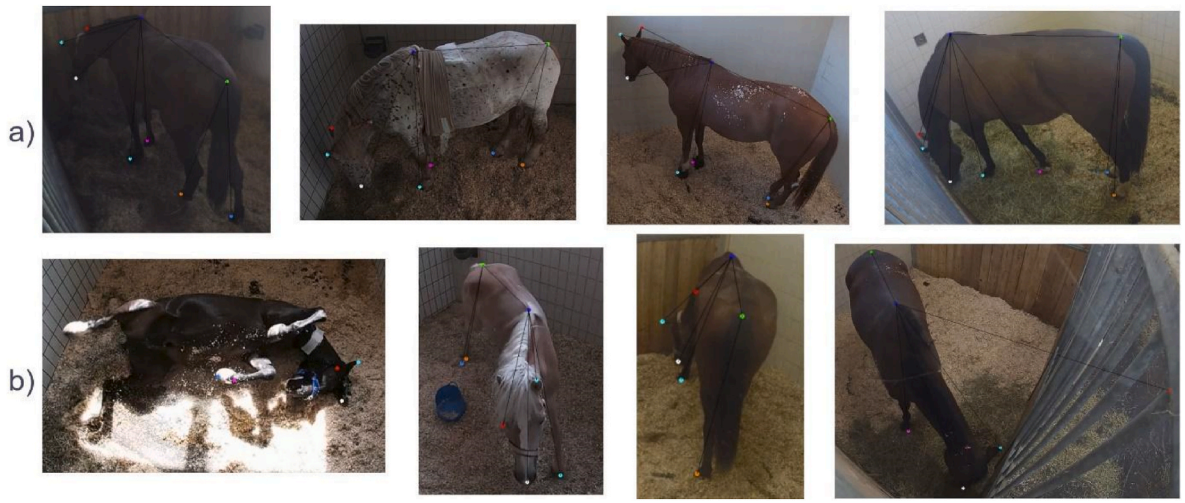

Fig. 9. Example of keypoint detection results. Model performance a) Good. b) Bad.

Table 6
F-Score of moving behaviour classification using LSTM, BiLSTM and Transformers models with single-output of 20 frames and multi-output. Each sample contains one behaviour. Comparison between same domain and leave-one-out.

Model	Methodology	Same domain	Leave one out CV
	Output	F-Score	Score
LSTM	Single output	0.911	0.676
	Multi-output	**0.942**	**0.715**
BiLSTM	Single output	0.929	0.698
	Multi-output	**0.965**	**0.724**
Transformers	Single output	0.923	0.657
	Multi-output	**0.931**	**0.685**

3.2.4. Computation complexity and performance

A comparative analysis of computation complexity and performance was undertaken to assess the impact of multi-output and multi-input configurations (Table 11). The results indicated that these configurations introduced additional hyper-parameters, parameters and increased computational complexity. Specifically, the multi-input approach led to a notable increase in the number of parameters due to the use of two windows, resulting in more parameters. Furthermore, when the BiLSTM model was implemented, the number of parameters increased more than twice compared with LSTM model, owing to the presence of two LSTM layers.

As predicted, the results revealed a positive correlation between the number of parameters and the training time, the higher the number of parameters, the longer was the training and inference time. Nevertheless, this correlation exhibited a relatively modest increase when the model was used for the inference process. The LSTM and Transformer models demonstrated comparable training times and yielded superior inference times compared to the BiLSTM model, owing to their architectural design.

4. Discussion

Highly accurate classification of horse behaviours in this study is directly dependent on accurate detection of keypoints. A 94.17% of PCK@0.2 was achieved in the detection of key body points in horses. When comparing the results of our study with the results of the other

Table 7
Balanced accuracy balanced classification comparison using LSTM, BiLSTM and Transformers models and their F-Score results for each model and behaviours using different windows and multi-input. Comparison of each sample containing only one behaviour with a real scenario when a sample can have one or more than one behaviour. Test data had samples chosen randomly and removed all the samples with more than one frame overlapped.

Behaviour							
Same Behaviour			Feeding	Lying	Moving	Resting	Unstable
Model	Input Size	B. Accuracy	F-Score	F-Score	F-Score	F-Score	F-Score
LSTM	20 frames	0.885	0.975	0.992	0.967	0.906	0.646
	60 frames	0.875	0.980	**1.000**	**1.000**	0.957	0.478
	20-60 frames	0.927	0.987	0.993	**1.000**	**0.963**	0.683
BiLSTM	20 frames	0.887	0.980	0.985	0.967	0.915	0.666
	60 frames	0.876	**0.988**	**1.000**	**1.000**	0.950	0.538
	20-60 frames	0.927	0.982	**1.000**	**1.000**	0.946	**0.697**
Transformers	20 frames	0.874	0.975	0.993	0.903	0.908	0.649
	60 frames	0.902	0.983	0.973	**1.000**	0.958	0.645
	20-60 frames	0.911	0.984	0.993	0.933	0.950	0.634
Real Scenario							
LSTM	20 frames	0.819	0.930	**1.000**	0.750	0.825	0.499
	60 frames	0.784	0.918	**1.000**	0.625	0.814	0.489
	20-60 frames	0.833	0.947	**1.000**	0.750	0.825	0.558
BiLSTM	20 frames	0.836	0.926	**1.000**	0.750	0.842	0.573
	60 frames	0.774	0.927	0.987	0.444	0.829	0.592
	20-60 frames	0.846	0.940	0.988	0.666	0.837	0.629
Transformers	20 frames	0.890	0.950	**1.000**	0.824	0.846	0.662
	60 frames	0.840	0.919	0.953	0.499	0.824	0.653
	20-60 frames	0.893	**0.953**	0.992	**0.875**	**0.842**	**0.689**

Table 8
Balanced accuracy classification comparison using LSTM, BiLSTM and Transformers models and their F-Score results for each model and behaviours using different windows and multi-input. Comparison when each sample contained only one behaviour and in a real scenario when a sample can have one or more than one behaviour. A leave-one-out cross-validation methodology was used to create the test dataset.

Behaviour							
Same Behaviour			Feeding	Lying	Moving	Resting	Unstable
Model	Input Size	B. Accuracy	F-Score	F-Score	F-Score	F-Score	F-Score
LSTM	20 frames	0.655	0.724	0.369	0.732	0.702	0.228
	60 frames	0.578	0.791	0.624	NA	0.715	0.221
	20-60 frames	0.656	0.842	**0.893**	NA	0.719	0.232
BiLSTM	20 frames	0.573	0.793	0.448	0.727	0.714	0.325
	60 frames	0.620	0.813	0.667	NA	0.679	0.279
	20-60 frames	0.659	0.846	0.802	NA	0.737	0.245
Transformers	20 frames	0.682	0.784	0.726	**0.797**	0.698	0.405
	60 frames	0.673	0.832	0.687	NA	**0.776**	**0.427**
	20-60 frames	0.726	**0.850**	0.862	NA	0.748	0.421
Real Scenario							
LSTM	20 frames	0.550	0.726	0.435	0.715	0.655	0.358
	60 frames	0.451	0.718	0.383	0.266	0.667	0.364
	20-60 frames	0.611	0.764	0.703	0.683	0.649	0.381
BiLSTM	20 frames	0.591	0.736	0.644	**0.724**	0.640	0.353
	60 frames	0.465	0.689	0.571	0.248	0.625	0.352
	20-60 frames	0.607	0.758	0.684	0.702	0.650	0.375
Transformers	20 frames	0.593	0.786	0.386	0.685	0.680	0.502
	60 frames	0.466	0.730	0.347	0.287	0.638	0.423
	20-60 frames	0.676	**0.822**	**0.749**	0.701	**0.709**	**0.507**

Table 9
Behaviour classification balanced accuracy results for each horse with LSTM, BiLSTM and Transformers models.

Horse	LSTM	BiLSTM	Transformers
#1	0.591	0.621	0.567
#2	0.676	0.624	**0.749**
#3	0.526	0.612	**0.643**
#4	0.452	0.483	**0.582**
#5	0.620	0.643	**0.691**
#6	0.630	0.631	**0.640**
#7	0.576	0.579	**0.633**
#8	0.668	0.613	**0.690**
#9	0.481	0.489	**0.527**
#10	0.561	0.513	**0.624**

studies, we found variable performance of key body point detection in the other studies. The detection of keypoints with the MPII dataset for pose estimation in humans, using the HRNet model by Sun et al. (2019), resulted in performance of 92.3% PCK, which was slightly below the results obtained in our study. However, it is important to consider that MPII dataset consists of images with one and more than one individual, leading to a more complex dataset. Additional studies have focused on estimating keypoints in dogs. Zhang et al. (2021) utilised lateral X-ray images and keypoint detection to aid veterinarians in diagnosing canine cardiac enlargement. In their study, HRNet was applied to detect 16 keypoints i.e., 12 located on vertebra and 4 on heart, achieving a PCK of 99.36% on their dataset. The results demonstrated a higher PCK due to the X-ray images being in a constant lateral position, which allows the dataset to be considered smaller and simpler. In another investigation, Zhu, Salgırlı, Can, Durmuş Atılgan, and Salah (2022) developed a method for automatically recognizing pain-related behaviours in dogs by utilizing spatial and temporal features. The study detected a total of 17 keypoints from various parts of the dog's body. The results of using

Table 10
Time budget error in tant per cent results for each horse with LSTM, BiLSTM and Transformers models.

Horse	LSTM	BiLSTM	Transformers
#1	**3.57**	3.82	12.01
#2	26.01	24.73	**24.45**
#3	19.02	**15.87**	17.6
#4	20.58	14.91	**12.28**
#5	4.02	3.795	**3.33**
#6	9.47	7.835	**4.93**
#7	19.30	22.23	**12.61**
#8	6.875	6.8	**5.94**
#9	27.53	28.17	**18.99**
#10	15.70	20.31	**15.46**

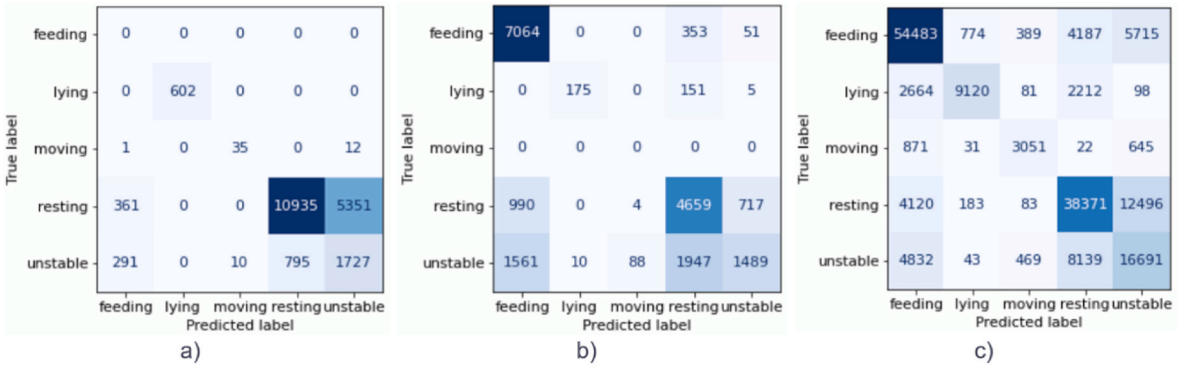

Fig. 10. Confusion matrix results for a) horse #1, b) horse #9 c) tall horses. Multi-input and multi-output Transformer model.

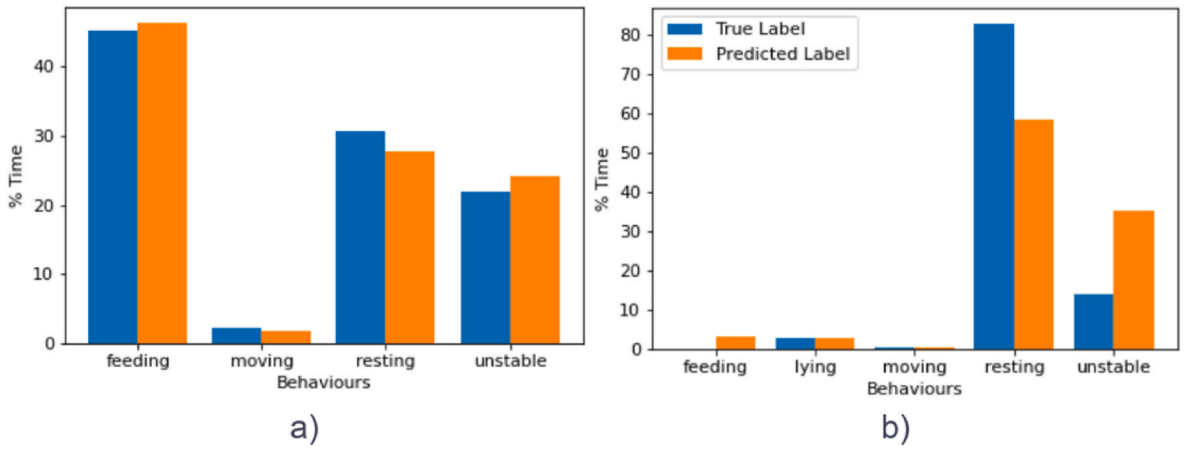

Fig. 11. Time budgets with a) horse #5 b) horse #2. Multi-input and multi-output Transformer model.

Table 11
Computational complexity, GPU memory consumption, and training and inference times of the models' strategies using a batch of 64 samples. Hardware: NVIDIA GeForce RTX 3090 24 GB GPU and Intel(R) Xeon(R) Gold 6226 CPU @ 2.70 GHz.

Model	Input size	Model Parameters	Training GPU time	Inference GPU time
LSTM	20 frames	113 K	8 ms/batch	3 ms/batch
	60 frames	389 K	10 ms/batch	4 ms/batch
	20-60 frames	502 K	19 ms/batch	5 ms/batch
BiLSTM	20 frames	352 K	8 ms/batch	3 ms/batch
	60 frames	770 K	16 ms/batch	5 ms/batch
	20-60 frames	992 K	29 ms/batch	8 ms/batch
Transformers	20 frames	153 K	8 ms/batch	3 ms/batch
	60 frames	315 K	9 ms/batch	3 ms/batch
	20-60 frames	469 K	18 ms/batch	5 ms/batch

the HRNet model revealed a PCK of 80.9% on their dataset and 81.5% on the TigDog dataset (Del Pero, Ricco, Sukthankar, & Ferrari, 2014). The resulting PCK was significantly lower compared to the results in our study. However, it should be noted that the dataset consists of images with different dog breeds in various environments, making it a much more complex dataset than ours. Therefore, it can be concluded that the results obtained in our study are consistent with those found in other studies, as the PCK values vary depending on the complexity of the dataset and the different thresholds applied for PCK calculation, such as 20% of the torso or 50% of the head, leading to varying levels of performance.

A promising next step to improve the performance of pose detection could be to include estimates of locations of all key body points when training the model, even if some key body points are occluded or not visible. Although the key points may not be visible, they can be annotated using knowledge about horse's anatomy. This would ensure that all keypoints are consistently detected even if not visible or occluded, which can potentially enhance the results of pose estimation in comparison to models that are trained without these key points. In our current study not visible and occluded key body points were not used for training of the HRNet model. Furthermore, it is imperative to obtain a larger-scale image dataset that is more representative of the real scenario with more horses with variables, sizes, shapes, and colours (Shorten & Khoshgoftaar, 2019). This dataset could be real images, synthetic images, or a combination of both. Synthetic images offer the benefit of simulating scenarios that are challenging to capture or recreate (Tremblay et al., 2018), thereby providing greater flexibility in the training and testing of models. This approach has been successfully applied by Mu, Qiu, Hager, and Yuille (2019) in the domain of animal detection and pose estimation using only synthetic images, where the model trained can detect keypoints reliably on real images with horses and tigers in the TigDog dataset achieving better generalization performance than models trained on real images across different domains in the Visual Domain Adaptation Challenge dataset (VisDA 2019).

In our study, we consistently encountered a single horse in each frame, as the analysis focused on stables with individual horses. Consequently, the use of bottom-up and top-down methods for keypoint detection (Jin et al., 2017), which are typically used when multiple subjects are present in the frame, becomes unnecessary. Instead, the entire frame serves as the input image for the HRNet model. However, it should be noted that some keypoints were detected outside the designated area corresponding to the horse, as illustrated in Fig. 9 bottom right frame. A potential avenue for future research could involve exploring a top-down approach, which entails incorporating a horse detector as a preliminary step, followed by estimating the precise locations of key body points within the identified horse region. By adopting a top-down structure, it is possible to address the issue previously mentioned, wherein keypoints were detected outside the horse's area. Furthermore, attention-based models such as Visual Transformers (Y. Liu et al., 2021), and ViTPose (Xu, Zhang, Zhang, & Tao, 2022) are currently state-of-the-art in the fields of computer vision and pose estimation achieving better results on the COCO keypoint detection dataset. It may be possible to adapt and apply these models to estimate horse pose and compare their performance against convolutional neural network models in the future research.

This study proposed a novel multi-input, multi-output classification methodology designed to tackle the challenges associated with the detection of behaviours. These challenges included the masking effect caused by behaviours with excessively long durations and the potential misclassification resulting from the simultaneous occurrence of multiple behaviours. In the first behaviour classification experiment with the aim to mitigate the risk of misclassification of overlapped behaviours, a multi-output approach was adopted comparing LSTM, BiLSTM and Transformer models. This approach facilitated a binary classification task focused solely on movements and a multiclass classification task to detect time budget behaviours. The implementation of this approach resulted in an improvement of around 3% in the F-score for moving behaviour (Table 6) compared to the single-output model i.e., false positives and false negatives were reduced.

A second experiment was conducted to evaluate LSTM, BiLSTM and Transformer models with different input window sizes, a short window, a long window, and multi-input with both window sizes. The input windows were configured with a short duration of 20 frames or 10 s and a longer duration of 60 frames or 30 s. The results demonstrated that

longer input windows produced higher balanced accuracy when the window contained only one behaviour in Table 7 same behaviour. This suggests that the longer the input window, the easier it is to predict the corresponding label. However, the results were reversed when the input window contained multiple behaviours or transitioned between different behaviours. This was because, as the window size decreased, the number of samples with transitions were reduced, and it became easier to obtain samples with only one behaviour, thus making the prediction more feasible. When the results of the models with one single input and multi-input were compared in Tables 7 and 8, the models with two simultaneous window sizes demonstrated improvement. The multi-input approach was found to produce better results as it could interpret both windows simultaneously. Although only two windows were used in this study, future investigations could explore different windows lengths and multiple windows sizes if necessary.

Our analysis indicates that classification of unstable behaviour resulted in lower values of F-Score, around 20%, when it was compared to the results of classification of the other behaviours considered in our experiment. This outcome can be attributed to the fact that such behaviour can cause a subtle swinging of the horse's weight. Additionally, the periods of resting between bouts of swinging can fluctuate, leading to misclassification of this behaviour similarly to resting in many instances. For future research, it is imperative to establish a more precise definition of this behaviour, i.e., annotate when the horse swings and define the maximum time that the horse doesn't move and annotate it as a resting period.

For future research focused on behaviour detection and classification in continuous data scenarios, it is recommended to apply multi-input configurations, as they have shown improvements in terms of the F-Score values for each behaviour (Tables 6 and 7). As for the use of multi-output configuration, it proved to be effective in specific cases within our dataset, where multiple behaviours occurred simultaneously. Thus, application of multi-output configuration might improve performance of behaviour detection in similar scenarios. However, it is crucial to consider that both solutions introduced in this study result in an increase of two to five times in the number of parameters in comparison to single-input models (Table 11), albeit without doubling the inference time. The input data used for these models were key points and their respective vectors, which significantly reduced the number of parameters, learning time, and inference time in comparison to using complete frames. Nevertheless, if dealing with larger datasets, such as full frames, further studies would be necessary to assess their feasibility and optimise their implementation for behaviour detection purposes.

Comparison of our results of time budget classification with the other studies was not possible as we couldn't find relevant studies for such comparison. Although (Alameer, Kyriazakis, & Bacardit, 2020; Alameer, Kyriazakis, Dalton, Miller, & Bacardit, 2020) tried to recognise if a pig was standing, sitting, lying sternal or laterally to determine animal welfare and in a second study to determine if the pig was feeding or not but, a time window of a single frame was used in both studies to recognise them and cannot be directly compared to our study, where the recognition of behaviour needs multiple frames. Another work presented by Cowton, Kyriazakis, and Bacardit (2019) made an estimation of the distance travelled, average speed, and idle time in pigs but for this purpose tracking methods were used, and it was analysed how many pixels changed in consecutive frames of video recordings. Finally, Shavit and Klein (2021) presented a similar work in humans for action recognition to classify behaviours such as walking, running, sitting, standing, jogging, biking, upstairs and downstairs with Transformers models and additionally using the accelerometer and gyroscope sensors installed in smartphones. Although the work may be similar, the dataset is different, whereas their dataset consists of accelerometer and gyroscope data, ours comprises videos and results cannot be compared with our work.

After evaluating the performance of three models applied in our study, we conclude that the Transformer-based methods were more resistant to domain changes than the LSTM-based models. Transformers got more stable values in accuracy and F-Scores and both metrics did not show larger drops of their values. Thus, we also conclude that Transformer-based methods were more robust than LSTM-based methods even though application of LSTM's gave higher accuracy values than Transformers in some horses. Nevertheless, it would be misleading to conclude that LSTM's superior robustness is due to the intrinsic properties of Transformers i.e., the self-attention mechanism. We can conclude that, in our dataset, Transformer-based methods learned better representations and generalised better across domains than other classification methods. Thus, the Transformers model was selected as the baseline for subsequent experiments.

5. Conclusions

This study compared the performance of Transformers, LSTMs, and Bidirectional LSTMs for the classification of the behavioural time budget in horses based on video data. The experiments conducted to analyse the classification results demonstrated that models with multi-input and multi-output configurations outperformed models with a single input and single output. The Transformer model exhibited slightly better performance compared to the LSTM-based models, with unstable behaviour presenting a challenge in classification accuracy. The leave-one-out cross-validation experiment showed a decline in F-Score values, indicating the need for further research to enhance the generalization ability of the models.

Overall, this research contributes to the field of automated behaviour recognition and classification in horses using computer vision and time series analysis. The proposed methodology has the potential to enhance the detection and classification of horse behaviours, which can be valuable for welfare monitoring, pain detection, and early detection of health problems in horses. Further advancements in dataset diversity, pose estimation techniques, and model architectures can lead to more accurate and generalizable results in future studies.

CRediT authorship contribution statement

Albert Martin-Cirera: Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Magdelena Nowak:** Supervision, Validation. **Tomas Norton:** Supervision, Validation, Writing – review & editing. **Ulrike Auer:** Supervision, Validation. **Maciej Oczak:** Investigation, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Rasool, G., & Ramachandran, R. P. (2022). Transformers in time-series analysis: A tutorial. *Circuits, Systems, and Signal Processing, 42*(12), 7433–7466. https://doi.org/10.1007/s00034-023-02454-8

Alameer, A., Kyriazakis, I., & Bacardit, J. (2020). Automated recognition of postures and drinking behaviour for the detection of compromised health in pigs. *Scientific Reports, 10*(1), 1–15. https://doi.org/10.1038/s41598-020-70688-6, 10(1).

Alameer, A., Kyriazakis, I., Dalton, H. A., Miller, A. L., & Bacardit, J. (2020). Automatic recognition of feeding and foraging behaviour in pigs using deep learning. *Biosystems Engineering, 197*, 91–104. https://doi.org/10.1016/J.BIOSYSTEMSENG.2020.06.013

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2D human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 3686–3693). https://doi.org/10.1109/CVPR.2014.471

Ashley, F. H., Waterman-Pearson, A. E., & Whay, H. R. (2005). Behavioural assessment of pain in horses and donkeys: Application to clinical practice and future studies. *Equine Veterinary Journal, 37*(6), 565–575. https://doi.org/10.2746/042516405775314826

Auer, U., Kelemen, Z., Engl, V., & Jenner, F. (2021). Activity time budgets—a potential tool to monitor equine welfare? *Animals, 11*, 850. https://doi.org/10.3390/ANI11030850, 11(3), 850.

Berger, A., Scheibe, K. M., Michaelis, S., & Streich, W. J. (2003). Evaluation of living conditions of free-ranging animals by automated chronobiological analysis of behavior. *Behavior Research Methods, Instruments, & Computers, 35*(3), 458–466. https://doi.org/10.3758/BF03195524/METRICS

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics, 1–3*, 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Boy, V., & Duncan, P. (1979). Time-budgets of camargue horses I. Developmental Changes in the time-budgets of foals. *Behaviour, 71*(3–4), 187–201. https://doi.org/10.1163/156853979X00160

Brooks, J. (2019). COCO annotator. https://github.com/jsbroks/coco-annotator.

Cangar, Ö., Leroy, T., Guarino, M., Vranken, E., Fallon, R., Lenehan, J., et al. (2008). Automatic real-time monitoring of locomotion and posture behaviour of pregnant cows prior to calving using online image analysis. *Computers and Electronics in Agriculture, 64*(1), 53–60. https://doi.org/10.1016/J.COMPAG.2008.05.014

Chen, C., Zhu, W., & Norton, T. (2021). Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. *Computers and Electronics in Agriculture, 187*, Article 106255. https://doi.org/10.1016/J.COMPAG.2021.106255

Chen, C., Zhu, W., Oczak, M., Maschat, K., Baumgartner, J., Larsen, M. L. V., et al. (2020). A computer vision approach for recognition of the engagement of pigs with different enrichment objects. *Computers and Electronics in Agriculture, 175*. https://doi.org/10.1016/J.COMPAG.2020.105580

Chen, C., Zhu, W., Steibel, J., Siegford, J., Wurtz, K., Han, J., et al. (2020). Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory. *Computers and Electronics in Agriculture, 169*, Article 105166. https://doi.org/10.1016/J.COMPAG.2019.105166

Clothier, J., Small, A., Hinch, G., Barwick, J., & Brown, W. Y. (2019). Using movement sensors to assess lying time in horses with and without angular limb deformities. *Journal of Equine Veterinary Science, 75*, 55–59. https://doi.org/10.1016/J.JEVS.2019.01.011

Contributors, M. (2020). Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose.

Cowton, J., Kyriazakis, I., & Bacardit, J. (2019). Automated individual pig localisation, tracking and behaviour metric extraction using deep learning. *IEEE Access, 7*, 108049–108060. https://doi.org/10.1109/ACCESS.2019.2933060

Del Pero, L., Ricco, S., Sukthankar, R., & Ferrari, V. (2014). Articulated motion discovery using pairs of trajectories. https://arxiv.org/abs/1411.7883v3.

Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). *Revisiting skeleton-based action recognition*. https://github.com/kennymckormick/pyskl.

Duncan, P. (1980). Time-budgets of camargue horses II. Time-budgets of adult horses and weaned sub-adults. *Behaviour, 72*(1–2), 26–48. https://doi.org/10.1163/156853980X00023

Feist, J. D., & McCullough, D. R. (1976). Behavior patterns and communication in feral horses. *Zeitschrift für Tierpsychologie, 41*(4), 337–371. https://doi.org/10.1111/J.1439-0310.1976.TB00947.X

Flannigan, G., & Stookey, J. M. (2002). Day-time time budgets of pregnant mares housed in tie stalls: A comparison of draft versus light mares. *Applied Animal Behaviour Science, 78*(2–4), 125–143. https://doi.org/10.1016/S0168-1591(02)00085-0

Frost, A. R., Schofield, C. P., Beaulah, S. A., Mottram, T. T., Lines, J. A., & Wathes, C. M. (1997). A review of livestock monitoring and the need for integrated systems. *Computers and Electronics in Agriculture, 17*(2), 139–159. https://doi.org/10.1016/S0168-1699(96)01301-4

Goodwin, D. (1999). The importance of ethology in understanding the behaviour of the horse. *Equine Veterinary Journal, 31*(S28), 15–19. https://doi.org/10.1111/J.2042-3306.1999.TB05150.X

Goodwin, D., Davidson, H. P. B., & Harris, P. (2002). Foraging enrichment for stabled horses: Effects on behaviour and selection. *Equine Veterinary Journal, 34*(7), 686–691. https://doi.org/10.2746/042516402776250450

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: An overview. https://arxiv.org/abs/2008.05756v1.

Hausberger, M., Fureix, C., & Lesimple, C. (2016). Detecting horses' sickness: In search of visible signs. *Applied Animal Behaviour Science, 175*, 41–49. https://doi.org/10.1016/J.APPLANIM.2015.09.005

Hausberger, M., Lerch, N., Guilbaud, E., Stomp, M., Grandgeorge, M., Henry, S., et al. (2020). On-farm welfare assessment of horses: The risks of putting the cart before the horse. *Animals, 10*, 371. https://doi.org/10.3390/ANI10030371, 10(3), 371.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Horn, M., Moor, M., Bock, C., Rieck, B., & Borgwardt, K. (2020). *Set functions for time series*. PMLR. https://proceedings.mlr.press/v119/horn20a.html.

Ijaz, M., Diaz, R., & Chen, C. (2022). Multimodal transformer for nursing activity recognition. In *IEEE computer society conference on computer vision and pattern recognition workshops, 2022-june* (pp. 2064–2073). https://doi.org/10.1109/CVPRW56347.2022.00224

Jin, S., Ma, X., Han, Z., Wu, Y., Yang, W., Liu, W., et al. (2017). *Towards Multi-Person Pose Tracking: Bottom-up and Top-down Methods*.

Lesimple, C. (2020). Indicators of horse welfare: State-of-the-Art. *Animals, 10*, 294. https://doi.org/10.3390/ANI10020294, 10(2), 294.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 8693 LNCS(PART 5)* (pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48/COVER

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A survey of Transformers. *AI Open, 3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

Liu, D., Oczak, M., Maschat, K., Baumgartner, J., Pletzer, B., He, D., et al. (2020). A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs. *Biosystems Engineering, 195*, 27–41. https://doi.org/10.1016/J.BIOSYSTEMSENG.2020.04.007

Liu, Y., Zhang, Y., Wang, Y., Hou, F., Yuan, J., Tian, J., et al. (2021). A survey of visual Transformers. *IEEE Transactions on Neural Networks and Learning Systems*. https://doi.org/10.1109/TNNLS.2022.3227717

Martin, B. R., Prescott, W. R., & Zhu, M. (1992). Quantitation of rodent catalepsy by a computer-imaging technique. *Pharmacology Biochemistry and Behavior, 43*(2), 381–386. https://doi.org/10.1016/0091-3057(92)90166-D

Mayes, E., & Duncan, P. (1986). Temporal patterns of feeding behaviour in free-ranging horses. *Behaviour, 96*(1–2), 105–129. https://doi.org/10.1163/156853986X00243

Mu, J., Qiu, W., Hager, G., & Yuille, A. (2019). Learning from synthetic animals. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 12383–12392). https://doi.org/10.1109/CVPR42600.2020.01240

Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. *Sensing and Bio-Sensing Research, 29*, Article 100367. https://doi.org/10.1016/J.SBSR.2020.100367

Nilsson, M., Herlin, A. H., Ardö, H., Guzhva, O., Aström, K., & Bergsten, C. (2015). Development of automatic surveillance of animal behaviour and welfare using image analysis and machine learned segmentation technique. *Animal, 9*(11), 1859–1865. https://doi.org/10.1017/S1751731115001342

Oczak, M., Viazzi, S., Ismayilova, G., Sonoda, L. T., Roulston, N., Fels, M., et al. (2014). Classification of aggressive behaviour in pigs by activity index and multilayer feed forward neural network. *Biosystems Engineering, 119*, 89–97. https://doi.org/10.1016/J.BIOSYSTEMSENG.2014.01.005

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S. H., Murthy, M., et al. (2019). Fast animal pose estimation using deep neural networks. *Nature Methods, 16*(1), 117–125. https://doi.org/10.1038/S41592-018-0234-5

Price, J., Catriona, S., Welsh, E. M., & Waran, N. K. (2003). Preliminary evaluation of a behaviour-based system for assessment of post-operative pain in horses following arthroscopic surgery. *Veterinary Anaesthesia and Analgesia, 30*(3), 124–137. https://doi.org/10.1046/J.1467-2995.2003.00139.X

Sarrafchi, A., & Blokhuis, H. J. (2013). Equine stereotypic behaviors: Causation, occurrence, and prevention. https://doi.org/10.1016/j.jveb.2013.04.068.

Shavit, Y., & Klein, I. (2021). Boosting inertial-based human activity recognition with Transformers. *IEEE Access, 9*, 53540–53547. https://doi.org/10.1109/ACCESS.2021.3070646

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1–48. https://doi.org/10.1186/S40537-019-0197-0/FIGURES/33

Si, C., Chen, W., Wang, W., Wang, L., & Tan, T. (2019). An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition, 2019-june* (pp. 1227–1236). https://doi.org/10.1109/CVPR.2019.00132

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. In *Proceedings - 2019 IEEE international conference on big data, big data 2019* (pp. 3285–3292). https://doi.org/10.1109/BIGDATA47090.2019.9005997

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and roc: A family of discriminant measures for performance evaluation. *AAAI Workshop - Technical Report, WS-06-06*, 24–29. https://doi.org/10.1007/11941439_114/COVER

Song, L., Yu, G., Yuan, J., & Liu, Z. (2021). Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation, 76*, Article 103055. https://doi.org/10.1016/J.JVCIR.2021.103055

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 5686–5696). https://doi.org/10.1109/CVPR.2019.00584, 2019-June.

Sweeting, M. P., Houpt, C. E., & Houpt, K. A. (1985). Social facilitation of feeding and time budgets in stabled ponies. *Journal of Animal Science, 60*(2), 369–374. https://doi.org/10.2527/JAS1985.602369X

Thorne, J. B., Goodwin, D., Kennedy, M. J., Davidson, H. P. B., & Harris, P. (2005). Foraging enrichment for individually housed horses: Practicality and effects on behaviour. *Applied Animal Behaviour Science, 94*(1–2), 149–164. https://doi.org/10.1016/J.APPLANIM.2005.02.002

Torcivia, C., & McDonnell, S. (2020). In-person caretaker visits disrupt ongoing discomfort behavior in hospitalized equine orthopedic surgical patients. *Animals: An Open Access Journal from MDPI, 10*(2). https://doi.org/10.3390/ANI10020210

Tremblay, J., Prakash, A., Acuna, D., Brophy, M., Jampani, V., Anil, C., et al. (2018). Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *IEEE computer society conference on computer vision and pattern recognition workshops, 2018-june* (pp. 1082–1090). https://doi.org/10.1109/CVPRW.2018.00143

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review, 53*(8), 5929–5955. https://doi.org/10.1007/S10462-020-09838-1/TABLES/1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5999–6009, 2017-December https://arxiv.org/abs/1706.03762v7.

Wang, M., oczak, M., Larsen, M., Bayer, F., Maschat, K., Baumgartner, J., et al. (2021). A PCA-based frame selection method for applying CNN and LSTM to classify postural behaviour in sows. *Computers and Electronics in Agriculture, 189*. https://doi.org/10.1016/J.COMPAG.2021.106351

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., et al. (2022). Transformers in time series: A survey. In *IJCAI international joint conference on artificial intelligence, 2023-august* (pp. 6778–6786). https://doi.org/10.24963/ijcai.2023/759

Xu, Y., Zhang, J., Zhang, Q., & Tao, D. (2022). ViTPose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems, 35*. https://arxiv.org/abs/2204.12484v3.

Yang, Y., & Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(12), 2878–2890. https://doi.org/10.1109/TPAMI.2012.261

Yarnell, K., Hall, C., Royle, C., & Walker, S. L. (2015). Domesticated horses differ in their behavioural and physiological responses to isolated and group housing. *Physiology & Behavior, 143*, 51–57. https://doi.org/10.1016/J.PHYSBEH.2015.02.040

Yin, X., Wu, D., Shang, Y., Jiang, B., & Song, H. (2020). Using an EfficientNet-LSTM for the recognition of single Cow's motion behaviours in a complicated environment. *Computers and Electronics in Agriculture, 177*. https://doi.org/10.1016/J.COMPAG.2020.105707

Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., & Tao, D. (2021). AP-10K: A benchmark for animal pose estimation in the wild. https://arxiv.org/abs/2108.12617v2.

Yu, B., Yin, H., & Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *IJCAI international joint conference on artificial intelligence, 2018-july* (pp. 3634–3640). https://doi.org/10.24963/ijcai.2018/505

Zhang, M., Zhang, K., Yu, D., Xie, Q., Liu, B., Chen, D., et al. (2021). Computerized assisted evaluation system for canine cardiomegaly via key points detection with deep learning. *Preventive Veterinary Medicine, 193*. https://doi.org/10.1016/J.PREVETMED.2021.105399

Zhu, H., Salgırlı, Y., Can, P., Durmuş Atılgan, D., & Salah, A. A. (2022). Video-based estimation of pain indicators in dogs. https://arxiv.org/abs/2209.13296v2.