Automatic equine activity detection by convolutional neural networks using accelerometer data

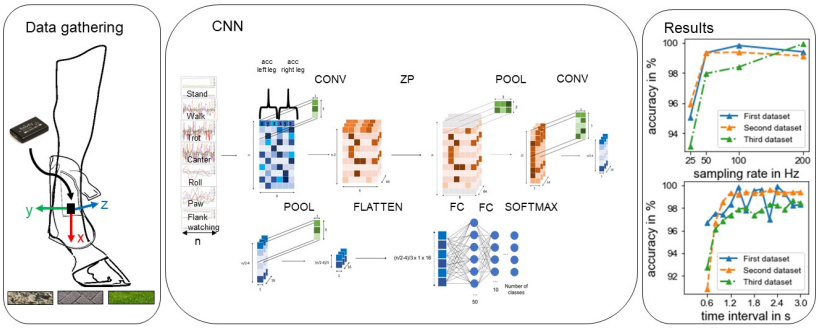
Anniek Eerdekens [a,*], Margot Deruyck [a], Jaron Fontaine [b], Luc Martens [a], Eli De Poorter [b], Wout Joseph [a]

[a] WAVES, Department of Information Technology, Ghent University - imec, Technologiepark-Zwijnaarde 126, B-9052 Ghent, Belgium

[b] IDLab, Department of Information Technology, Ghent University - imec, Technologiepark-Zwijnaarde 126, B-9052 Ghent, Belgium

[*] Corresponding author. E-mail address: anniek.eerdekens@ugent.be (Anniek Eerdekens)

Graphical Abstract



Abstract

In recent years, with a widespread of sensors embedded in all kind of mobile devices, human activity analysis is occurring more often in several domains like healthcare monitoring and fitness tracking. This trend did also enter the equestrian world because monitoring behaviours can yield important information about the health and welfare of horses. In this research, a deep learning-based approach for activity detection of equines is proposed to classify seven activities based on accelerometer data. We propose using Convolutional Neural Networks (CNN) by which features are extracted automatically by using strong computing capabilities. Furthermore, we investigate the impact of the sampling frequency, the time series length and the type of underground on which the data is gathered on the recognition accuracy and evaluate the model on three types of experimental datasets that are compiled of labelled accelerometer data gathered from six different subjects performing seven different activities. Afterwards, a horse-wise cross validation is carried out to investigate the impact of the subjects themselves on the model recognition accuracy. Finally, a slightly adjusted model is validated on different amounts of 50 Hz sensor data.

A 99% accuracy can be reached for detecting seven behaviours of a seen horse when the sampling rate is 25 Hz and the time interval is 2.1 s. Four behaviours of an unseen horse can be detected with the same accuracy when the sampling rate is 69 Hz and the time interval is 2.4 s. Moreover, the accuracy of the model for the three datasets decreased on average with about 4.75% when the sampling rate was decreased from

200 Hz to 25 Hz and with 5.27% when the time interval was decreased from 3 s to 0.6 s. In addition, the classification performance of the activity "walk" was not influenced by the type of underground the horse was performing this movement on and even the model could conclude from which underground the data was gathered for three out of four undergrounds with accuracies above 93% at time intervals higher than 1.2 s. This ensures the evaluation of activity patterns in real world circumstances. The performance and ability of the model to generalise is validated on 50 Hz data from different horse types, using ten-fold cross validation, reaching a mean classification accuracy of 97.84% and 96.10% when validated on a lame horse and pony, respectively. Moreover, in this work we show that using data from one sensors is at the cost of only 0.24% reduction in accuracy (99.42% vs 99.66%).

Keywords

Deep learning; Convolutional Neural Networks, Equines, Behaviour classification; Accelerometer

1. Introduction

Monitoring behaviour can yield important information about the health and welfare of horses (van Loon and Van Dierendonck, 2015). Direct observation of horse behavior is labour-intensive and is mainly based on intuition derived from previous experiences, which involves subjective decisions. To solve this kind of issues, different technologies are developed to detect various parameters such as activity, elevation, heart rate and so on from which conclusions can be drawn regarding the behaviour of the horse (Langrock et al., 2012), (Burla et al., 2014), (Bidder et al., 2014). In particular, wearable accelerometers have been tested for the determination of gaits by definition of distinct acceleration value ranges for stand, walk, trot and canter but not yet to detect other behaviours such as rolling, pawing and flank watching (Burla et al., 2014). In addition to accelerometers, researchers have suggested the use of various machine learning tools to classify accelerometer data more accurately. A disadvantage of the proposed methods is that feature extraction is

still necessary. A convolutional Neural Network (CNN) has the advantage of automatic features extraction by using strong computing capabilities. Deep learning-based classifiers can learn features and achieve better accuracy. For example, (Zhao et al., 2019) uses the deep CNN features for ground-based cloud image classification. The results show that the cloud classification accuracy of CNN improved significantly, demonstrating the superiority of CNN over hand-engineered features. Besides high accuracy and good generalisation, one main advantage of this way of working is that after a deep learning model is designed, it is trained in an end-to-end fashion, thus completely removing the need of manual feature engineering (Ignatov, 2018). In recent years, CNNs have shown excellent performance on classification problems when large-scale labelled datasets are available (Um et al., 2017). Studies demonstrated that deep learning models are able to learn and discriminate among human activities ranging from sitting, walking, climbing upstairs, walking downstairs and falling, among others but are to the authors' knowledge not yet applied for the detection of equine activities (Ravi et al., 2017).

In this work, as a novelty, an experimentally validated CNN is proposed to automatically detect seven distinct activities of equines by using data from two accelerometers for the first time to the author's knowledge. Further novelties include the analysis of sampling rate, time series length and investigation of the influence of the underground. Also experimental data for six horses and seven activities wearing two accelerometers has been gathered and annotated.

The rest of the paper is organised as follows. Section 2 deals with the methodology and the proposed deep learning model. Results of the experimentally validated model are presented in Section 3. Finally, conclusions are drawn in Section 4.

2. Materials and method

2.1. Animals and training arena

Measurements were conducted between November 2018 and April 2019 in a horse farm in Zutendaal, Belgium with six adult horses of different breeds. All details about the subjects can be found in Table 1. This

Subject number	Breed class	Height at withers (cm)	Gender	Age	State	Shoeing
1	Warmblood	172	Mare	7	Healthy	Barefoot
2	Warmblood	167	Gelding	11	Healthy	Barefoot
3	Warmblood	181	Mare	17	Lame	Barefoot
4	Warmblood	168	Mare	19	Healthy	Barefoot
5	Friesian	159	Mare	12	Healthy	Shoed
6	Pony	116	Gelding	15	Healthy	Barefoot

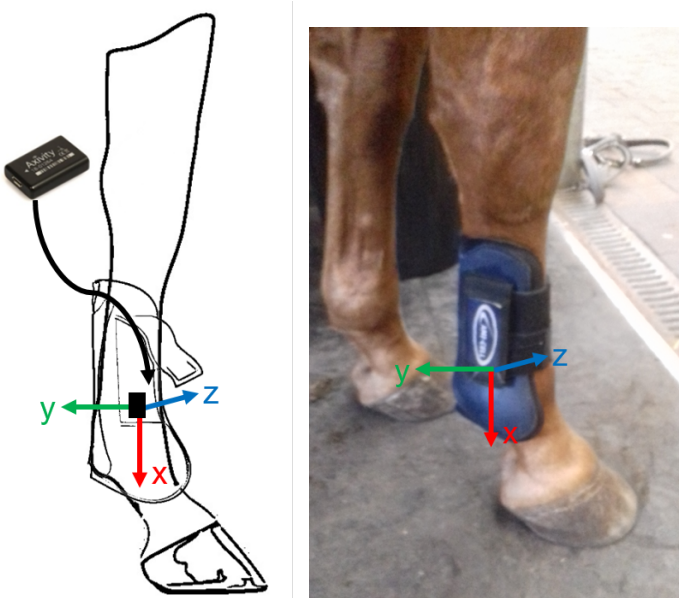
Table 1: Participating horses with breed class, height at withers, gender, age, state and type of shoeing.

variety of horses is suitable for our research since the difference in characteristics will contribute to the generalization of the machine learning model because accelerometer data patterns will be different for the different subjects. For example, the mean acceleration value per second during the gaits trot and walk are higher for ponies than for horses (Burla et al., 2014). Also, lame horses have asymmetrical gait patterns because they consistently shorten the cranial (forward) phase of stride (Davidson, 2018). The exercising for data recording is carried out by the owners or familiar riders at a local training arena with a size of 25 m x 38 m and a track surface of sand mixed with GEOPAT polyflakes. A minority of the data is gathered on a meadow and a clinker brick underground.

2.2. Data collection procedure

All six subjects, are wearing two single triaxial Axivity AX3 accelerometers (Axivity Ltd, Newcastle, United Kingdom), one on each front leg, as depicted in Figure 1. They were exercised in the different gaits walk, trot and canter for about 15 min each, either ridden or longed. The gait walk is also measured on a field and hard underground for horse 2. Horse 2 and 4 performed in addition other activities like rolling, pawing and flank watching. The orientation of the right accelerometer when the horse is standing is shown in Figure 1 and the three colored axis indicate the orientation of the accelerometer axis. This orientation was respected for all horses since a study (Thompson et al., 2018) revealed that the highest accuracies for detecting gaits could be reached at this location using an accelerometer. For successful data capturing the AX3

is securely fastened with the use of VELCRO stick on circles to the tendon boot with minimal room for vibration, slip or twist; to preserve that only the motions of the horse are captured. This is in contrast to many existing products that focus on easy installation at non-appropriate locations, at the cost of reduced accuracy. A second accelerometer is attached in the same way to the left leg. Observations on the activities



(a) Technical illustration (b) Real life attachement

Figure 1: Position and orientation (X, Y, and Z axes) of the right accelerometer

of the horses were made with video recordings at the same time as data from the sensors is collected. Table 2 lists the considered activities in this study with their descriptive definitions and the number of samples taken. All the data is labelled based on the video recordings using ELAN since it is difficult to use direct observation in combination with training of the horse. ELAN is a tool that allows such type of labelling procedure and is used by animal scientists for the video analysis and codification of images (Brugman et al., 2004), (Liebal et al., 2013). Annotations can be made by selecting the length of the segment where the behaviour is performed and typing the annotation.

2.3. Accelerometer data

Accelerometers fitted to the lateral side of the tendon boot with a size of 23 x 32.5 x 7.6 mm and a weight of 11 g are used, as shown in Figure 1.

Observed activities	Description	Samples	Subj.
Stand	The horse is standing on at least three legs with no movement to another place.	92121 (9.61%)	1-6
Walk	The horse performs a four beat gait with its legs following this sequence: left hind leg, left front leg, right hind leg, right front leg, leaving three feet on the ground.	406939 (42.43%)	1-6
Trot	The horse performs a two beat diagonal gait where the diagonal pairs of legs move forward at the same time with a moment of suspension between each beat.	327015 (34.10%)	1-6
Canter	The canter is a three beat gait. This gait starts with the hind leg then leads to the front in a rocking motion. This gait has a period of suspension after each stride.	110706 (11.54%)	1-6
Roll	The horse starts in a lying position on the side called "lateral recumbency" and rotates the body over its back, alternately from one side to another, remaining parallel to the performing surface.	11884 (1.24%)	4
Paw	The horse scrapes the ground with a forelimb.	5948 (0.62%)	4
Flank watching	The horse looks at its side or flank.	4462 (0.47%)	2

Table 2: Description of the observed activities with the relative and absolute number of samples and the subjects performing the activity (Sutton et al., 2013).

These log data with configurable sampling rates ranging from 12.5 Hz to 3200 Hz. The data logger is powered by a 150 mAh lithium–polymer battery, rechargeable via USB connection, which enables measurements over 30 days at 12.5 Hz and 14 days at 100 Hz. Acceleration is measurable on x-, y-, z-axes with a maximum sensitivity of $\pm 16g$ [$g = m/s^2$]. Setup and configuration of the AX3 sensors for recording is done with the AX3 OMGUI Configuration and Analysis Tool, which is an open source application. Data is recorded on an integrated memory with a capacity of 512 MB. It was transferred to a computer after recording via USB connection and stored in a Continuous Wave Accelerometer format.

An attachment convention for device orientation assists in consis-

	25 Hz	50 Hz	100 Hz	200 Hz	1600 Hz
Time measured [s]	2752	5492	3006	2560	417
Number of subjects	3	6	3	3	1
Number of behaviours	4	7	4	5	4

Table 3: Total time of movement data, number of subjects and number of investigad behaviours for each sampling rate of the merged accelerometer data.

tent and comparable datasets being gathered. The orientation of the accelerometer respected for all horses is depicted in Figure 1 with the USB port configured to point towards the ground as is suggested by the AX3 user manual. The AX3 has a built in, real-time clock (RTC) and calendar which provides the time base for the recorded acceleration data. These sensor data are sampled at four different sampling rates i.e., 25 Hz, 50 Hz, 100 Hz and 200 Hz. Each AX3 was set to record with a range of $\pm$ 8g for all the datasets except for one high sampling rate measurement at 1600 Hz the range is increased to $\pm$ 16g since this measurement was necessary for another research topic. Table 3 gives an overview of the total time measured, the number of subjects and the number of investigated behaviours at each sampling rate. Other captured behaviors such as cross canter, kicking backwards, trot to canter, etc are removed from the final dataset.

2.4. Machine learning model

A multilayer convolutional network as depicted in Figure 2, is used with two convolutional layers, which are followed by max-pooling layers, and two fully connected layers. The output of the last fully-connected layer is fed to a 7-way softmax layer which produces a distribution over the seven class labels: stand, walk, trot, canter, roll, paw, flank watching. The first convolutional layer filters the $n \times 6 \times 1$ input acceleration data with 64 kernels of size 3×1 and stride 1. The L2 regularization technique is used in this layer with a weight decay coefficient of 0.01 (Mallouh et al., 2019). After the first convolutional layer a zero-padding is used such that the output has the same length as the original input. Then a max-pooling operation is done. The second convolutional layer takes as input the (pooled) output of the first convolutional layer and filters it with 16 kernels of size 5×2 and stride 1. Both layers contain an

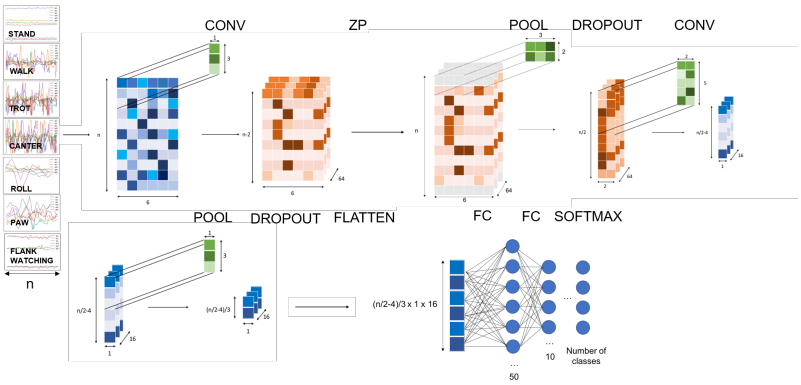


Figure 2: The structure of the CNN with n the number of samples as input.

activation layer using rectified units (ReLUs) and dropout of 0.55 is used (Srivastava et al., 2014). The Adam optimizer is used for training the neural network through back propagation. Training is done for 400 epochs, with an early stopping criterion of halting training when there is no increase in accuracy during the last 60 epochs (Ronao and Cho, 2016). The training set is used to train the model, while the validation set is only used to evaluate the model's performance. Table 4 shows the experimental setup.

Parameter	**Value**
The size of the input vector	n
The number of input channels	6
The number of feature maps	64-16
Filter size	3×1 - 5×2
Stride	1
Pooling size	2×3 - 3×1
Activation function	ReLu and Softmax
Weight decay	0.01 (L2 regularization)
The probability of dropout	0.55
Maximum epochs	400
Optimization (back propagation)	Adam optimizer

Table 4: Experimental setup.

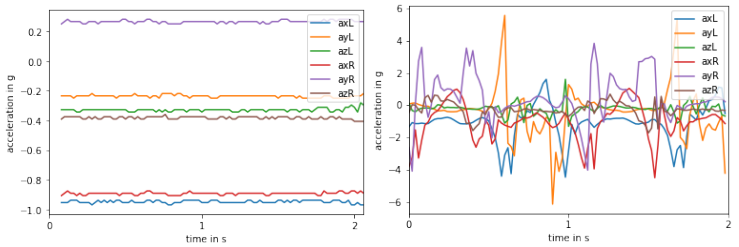
Correct selection of the evaluation criteria is crucial for evaluating the merits of a model. In this work, the overall model accuracy and confusion matrices are considered for the model performance assessment. The overall model accuracy is the number of true positive instances of all behavioural classes divided by the total number of instances in the test set. Here, true positive is the number of instances where the behaviour was correctly classified by the algorithm using video-observations as reference. A target accuracy of 99% or higher is

the objective. A paired Wilcoxon signed-rank test is used to obtain the p-values for determining statistical significance. If the p-value ≤ 0.05, there is statistical evidence that the reached accuracies are not equal. A normalized confusion matrix is used to evaluate the designed model where the diagonal elements represent the percentage for which the predicted label is equal to the true label, while off-diagonal percentages are those that are mislabelled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions. The normalized confusion matrices presented in this paper contain merged training and validation data.

3. Results

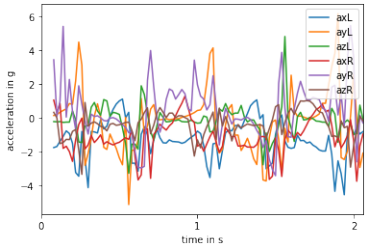
Figure 3 illustrates exemplar two second data windows of the four gaits and the other behaviours, from the left and right accelerometer worn on the lateral side of the tendon boot.

To train the convolutional neural network, separate training and validation sets are needed and can be selected in various ways. If all accelerometer data of one subject is not in the training set, but in the test set, the subject is "unseen" in terms of network training. That means the test results will be a good indication of performance against completely new subjects. First, a training and validation set are obtained by automatically splitting the training and the validation data with a fixed ratio of 66/34 so the model is validated on data from a seen horse referred to as the 'First dataset'. Secondly, the 50 Hz dataset, which contains all seven considered behaviours, is resampled to 25 Hz, 100 Hz and 200 Hz and merged with the original dataset at that sampling rate referred to as the 'Second dataset'. The model can then be assessed for any behaviour at each sampling rate. Again automatic split testing is used to obtain the training and validation set so the model is again validated on data from a seen horse. Finally, the separation of the training and validation data is attained manually and as a result the model is validated on data from an unseen horse referred to as the 'Third dataset'. In this case data from the lame horse is used to validate our model while it is trained on healthy horses, to further asses the

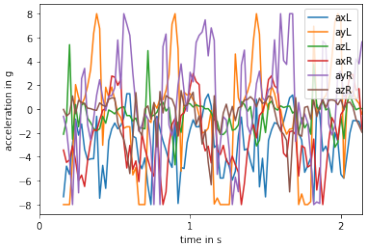


(a) Stand

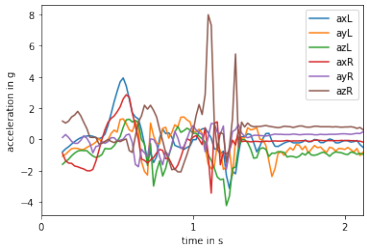
(b) Walk



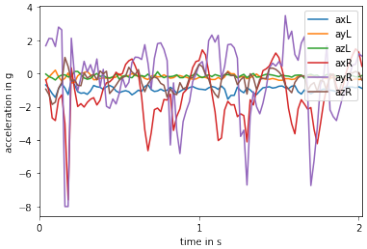
(c) Trot



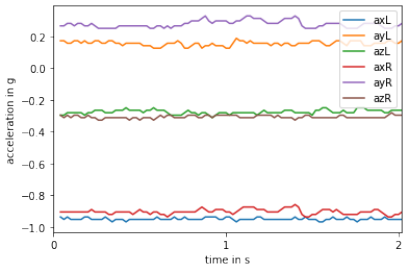
(d) Canter



(e) Roll



(f) Paw



(g) Flank watching

Figure 3: Typical accelerometer patterns of (a) stand, (b) walk, (c) trot, (d) canter, (e) roll, (f) paw and (g) flank watching in a 2 s window. The blue, yellow, green lines represent X,Y,Z signals from the left accelerometer and the red, purple and brown lines represent X,Y,Z signals from the right accelerometer , respectively.

generalization of the model. Table 5 gives an overview of the train and test sets, the total time of movement data in seconds for each behavior and the subjects present in each dataset at each sampling rate.

Set			25 Hz	50 Hz	100 Hz	200 Hz
1	66% train 34% test	S	259	450	140	189
		W	1251	2292	843	830
		T	629	1368	916	671
		C	170	398	311	219
		R	0	62	0	41
		P	0	51	0	0
		F	0	80	0	0
		Subjects	1-3	1-6	1-3	2-4
2	66% train 34% test	S	709	450	590	639
		W	3543	2292	3135	3122
		T	1997	1368	2284	2039
		C	568	398	709	617
		R	62	62	62	103
		P	51	51	51	51
		F	80	80	80	80
		Subjects	1-6			
3	train	S	209	424	132	176
		W	1037	2137	720	646
		T	515	1270	816	582
		C	144	375	281	208
		Subjects	1,2	1,2,4,5,6	1,2	2,4
	test	S	50	26	8	13
		W	214	155	123	184
		T	114	98	100	89
		C	26	23	30	11
		Subject	3			

Table 5: The train and test set, the total time of movement data in seconds per behavior (S = stand, W = walk, T = trot, C = canter, R = roll, P = paw, F = flank watching) and the subjects for each sampling rate present in the first dataset, second and third dataset.

3.1. Effects of the sampling rate

The first topic of investigation was the effect of the sampling rate on the classification accuracy. In Figure 4 the mean performance of the CNN for a time interval ranging from 0.6 s to 3 s with increasing sampling rate is depicted for the three datasets. The number between the brackets indicates the number of behaviours that are taken into account in the training and validation of the CNN. For all datasets the accuracy increases (on average from 94.74% to 98.88%) when the sampling rate is increased from 25 Hz to 50 Hz ($p\text{-}value \leq 0.05$). From 100 Hz to 200 Hz, the accuracy for two out of three datasets decreases decreases on average from 99.60% to 99.27% ($p\text{-}value \leq 0.05$). As can be concluded from this graph, for a sampling rate of 25 Hz, the CNN performs the best on the second dataset i.e. when all behaviours and all horses are taken into account since in the first and third dataset the misclassifi-

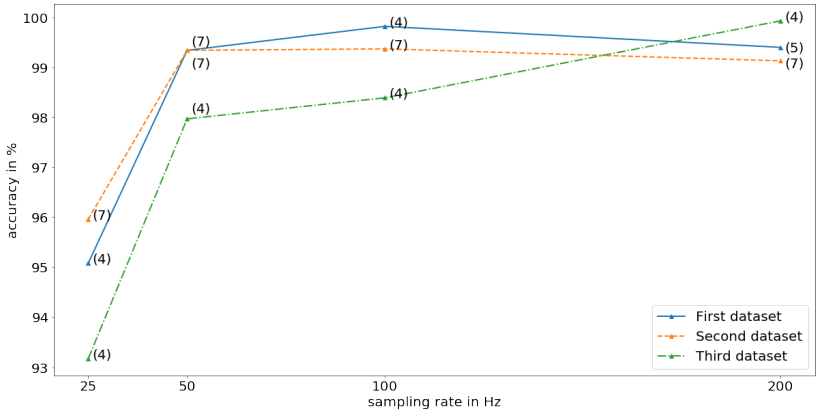
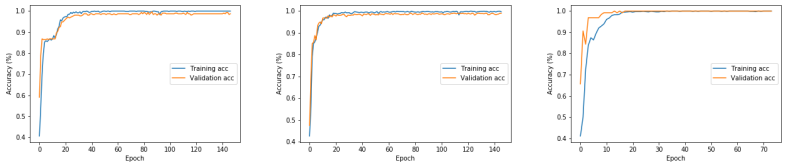


Figure 4: Mean performance of the convolutional neural network with increasing sampling rate for three datasets. The number between the brackets is the number of different activities available in the dataset.

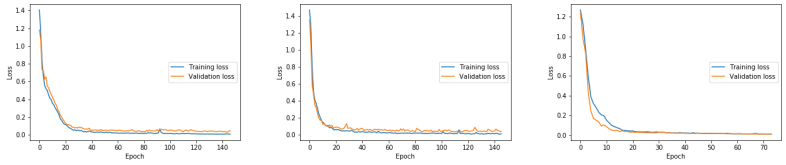
cation of 'canter' brings down the average accuracy at this sampling rate. The CNN validated on the data of the lame horse performs the least in the sampling rate range from 25 Hz to 100 Hz. At a sampling rate of 200 Hz, the CNN performs best when the training and validation data are split up by hand and the model trained on all behaviours performs the least ($p\text{-}value \leq 0.05$). A possible explanation could be that it is necessary to increase the model complexity at this sampling rate to reach the same accuracy in predicting seven behaviours as in predicting four behaviours. As is depicted in Figure 5 an increase in the number of epochs after which the model is halted is not going to further increase the model accuracy for the first and second dataset since one can see that the loss is remaining more or less constant after 40 epochs. The third dataset fits perfect since the training and validation loss are almost equal. Moreover, the accuracy of the model for the three datasets decreased on average with about 4.75% when the sampling rate was decreased from 200 Hz to 25 Hz. This decrease in the ability of accelerometers to identify locomotion behaviour patterns when the sampling rate decreases was also remarked when monitoring cows' behaviours (Benaissa et al., 2017).

3.2. Effect of the time interval

The second investigated matter was how the time window size influenced model performance for the three datasets. For this purpose we varied the time series length between 0.6 s and 3 s with a step size of



(a) First dataset (b) Second dataset (c) Third dataset



(d) First dataset (e) Second dataset (f) Third dataset

Figure 5: Accuracy and loss plots of the training set (blue) and validation set (orange) for a 2 s time interval at a sampling rate of 200 Hz.

0.2 s. For each value the mean performance of the CNN for the three datasets for sampling rates between 25 Hz and 200 Hz are presented in Figure 6. The mean duration of the behaviours are annotated with black striped lines except for the flank-watching movement since the mean duration of this behaviour lies outside the investigated time intervals at 4.866 s. From the videofiles combined with the accelerometer data, the mean duration of each activity was calculated by taking 10 samples of each horse according to their description in Table 2. It is important to take the mean duration of each behaviour in consideration since this could affect the performance of the model.

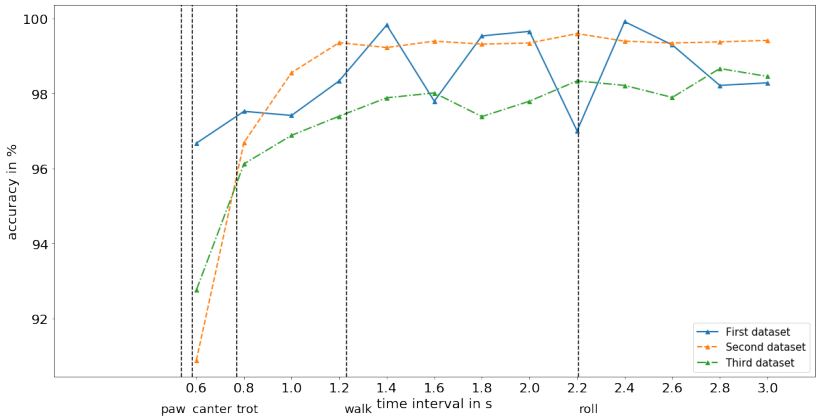


Figure 6: Mean performance of convolutional neural network with increasing time interval for the three datasets.The black striped lines indicate the mean duration of each activity.

Figure 6 shows that larger time intervals do not necessarily lead

to better classification results. While an increase in the time interval from 0.6 s to 1.2 s (a full walk cycle) gains a significant performance boost of 4.92% on average for all datasets ($p\text{-}value \leq 0.05$), its further growth introduces only moderate improvements for the third dataset of 1.06% and no improvements for the second dataset ($p\text{-}value > 0.05$). At lower time intervals (n <1.2 s) the gaits which have the biggest share (97,68%) in the dataset get misclassified more often. The first dataset is unstable due to the full misclassification of 'canter' at a sampling rate of 25 Hz in 8 out of 13 time intervals. In most of this cases 'canter' is classified as 'trot' and to a lesser extend as 'walk' which means that the model learned the wrong features or generalized not well with the learned features. As can be noticed the mean accuracy for the third dataset lies lower than for the first and second dataset due to again a high misclassification of 'canter' at a sampling rate of 25 Hz. For the first and second dataset, the largest contributors to a lower accuracy are the misclassification of 'canter', 'roll', 'paw' and 'flank-watching'. Moreover, the accuracy of the model for the three datasets decreased on average with about 5.27% when the time interval was decreased from 3 s to 0.6 s. These findings are in agreement with the results of (Ignatov, 2018), where the dependence of human activity recognition accuracy by convolutional neural networks using accelerometer data and the time window length was investigated.

3.3. Combination of the time interval and sampling rate

A third analysis examined on which combination of lowest time interval and sampling rate the model was best performing. For this purpose accuracy surface plots as depicted in Figure 7 for the three datasets as function of the time interval and sampling rate were generated by fitting a polynomial of degree two through the obtained datapoints indicated as blue dots. The low predicted accuracies are indicated with the colour blue and the high ones with the colour red. The combinations that are the least performing for the three datasets are observed in the region where both sampling rate and time interval are low.

As can be seen from the contourplots shown in Figure 8 a 100%

accuracy is reached in the red region. As indicated with yellow cross markers, the combinations that gain an accuracy of 100% at the lowest sampling rate and the shortest time interval are for the first dataset observed in the region where the value of the sampling rate ranges between 64 Hz at a time interval of 2.05 s and 170 Hz at a time interval of 0.85 s, for the second dataset in the region where the value of the sampling rate ranges between 36 Hz at a time interval of 2.4 s and 170 Hz at a time interval of 1 s, for the third dataset in the region where the value of the sampling rate ranges between 90 Hz at a time interval of 2.3 s and 170 Hz at a time interval of 1 s. The lowest sampling rate together with the corresponding length of time interval for three levels of accuracy for the second and third dataset are listed in Table 6. A 99% accuracy can thus be reached with a sampling rate of 25 Hz and a time interval 2.1 s for detecting seven activities of a seen horse or the movement of an unseen horse resembling the data in the training set.

	Second dataset (seen horse)			Third dataset (unseen horse)		
	98%	99%	100%	98%	99%	100%
f (Hz)	25	25	36.5	52.5	69	90
n (s)	1.8	2.1	2.4	2.5	2.4	2.3

Table 6: Time interval and sampling rate predictions for a seen horse and an unseen horse for three levels of accuracy.

3.4. Effects of the underground

A fourth analysis explored how the model accuracy was influenced by the type of underground the horse was walking on. Model accuracy for the class walk is studied for four different surfaces: dry sand mixed with polyflakes, wet sand mixed with polyflakes, meadow and a hard brick underground. The normalized confusion matrices are depicted in Figure 9.

As can be seen from the normalized confusion matrices for different time intervals, the class walk on a wet underground and on a dry underground get classified with an accuracy above 98% for every time interval. The class walk on a hard underground reaches accuracies higher than 86%. The class walk on a field swings between 15% and 86% classification accuracy. As can be concluded from the results presented in the normalized confusion matrices, the data gathered from

different undergrounds is significantly different so that the model could conclude from which underground the data was gathered for three out of four undergrounds.

Normalized confusion matrices with all activities included are shown in Figure 10.

As can be concluded from the confusion matrices, at small time intervals, more misclassification is taking place than at higher time intervals ($n \geq 1.2$ s). At those higher time intervals 'Walk-F' is performing the worst with accuracies swinging between 1% and 74% since it gets misclassified as 'Walk-H'. The other 'walk classes' get classified with high accuracies between 93% and 100% at higher time intervals. All the walk movements get classified as walk, independent of the underground, at any time interval. The other movements that are now included get classified in a few cases as one of the 'walk classes'. To the best of authors' knowledge, the influence of the surface on the activity classification performance of a CNN based on accelerometer data has not been studied previously and so no comparison with literature could be made.

3.5. Horse-wise cross validation

Examination of the generalizing capabilities of the model was executed by inspecting how model accuracy was influenced by the type of validation horse. A ten-fold leave one out cross-validation strategy was used. Therefore, data collected on five horses was used to train the system and then the system was tested by classifying the data of the sixth horse accordingly. The 50 Hz original data set was split manually into training and test sets, with the CNN performance investigated on unseen data. Since not all horses practiced every behaviour, only the four movements i.e. 'stand', 'walk', 'trot' and 'canter' performed by each horse are investigated for a time window of 2.5 s. The number of instances per class for each horse are shown in Table 7. The performance of the CNN validated on the six subjects is presented as boxplots in Figure 11.

As can be concluded from the boxplots, the model validated on Horses 1-2-4-5 reaches all mean accuracies above or equal to 99.65%,

	Classes			
	S	W	T	C
Horse 1	33	258	99	4
Horse 2	22	157	90	48
Horse 3	7	55	33	7
Horse 4	62	163	121	49
Horse 5	7	129	113	22
Horse 6	19	104	53	16

Table 7: Movement class instances at a 2.5 s time window of the studied movements for each horse (S = stand, W = walk, T = trot and C = canter).

while the validation of the model on Horses 3 and 6 is performing the least. Those results are in line of expectation since Horse 3 is a lame horse with asymmetrical gait patterns and Horse 6 is a pony with higher mean acceleration values during the gaits 'walk' and 'trot'. The overall mean validation accuracy of the model validated on the lame horse is 97.84% due to the misclassification mainly of 'trot' and to a lesser extend 'canter'. For the pony, the overall mean validation accuracy of the model is 96.10% and the classes that are least performing are again 'trot' and 'canter'.

3.6. Effect of the number of sensors

A final analysis examined how model accuracy was influenced by the number of sensors. Ten-fold cross-validation was used with different amounts of sensor data. The original 50 Hz data set was split automatically into training and test sets, with CNN performance examined for a time window of 2.5 s. The second convolutional layer size was modified from 5×2 to 5×1 to meet the required variable dimensions between input, hidden, and output layers. Performance validation for the CNN using one or two sensors is presented as boxplots in Figure 12.

As can be concluded from the boxplots, the model validated on two sensors reaches a mean accuracy of 99.66% while with data from one sensor a mean accuracy of 99.42% is reached ($p\text{-}value \leq 0.05$). This decline in accuracy is not attributable to the performance of one specific class.

4. Conclusion

In this study we propose a solution for a horse activity recognition problem that is based on Convolutional Neural Networks with the use of accelerometer time series. It has the benefits of using short recognition intervals of size up to 2.1 s and small sampling rates up to 25 Hz for reaching accuracies of 99% and requiring no feature engineering.

To evaluate the performance of the considered approach we tested it on three experimental datasets. The obtained results demonstrate that the proposed CNN-based model establishes high accuracies at a lot of time intervals and sampling rates. A reduction in the sampling rate and time interval length did reduce the overall classification accuracy of the model on average with 4.75% and 5.27%, respectively. The experiment has further emphasized an architecture that can be applied not only to different subjects, but can be used in different measurement conditions such as on different types of undergrounds. Also, data from one accelerometer appears to be sufficient to classify seven behaviours of six different horses with an overall mean accuracy above 99%.

Our suggested approach demonstrates superior potential in most cases as shown by the above experimental results, but the main limitations of this study are the number of horses (six in this study) with data of only one pony present in our dataset and the fact that not every horse practices all behaviors.We conjecture that, with more training data of different breeds, our behavior detector will be more robust to these different cases.

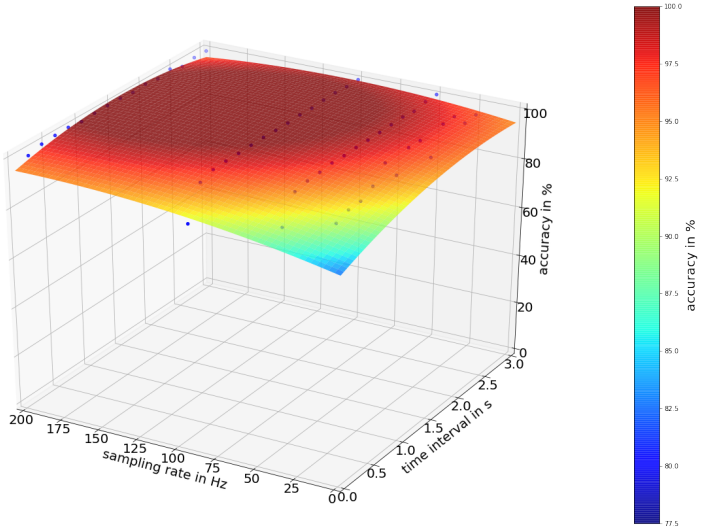
Future work will include capturing and analyzing more behaviours which horses are performing during training or related to horses experiencing an episode of colic like: kicking the abdomen, stretching and attempting to lie down. Also, further investigation needs to be done concerning the eating and drinking behaviour since this could give extra information about the well-being of the horse. In addition, activity measurements could be performed to conclude if a horse is agitated or depressed. Further study at lower sampling rates and a reduction in the number of accelerometer axis is needed since this could reduce computational cost, storage load and energy use and therefore available datasets can be resampled and re-analyzed. Also, we need further study

for the analysis of the features extracted automatically by the convent and compare them with the well-known hand-crafted features. Further study on the characteristics of the used CNN and utilizing larger datasets should be conducted.

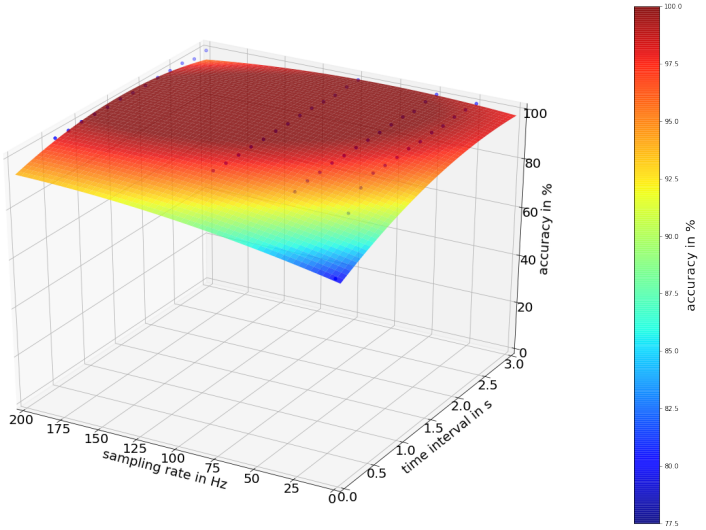
Acknowledgments

M. Deruyck is a Post-Doctoral fellow of the FWO (Research Foundation - Flanders, Belgium).

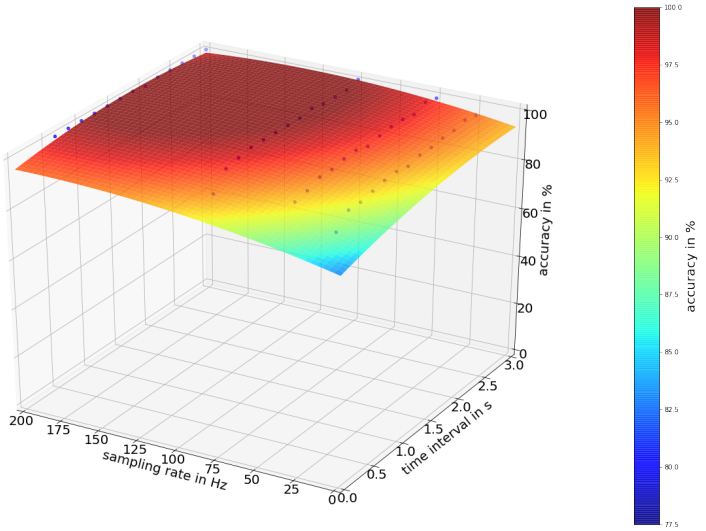
This work was executed within the imec.icon project Hoof-MATE, a research project bringing together academic researchers and industry partners. The Hoof-MATE project was co-financed by imec and received project support from Flanders Innovation Entrepreneurship (project nr. HBC.2018.0536).



(a) First dataset

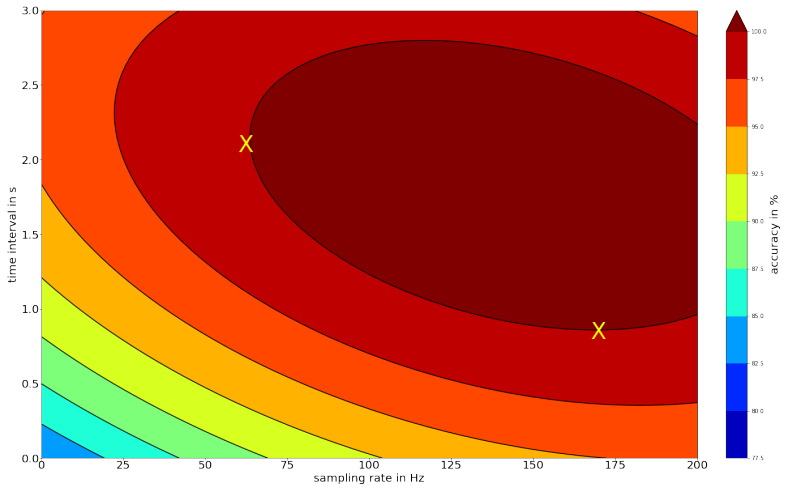


(b) Second dataset

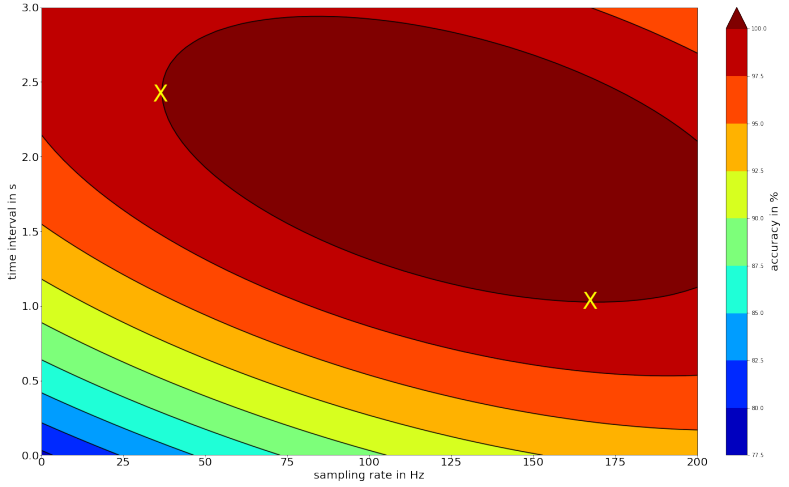


(c) Third dataset

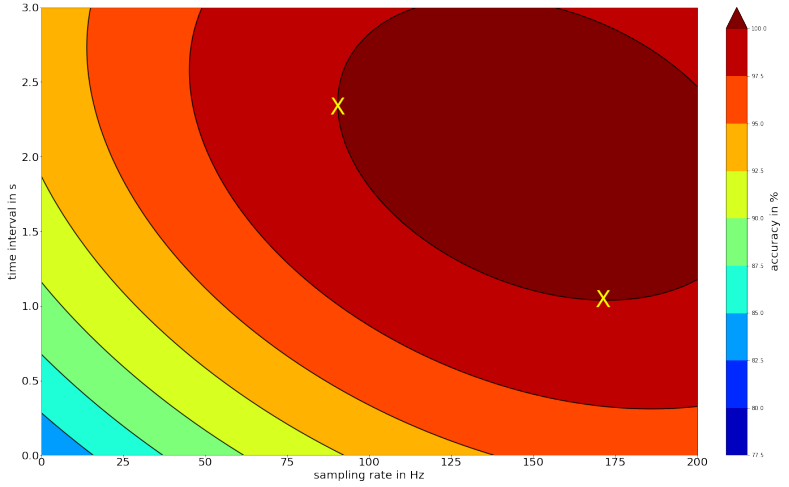
Figure 7: Accuracy surface plot as function of the sampling rate and the length of the time interval for three datasets.



(a) First dataset

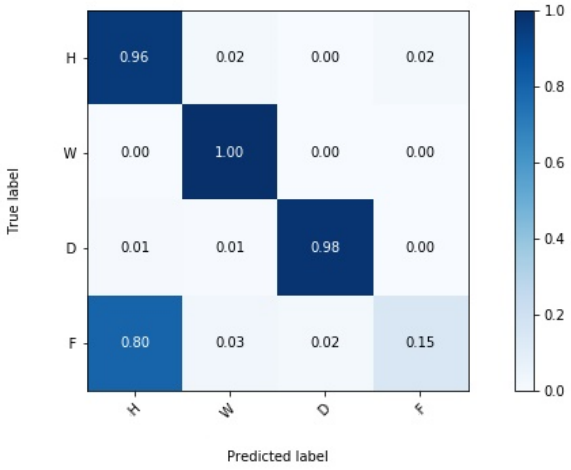


(b) Second dataset

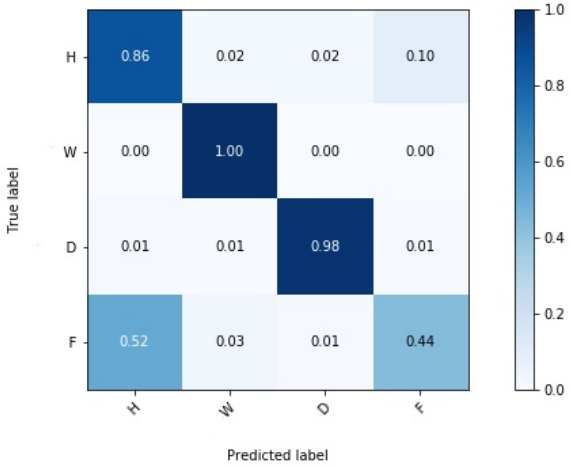


(c) Third dataset

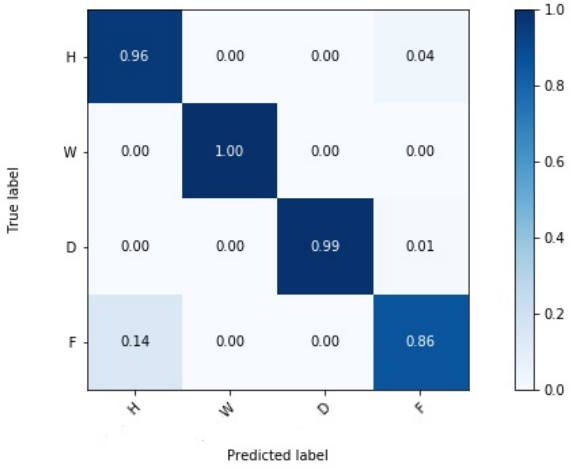
Figure 8: Accuracy contour plot as function of the sampling rate and the length of the time interval for three datasets.



(a) n = 0.6 s

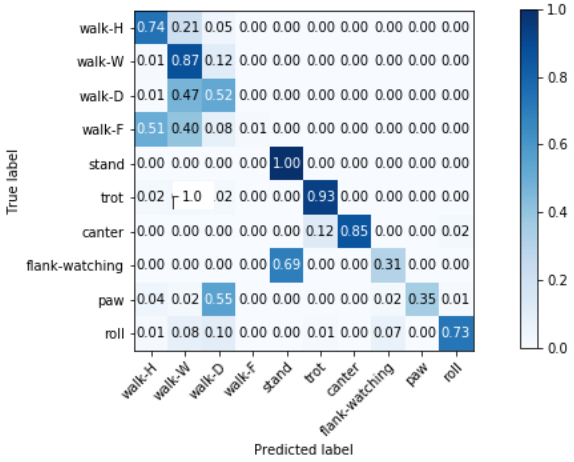


(b) n = 1.2 s

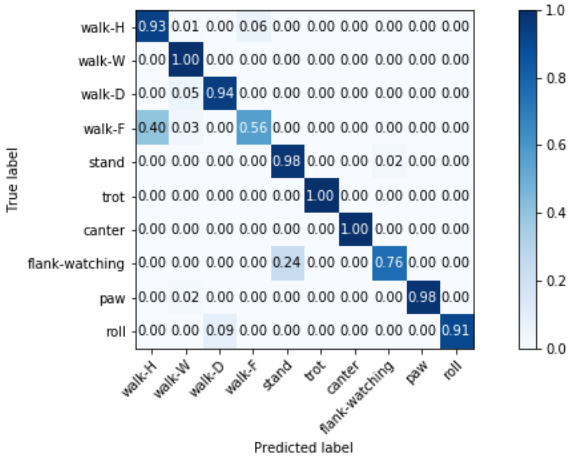


(c) n = 2.4 s

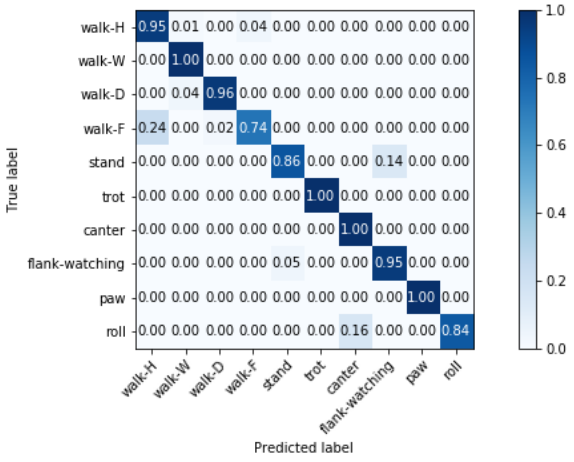
Figure 9: Normalized confusion matrix for training and test set of the behaviour 'walk' at a sampling rate of 50 Hz for different time intervals and four types of underground (H= hard, W = wet, F= field and D = dry).



(a) n = 0.6 s



(b) n = 1.2 s



(c) n = 2.4 s

Figure 10: Normalized confusion matrix for training and test set at a sampling rate of 50 Hz for different time intervals and four types of underground (H= hard, W = wet, F= field and D = dry) including all activities.

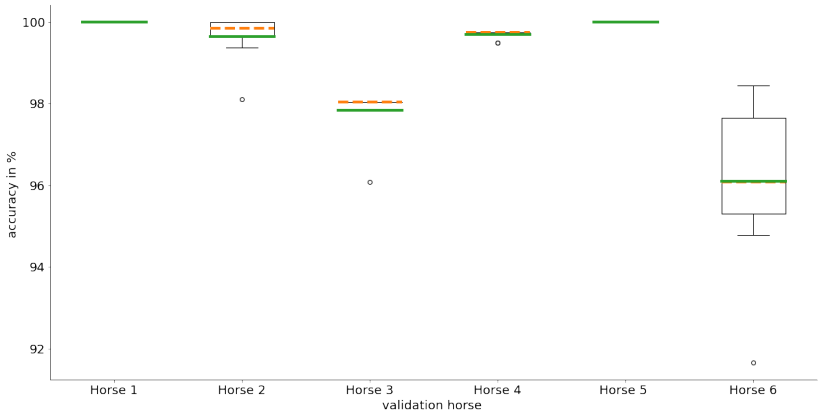


Figure 11: Performance of the convolutional neural network validated for six horses with three types of breed classes horse (Horse 1-4), Friesian horse (Horse 5) and Pony (Horse 6), given as boxplots with mean (solid green line), medians (dashed orange line), interquartile, absolute ranges and outliers.

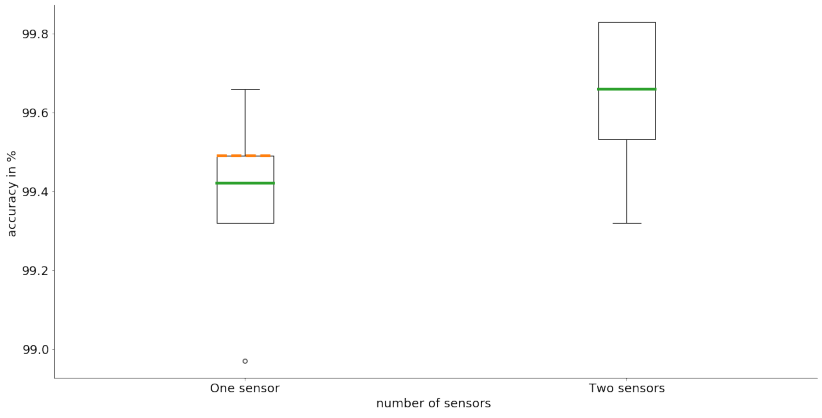


Figure 12: Performance of the convolutional neural network validated on one and two sensors, given as boxplots with mean (solid green line), medians (dashed orange line), interquartile, absolute ranges and outliers.

References

Benaissa, S., Tuyttens, F. A., Plets, D., de Pessemier, T., Trogh, J., Tanghe, E., Martens, L., Vandaele, L., Van Nuffel, A., Joseph, W., and Sonck, B. (2017). On the use of on-cow accelerometers for the classification of behaviours in dairy barns. *Research in veterinary science*.

Bidder, O. R., Campbell, H. A., Gómez-Laich, A., Urgé, P., Walker, J., Cai, Y., Gao, L., Quintana, F., and Wilson, R. P. (2014). Love thy neighbour: automatic animal behavioural classification of acceleration data using the k-nearest neighbour algorithm. *PloS one*, 9(2):e88609.

Brugman, H., Russel, A., and Nijmegen, X. (2004). Annotating multimedia/multi-modal resources with elan. In *LREC*.

Burla, J.-B., Ostertag, A., Westerath, H. S., and Hillmann, E. (2014). Gait determination and activity measurement in horses using an accelerometer. *Computers and electronics in agriculture*, 102:127–133.

Davidson, E. J. (2018). Lameness evaluation of the athletic horse. *Veterinary Clinics: Equine Practice*, 34(2):181–191.

Ignatov, A. (2018). Real-time human activity recognition from accelerometer data using convolutional neural networks. *Applied Soft Computing*, 62:915–922.

Langrock, R., King, R., Matthiopoulos, J., Thomas, L., Fortin, D., and Morales, J. M. (2012). Flexible and practical modeling of animal telemetry data: hidden markov models and extensions. *Ecology*, 93(11):2336–2342.

Liebal, K., Waller, B. M., Slocombe, K. E., and Burrows, A. M. (2013). *Primate communication: a multimodal approach*. Cambridge University Press.

Mallouh, A. A., Qawaqneh, Z., and Barkana, B. D. (2019). Utilizing cnns and transfer learning of pre-trained models for age range classification from unconstrained face images. *Image and Vision Computing*.

Ravi, D., Wong, C., Lo, B., and Yang, G.-Z. (2017). A deep learning approach to on-node sensor data analytics for mobile or wearable devices. *IEEE journal of biomedical and health informatics*, 21(1):56–64.

Ronao, C. A. and Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Sutton, G. A., Dahan, R., Turner, D., and Paltiel, O. (2013). A behaviour-based pain scale for horses with acute colic: scale construction. *The Veterinary Journal*, 196(3):394–401.

Thompson, C. J., Luck, L. M., Keshwani, J., Pitla, S. K., and Karr, L. K. (2018). Location on the body of a wearable accelerometer affects accuracy of data for identifying equine gaits. *Journal of equine veterinary science*, 63:1–7.

Um, T. T., Babakeshizadeh, V., and Kulić, D. (2017). Exercise motion classification from large-scale wearable sensor data using convolutional neural networks. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 2385–2390. IEEE.

van Loon, J. P. and Van Dierendonck, M. C. (2015). Monitoring acute equine visceral pain with the equine utrecht university scale for composite pain assessment (equus-compass) and the equine utrecht

university scale for facial assessment of pain (equus-fap): a scale-construction study. *The Veterinary Journal*, 206(3):356–364.

Zhao, X., Wei, H., Wang, H., Zhu, T., and Zhang, K. (2019). 3d-cnn-based feature extraction of ground-based cloud images for direct normal irradiance prediction. *Solar Energy*, 181:510–518.