2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT) | 979-8-3503-4649-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/DCOSS-IOT58021.2023.00029

Terrain type detection for smart equine gait analysis systems using inertial sensors and machine learning

Jeanne I.M. Parmentier
Dept. of Clinical Sciences
Utrecht University
Utrecht, The Netherlands
Pervasive Systems Group
University of Twente
Enschede, The Netherlands
j.i.m.parmentier@uu.nl

Filipe M. Serra Bragança
Dept. of Clinical Sciences
Utrecht University
Utrecht, The Netherlands
f.m.serrabraganca@uu.nl

Elin Hernlund
Dept. of Anatomy, Physiology and Biochemistry
Swedish University of Agricultural Sciences
Uppsala, Sweden
elin.hernlund@slu.se

Berend Jan van der Zwaag
Pervasive Systems Group
University of Twente
Enschede, The Netherlands
b.j.vanderzwaag@utwente.nl

***Abstract*—Lameness, limping due to pain, is a significant welfare issue for horses. Veterinarians typically evaluate horses on two terrain types (hard and soft, e.g., asphalt and sand) that are known to affect the observed degree of lameness based on the origin/location of the pain. In the past years, whole-body inertial measurement units (IMU)-based gait analysis systems were developed to support diagnostics and monitor locomotion changes over time. Movement direction and gait (walk, trot) are automatically labeled, resulting in smart and easy-to-use systems. However, terrain types are not detected, leading to information loss. In this work, we explored terrain classification tasks with equine IMU data and machine and deep learning. Using the data of 111 horses equipped with IMU sensors (withers, pelvis, front, and hind limbs), we compared different features-based (FT) and time-series-based (TS) classifiers (train-test ratio: 0.7–0.3). In order to reduce the computational costs of the future system, we also evaluated the performance (F1 score) of the classifiers with different sampling frequencies (10 to 200Hz) and different IMU combinations (body and limbs). Our Convolutional Neural Network models accurately classified terrain types with only one IMU placed on the front limb. Downsampling the signals led to similar results, thus enabling real-time applications.**

***Index Terms*—Artificial intelligence, horse, surface, inertial measurement units, context awareness**

I. Introduction

Lameness (limping due to pain) is one of the main reasons for veterinary consultations for horses [1]. During lameness exams, veterinarians observe the horse's locomotion at different gaits (mostly, walk and trot), directions of the movement (straight lines, circles), and terrains (hard/undeformable, such as concrete, asphalt, or bricks and soft/deformable, such as sand or sand-fiber mixtures) to try to pinpoint from which limb (left/right front or hind) the lameness originates from [2]. In the past decades, equine gait analysis systems (EGAS) using, among others, inertial measurement units (IMU) have been developed to objectively quantify locomotion and are used in both clinical and research settings [3], [4]. Gait, direction, and terrain types are known to have an effect on different gait variables interesting for veterinarians and extracted by EGAS [5], [6] and are thus important to keep track of, as they might give indications related to where the location of pain is and which underlying pathology is causing it [7]. While direction and gait are already automatically detected and labeled by some systems [4], [8], terrain type is not, leading to information loss or increased time consumption for the users. In a study from Hardeman et al. [9], one reason for not adopting gait analysis technology in veterinary practice was the time consumption and complicated usability. Thus, automated terrain detection in EGAS and other embedded applications can have several benefits:

- Development of context-aware EGAS, with easier use and increased adoption rate, supporting veterinary diagnosis during lameness exams and health check-ups
- Development of mobile gait monitoring systems for injury prevention, in the context of training
- Expansion of the available metadata, thus an increase of databases' quality

In this study, we explored the minimal IMU setup (all, withers and pelvis, pelvis only, pelvis and one front or hind limb, front and hind limbs, front or hind limb only) needed to reach an accurate terrain classification, with all gaits pooled together (walk and trot). We also looked at the effect of downsampling the data on the classification performance. Finally, we compared different feature-based and time-series-based classifiers in terms of classification performance, but also computational costs and model weight for embedded and real-time classification.

The rest of the paper is organized as follows: Section II reviews the related works on terrain classification using (human) locomotion data. Section III describes our methodology: data collection and data labeling, data processing and downsampling, features extraction and selection, feature-based and time-series classifiers used, performance and computational costs evaluation. Section IV presents our terrain classification results and the effect of using different IMU sets, sampling frequencies and classifiers. It also describes the computational costs when using one IMU data as input. Section V provides a discussion of the results. Finally, Section VI concludes and discusses future perspectives.

2325-2944/23/$31.00 ©2023 IEEE
DOI 10.1109/DCOSS-IoT58021.2023.00029

II. Related works

Terrain classification is a research topic that has applications in health and training monitoring for humans or in autonomous vehicles, legged robots, and drones. Terrains are usually classified using different types of data sources as input, such as optical cameras [10], or wearables [11]. We focus on studies that used wearables (i.e., IMU or accelerometers alone) data to explore terrain classification tasks.

To the authors' knowledge, only one study looked at the terrain classification task for horses [12]. In this previous work, we benchmarked different feature selection algorithms (FSA) and trained linear Support Vector Machine (SVM) models with different amounts of selected features for binary classification of the terrain type (Hard or Soft), for two gaits (walk and trot) separately. While optimizing classification performance was not the goal of this paper, our best models trained with trot data reached F1 scores above 90% when training with 50 features and more, while the best walk models reached F1 scores above 80% when training with 50 features and more (all IMU locations combined).

Several studies were conducted in the human field to explore terrain types and terrain regularity. Dixon et al. [11] compared feature-based classifiers (Gradient Boosting decision trees) and time-series-based classifiers (Convolutional Neural Networks) to classify terrains, using accelerometer data collected during running tasks. They attached accelerometers to the tibia and the lower-back of their participants, who had to run over three types of terrain (woodchip trail, synthetic track, concrete road). Their classification performances ranged from 86.7% to 97.0%, depending on the training input used (IMU location, number of features).

Hashmi et al. [13] also looked at terrain classification with data of human walk, with IMU attached to the chest or the lower-back. They trained SVM and Random Forest (RF) models to classify between different terrain categories: [indoor or outdoor], [hard or soft], [concrete or grass or asphalt or soil or tiles]. Their binary hard-soft classification reached an accuracy of 92.08% when training a RF classifier with the 200 lower-back IMU features but dropped to 79.87% when training an SVM with a subset of 100 angular velocities features of the same lower-back IMU.

Hu et al. [14] collected data of IMU attached to the right wrist, lower-back, bilateral thighs and bilateral shanks to classify uneven cobblestone, flat bank surface, stairs, and slope surface. They trained Convolutional Neural Network (CNN) and Long-Short-Term Memory (LSTM) neural network models with different IMU sets for this task. Their best classification performances were obtained when training with all sensors (wrist, lower back, bilateral thighs, bilateral shanks) for both CNN and LSTM (F1 score: 90%) while worst results were obtained when training with lower-back (LSTM F1 score: 79%) and left thigh (CNN F1 score: 71%).

Shah et al. [15] used a feedforward neural network and a CNN to classify features and raw signals extracted from IMU walk data (wrist, trunk, bilateral thighs, bilateral shanks) into nine terrain classes (flat-even, slope-up, slope-down, stairs-up, stairs-down, cobblestone, grass, banked-left, banked-right). Their best results were obtained with the right shank or lower back IMU data (F1 score: 78%).

The equine study [12] trained for both gaits separately and used only a linear kernel SVM classifier. We compared the features selected by different feature selection algorithms and showed that in horses, gyroscope features were not useful for this task. All the human studies focused on one type of gait (walk or running) for bipedal locomotion on straight lines. In these studies, data were segmented either with sliding windows [11], [12] or based on gait cycle [13]–[15]. They all compared the classification performances when training with different IMU location sets but did not look into the effect of sampling frequencies and did not look at the computational costs of their methods.

In this work, we explored binary terrain classification with machine learning and deep learning models for horse gait analysis applications, comparing different IMU location sets and sampling frequencies, regardless of the gait and the direction of the movement. We also evaluated the computational costs of our models for future embedded and real-time applications.

III. Methodology

Data processing, classifiers training and evaluation were performed in MATLAB 2022b (MathWorks, Natick, Massachusetts, USA). Statistical comparisons were conducted in RStudio v2.2.2 with packages *ggplot2* version 3.4.0 *nlme* version 3.1.157 and *emmeans* version 1.7.3.

A. Data Collection

Retrospective data from 111 horses were included in this study, including data from different published [16] and unpublished lameness induction studies, as well as data collected for healthy locomotion studies [12], [17]. All horses were led in-hand at different gaits (walk and trot), directions (straight line, right and left circles), and over different terrain types (undeformable, hard; deformable, soft), at different locations. Horses were equipped with six inertial measurement units (IMU) (ProMove-Mini, Inertia Technology B.V., The Netherlands), forming a network of wireless sensors synchronized within a precision of 100 ns as described in [4]. IMUs were fixed at the withers level with a custom girth, to the pelvis with double-sided tape, and to the lateral aspect of front and hind lower limbs (cannon bones) with custom pockets fixed to brushing boots, shown in Fig. 1. Each IMU was sampled at 200Hz and embedded a gyroscope ($\pm 2000 dps$), low-g ($\pm 8g$), and high-g ($\pm 200g$) accelerometers.

B. Data Labeling

1) Direction: Direction of the gait was defined using the yaw angle of the pelvis IMU. Z-axis of the gyroscope was first processed to obtain the yaw angle [18], which was then filtered (zero-phase 4th order low-pass Butterworth filter, cut-off frequency: 1Hz). A moving-average filter was applied to further smooth the signal. Turns were detected using a

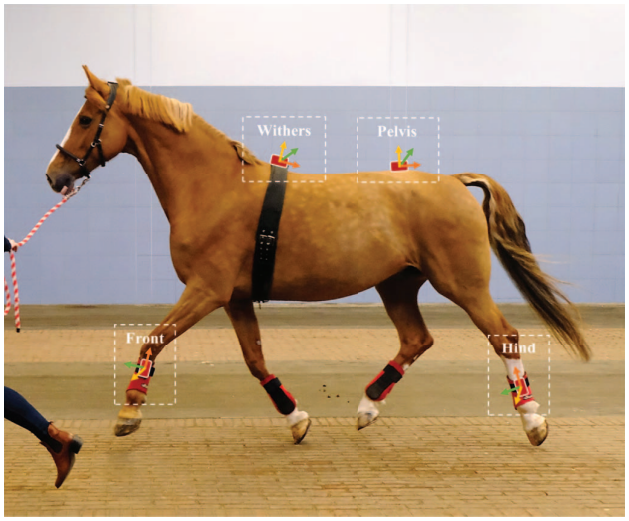

Fig. 1. Horse equipped with the gait monitoring system. Red rectangle: IMU; orange arrow: x-axis; green arrow: y-axis; yellow arrow: z-axis.

simple threshold ([-10, 10]). Positive and negative turns correspond to left and right turns, respectively. Detected turns were then classified as circles if longer than a certain duration, discarded otherwise. An example of direction labeling results is shown in Fig. 2.

2) Gait: Gait was automatically labeled based on the work from [8]. This automated labeling process requires input data from at least the pelvis and limbs IMU to provide a robust classification of the walk and trot segments used in this work.

3) Terrain: Terrains encountered were either undeformable (e.g., bricks or asphalt) or deformable (e.g., sand mix or forest dirt), as shown in Fig. 3. During data collection, terrains were manually noted alongside the timestamps. Data collected on undeformable terrains were thus labeled as "Hard" and data collected on deformable terrains as "Soft".

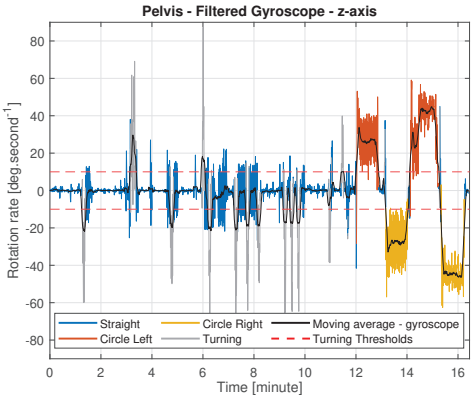

Fig. 2. Example of a trial labeled with direction.

Fig. 3. Example of Hard (left) and Soft (right) terrains used in this study.

C. Data Processing

In our previous work [12], we showed that the gyroscope data had low interest in the terrain classification task for horses data. Thus, for both the feature- and time-series-based classifiers, we used only 3D accelerometer data. For simplification purposes, as quadrupedal vertebrates show movement symmetry between limb pairs [19], we also use only the left front and left hind limb IMU data.

1) Merging Low- and High-g Accelerometer Data: For each IMU, the low-g and high-g accelerometer signals were merged to obtain a low noise and high-range signal [4]. The 3D merged acceleration (in g) was defined as:

$$Acceleration_{merged} = \begin{cases} Acceleration_{low-g}, & \text{if } |Acceleration| < 8g \\ Acceleration_{high-g} & \text{otherwise.} \end{cases} \quad (1)$$

2) Data Downsampling: To evaluate the minimal sampling frequency (Fs) required for accurate classification, we downsampled our data with different factors: 2, 4, 10, and 20; thus simulating sampling frequencies of 100Hz, 50Hz, 20Hz, and 10Hz. An example of downsampled front limb and pelvis 1D accelerometer signal is shown in Fig. 4. The MATLAB function *downsample* was used for all signals, using the resampling factors defined above.

3) Features and Time-Series Extraction: For each segment of different directions and gaits, we used a sliding window of 2 seconds with a 50% overlap. In total, 33 442 windows were extracted. For the feature-based classifiers, time and frequency domain features were extracted from each window, forming feature vectors. The different features are described in Table I. Fast Fourier Transform (FFT) coefficients' magnitude and phase were obtained with the procedure described in [20]. For the time-series-based classifiers, the 3D accelerometer signals of each used IMU were stacked to form an input matrix of (3D accelerometer signals$_{x,y,z}$ x n IMU) x (Fs x 2 seconds). Each vector and each matrix were then associated with a "Hard" or "Soft" terrain label as described in section III-B3.

4) Classifiers: We benchmarked five feature-based machine learning classifiers (Gradient Boosted Tree Ensemble (GB-Tree), Linear Discriminant Analysis (LDA), Fully Connected Neural Network (FC NN), linear kernel SVM, and

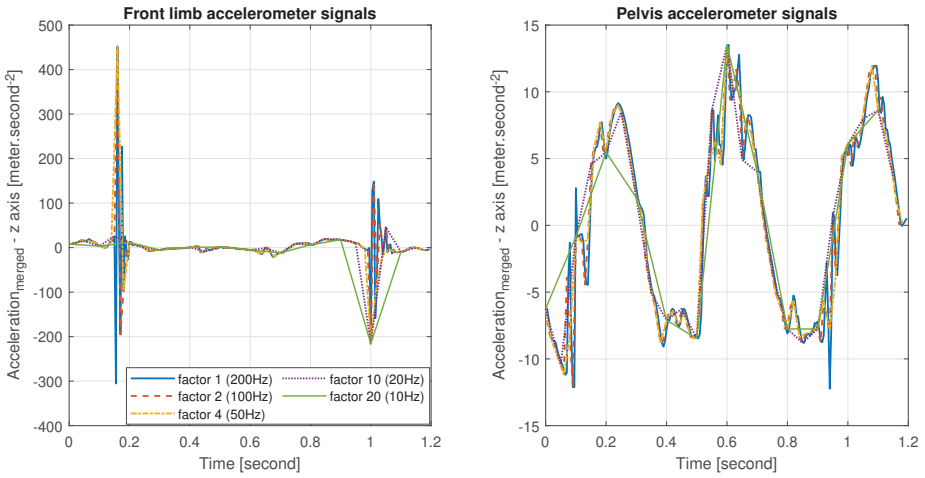

Fig. 4. Example of downsampled accelerometer signals for front limb (left) and pelvis (right) IMU.

TABLE I
DESCRIPTION OF THE EXTRACTED FEATURES.

Feature Names	Description	Domain
FFT magnitude	Six first FFT coefficient magnitudes	Freq.
FFT phase	Six first FFT coefficient phases	Freq.
Power	$\frac{1}{N}\sum_{n=1}^{N}\|x_n\|^2$	Freq.
Min	Minimum value of the window	Time
Max	Maximum value of the window	Time
Mean ($\overline{x}$)	$\frac{1}{N}\sum_{n=1}^{N} x_n$	Time
Standard Deviation (σ)	$\sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n-\overline{x})^2}$	Time
Median	Median value of the window	Time
Variance	$\frac{1}{N}\sum_{n=1}^{N}(x_n-\overline{x})^2$	Time
Skewness	$\frac{1}{N}\sum_{n=1}^{N}(\frac{x_n-\overline{x}}{\sigma})^3$	Time
Kurtosis	$\frac{1}{N}\sum_{n=1}^{N}(\frac{x_n-\overline{x}}{\sigma})^4$	Time
25^{th} & 75^{th} percentiles	25^{th} and 75^{th} of the window	Time
Gait	1 if walk, 0 if trot	Time

TABLE II
HYPERPARAMETERS USED FOR THE DIFFERENT CLASSIFIERS.

Model	Hyperparameter[a]	Value
Quad. SVM	Kernel	Quadratic
GB-Tree	Maximum Decision Splits	49
	Ensemble method	AdaBoostM1
	Learning Cycles	10
CNN	Layers architecture	input, 1D conv., reLu, layer normalization, 1D con., reLu, layer normalization, global average pooling, fully connected, softmax, classification
	Loss-function	Cross-entropy
	Filter size	3
	Number of filters	32 for the first 1D conv., 64 for the second 1D conv.
	Solver	adam
	Sequence padding	left
	Epochs	20
	Mini-batch size	one 100^{th} size of the training set
LSTM	Layers architecture	input, biLSTM, dropout, biLSTM, Fully connected, softmax, classification
	Loss-function	Cross-entropy
	Hidden units	50 per biLSTM layer
	Solver	adam
	Dropout factor	0.30
	Initial learning rate	0.02
	Epochs	20
	Mini-batch size	1/100 size of the training set

[a]If a hyperparameter is not mentioned in this table, the MATLAB default values were used.

quadratic kernel SVM (Quad. SVM)) and two time-series based deep learning classifiers (CNN and LSTM). The different models and their hyperparameters are presented in Table II. Hyperparameters were selected after preliminary experiments. For clarity purposes, the results of only Quad. SVM, GB-Tree, CNN, and LSTM will be presented in this work.

5) Datasets Definition and Classifiers Training: We defined a total of 40 datasets characterized by their input ([8 IMU sets] x [5 Fs sets]) to train and evaluate the different classifiers. The IMU sets and Fs sets are described in Table III. For each dataset, the extracted feature vectors/matrices were divided into training and testing sets, including 70% and 30% of the horses respectively. This means that feature vectors/matrices from horses used for training were not used for testing the classifiers.

The training features vectors/matrices were used to calculate the scaling coefficients (min and max) and then scale both training and test features between 0 and 1. For each dataset, the scaled training features were then ranked using a FISHER FSA [21]. Briefly, this algorithm eliminates redundant features and aims to find the combination of features that maximize the distances between data points in different classes, while minimizing the distance between data points of the same class. The first 50 features were kept to train and test the feature-based classifiers.

In order to correct the class imbalance in the training set, random feature vectors/matrices of the minor class (Soft) were duplicated.

6) Performance Evaluation: For each dataset, the random horses training-test partition was performed five times leading to five training-test iterations. For each iteration, the F1 score

was calculated for both Hard and Soft classes as a positive class, and the average performance was evaluated using the overall F1 scores between both classes, as follows:

$$F1score(i)_{Hard,\ Soft} = \left(\frac{precision(i)_{Hard,\ Soft} * recall(i)_{Hard,\ Soft}}{precision(i)_{Hard,\ Soft} + recall(i)_{Hard,\ Soft}}\right) * 100 \quad (2)$$

$$F1score(i)_{overall} = \left(\frac{F1score(i)_{Hard} + F1score(i)_{Soft}}{2}\right) \quad (3)$$

where i is the training-test iteration.

The effects of IMU location and sampling frequency on the F1 scores were then evaluated by modeling the F1 scores with linear mixed models (fixed effects: IMU location, sampling frequency; interaction effect: IMU location – sampling frequency; random factor: training iteration). Pairwise *post hoc* comparisons p-values were corrected for multiple comparisons with the *fdr* method. The linear mixed model's results for the CNN are shown in Tables IV, V and VI, in the Annex section.

7) Computational Costs Evaluation: Computational costs were evaluated in processing step durations (in seconds) for FT and TS classifiers with only one IMU (Front) and for all sampling frequencies. The data of one testing horse from iteration *i* = 5 was used (318 windows of 2s), representing the rest of the population. The durations were evaluated for each step using the *tic* and *toc* functions of MATLAB. The computer used (Windows 10 Pro 64-bit) had an Intel core i7-8650U with 32GB of RAM and a clock speed of 1.9GHz.
For FT classifiers, we quantified the time needed for the following steps:

- Data preprocessing
- Computation of the 50 selected features*
- Prediction

*In order to have fair comparisons, we kept the 50 most selected features among all FT classifiers using only Front limb IMU data, regardless of the classifier (Quad. SVM and GB-Tree) and sampling frequency. These include Gait, 3D magnitude of the 6th first FFT coefficients, 3D kurtosis and skewness values, 3D maximum and minimum values, 3D standard deviations and variances, 3D power, 3D 25th and 75th percentiles values, median and mean values for the x and z axes. For the TS classifiers, we quantified the time needed for the following steps:

- Data preprocessing
- Extraction and stacking of the selected IMU signals
- Prediction

All classifiers were also compared in size (megabytes, MB) as the selected classifier will have to be embedded in the IMU board and should be lightweight.

TABLE III
DESCRIPTION OF THE IMU AND SAMPLING FREQUENCY SETS USED FOR THE DIFFERENT DATASETS.

Set name	**Number of IMU**	**IMU location**
All	4	Withers, Pelvis, Front, Hind
W-P	2	Withers, Pelvis
P	1	Pelvis
P-F	2	Pelvis, Front
P-H	2	Pelvis, Hind
F-H	2	Front, Hind
F	1	Front
H	1	Hind
Set name	**Sampling frequency (Hz)**	**Downsampling factor**
10	10	20
20	20	10
50	50	4
100	100	2
200	200	1

IV. RESULTS

The results are presented with the median overall F1 scores values (see Eq. 3) over five training-test iterations, for each model and dataset, as stated in section III-C6. For the CNN models, statistical comparisons are also presented.

A. Effect of IMU Location

Fig. 5 shows the classification performance results for each IMU set and classifier. Using upper-body IMU sets, i.e., Withers-Pelvis (W-P) and Pelvis (P), did not yield sufficient classification performances for feature-based classifiers (median F1 scores below 70%). The CNNs and LSTMs worst performances were also obtained with W-P and P. Overall, including Limb data (All, P-F, P-H, F-H, F, H) achieved better performances than using upper-body data (W-P, P) only. Using Limb data only (F-H, F, H) had the best performance across classifiers but also reduced the variations between the training-testing iterations. For CNNs, there were no significant differences at 200Hz when training with All, P-F, F-H, F, or H (p-value$>$0.05). Training with W-P was the worst compared to All (-15.41%, t(156)=-13.87, p=.000). Training with P was also significantly lower than with All (8.09%, t(156)=-7.34, p=.000). Detailed results for CNNs and IMU locations are presented in Table V (Annex).

B. Effect of Sampling Frequency

Fig. 5 shows the classification performance results for each Fs set and classifier. Sampling frequency did not have an effect on the FT classifiers when using the upper-body IMU sets (W-P, P). Sampling frequency also did not have an influence on the CNNs trained with W-P. For all FT models and for CNNs, training with higher sampling frequencies increased the F1 scores for the other IMU sets. Overall, LSTM classifiers performances decreased when training with sampling frequencies below 100Hz. For CNNs, there were no significant differences when training with 200 and 100Hz or 200 and 50Hz with All, W-P, P-F, P-H, F-H, and F sets (p-value$>$0.05). For the F set, training with 20 and 10Hz was significantly lower than with 200Hz (-10.35%, t(156)=-6.96, p=.000 and -15.23%, t(156)=-12.81, p=.000 respectively). Detailed results for CNNs and sampling frequencies are presented in Table VI (Annex).

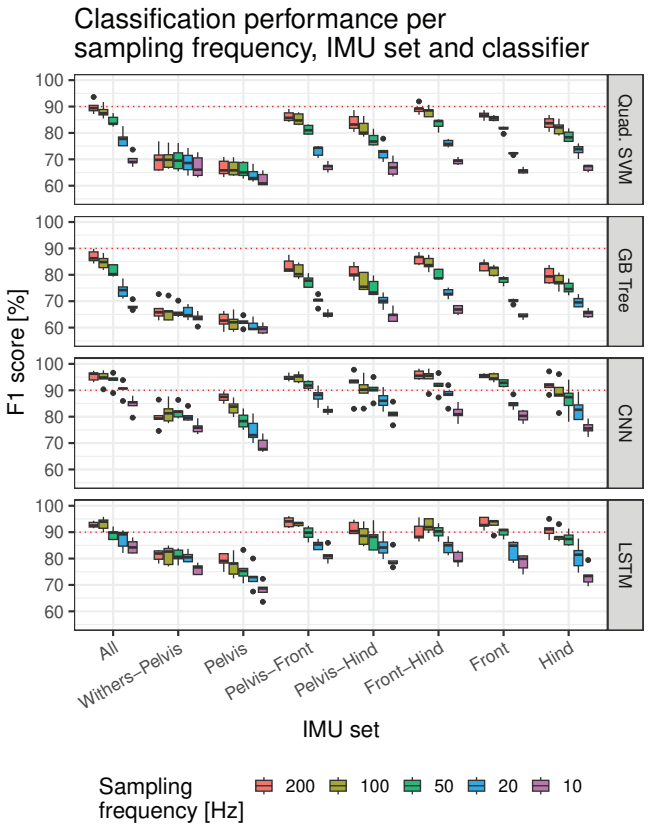

Fig. 5. Classification performance (F1 scores) per sampling frequency, IMU set, and classifier used for training. The dotted red lines represent F1 scores of 90%.

C. Feature- or Time-Series-Based Classifiers

Overall, CNNs was the most stable classifier (small variations in the performances) and had the highest F1 scores (>92%) for All, Pelvis-Front, Front-Hind and Front with sampling frequencies of 50Hz and above. Among the feature-based classifiers, the Quad. SVMs performed best with the Front-Hind IMU set, at 100Hz and 200Hz (F1 scores: 89% and 92% respectively). When using only the Pelvis IMU data, CNN classifiers were also able to reach good performances at 100Hz and 200Hz (F1: 84% and 88% respectively) while the Quad. SVMs had lower performances (F1: 71% for both). Using only Front IMU data, CNNs had a median F1 score of 93% at 50Hz and 95% at 200Hz, while Quad. SVMs had a median F1 score of 82% and 87% with the Front IMU data respectively.

D. Computational Costs Estimation

Computational costs estimation results are presented in Fig. 6. The computational costs were evaluated with the data of one testing horse from iteration $i = 5$ as described in section III-C7. The preprocessing step did not depend on the classifier type and was shorter for lower sampling frequencies (0.05s at 10Hz, 0.08-0.09s at 200Hz). The longest step was the feature

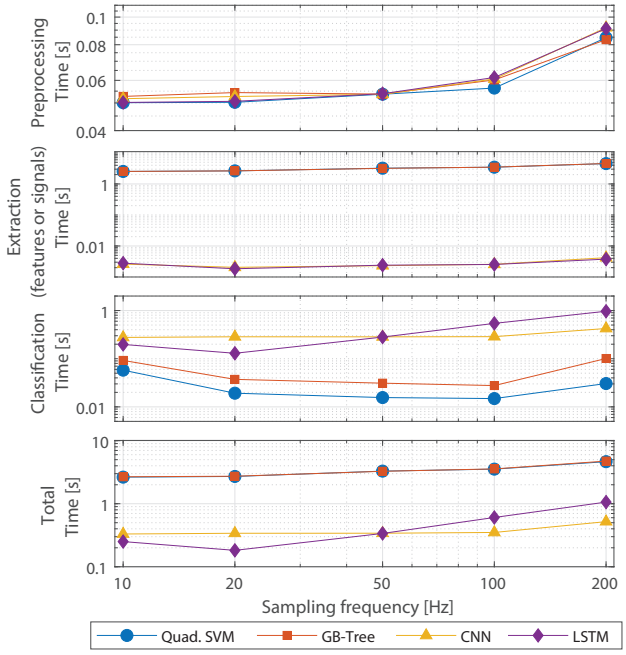

Fig. 6. Computational duration of the different steps and total duration in seconds needed for terrain classification, for each classifier and sampling frequency. Note: the x- and y- axes are in logarithmic scale.

extraction for the feature-based classifiers (Quad. SVM and GB-Tree), which could last up to 4.56s with data of 200 Hz. For CNN and LSTM, the equivalent step is the extraction of the signals, which consists in stacking the different signals and lasts up to 0.004 s. Classification durations were low for all classifiers (Quad. SVM: <0.1s, GB-Tree: <0.32s, CNN: <0.42s) but could reach up to 0.97s for the LSTM, at 200Hz. When summing up all durations, CNN was the fastest with 0.52s with 200Hz data and 0.33s with 10Hz data, a duration decrease of 33.5%, and Quad. SVM was the slowest (200Hz data: 4.75s, 10Hz data: 2.68s, thus a decrease of 43.6%).

For CNN and LSTM, the final model size is small and not influenced by the sampling frequency used when training (0.03 and 0.34MB respectively), as shown in Fig. 7. GB-Tree models' sizes slightly increased with increasing sampling frequency, with values ranging from 20.38MB to 20.50MB. Quad. SVM model' sizes decreased with increasing sampling frequencies, from 16.44MB at 10 Hz to 12.83MB at 200 Hz. This is explained by the fact that Quad. SVM classifiers trained with lower sampling frequencies compute more support vectors to separate the data, as less information is available.

V. Discussion

This study is the first one to explore different IMU sets and sampling frequencies needed for terrain classification with equine data and more broadly (living) quadruped locomotion analysis. We also classified data regardless of the gait, while other studies separated gaits or had only one gait type in

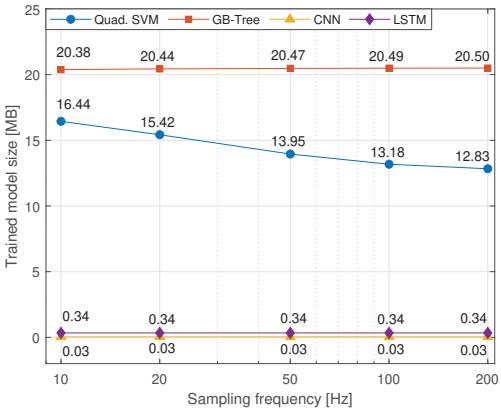

Fig. 7. Size in Megabytes (MB) of the models trained with Front limb data, using different sampling frequencies. Note: the x-axis is in logarithmic scale.

their datasets. We showed that by combining a minimal setup of IMU and machine or deep learning models, it is possible to accurately and quickly classify two terrain types encountered during lameness examinations, namely hard and soft, regardless of the gait (walk or trot).

A. IMU location

Our results show that using either all IMU (upper-body and limbs) or limbs only yielded similar classification performances when using CNNs. When using upper-body nodes only (W-P or P), our FT models were not able to classify terrains properly, which is in accordance with previous work in horses [12]. In the human literature, similar results were found as GB-trees were better when combining lower-back and tibia/shank data while CNNs were able to accurately classify using tibia/shank data only [11], [14]. However, in this work when using Pelvis data only, the CNNs performances were good but still insufficient (F1 score <90%). Terrains are known to influence how horses display lameness-related asymmetries at the pelvis at trot [6]. Our dataset also includes healthy horses which might not show sufficient differences between terrains. Moreover, at trot horses' limbs can be described with a mass-spring model [22]. Most of the impact vibrations are absorbed by the limbs, thus reducing the signals and feature differences between soft and hard terrains at the pelvis level. Our dataset also contained walk, described by the inverted pendulum model and presenting tripedal and bipedal support, further damping vibrations when compared to the unipedal support of the human running. In future works, the use of explainable artificial intelligence methods such as Grad-CAM [23] could be useful to better understand where in the signals the terrain-related differences appear.

B. Downsampling

None of the related works looked at the effect of sampling frequency on classification performances. Their sampling frequencies ranged from 75Hz [13] to 1024Hz [11]. Since softer terrains reduce the contact impact compared to hard terrains, higher sampling frequency would enable the capture of high-frequency components relevant to terrain classification tasks. We thus expected a decrease in classification performance when downsampling the signals, which is equivalent to reducing the high-frequency information related to impact. Feature-based classifiers were more affected by downsampling than signal-based classifiers, most likely because the selected features are too sensitive to the amount of information available in the signals. On the other hand, we showed that a good performance could still be obtained when using Front limb data sampled at 50Hz with CNNs and LSTMs, thus reducing the CPU (or GPU when available) usage and slightly reducing the execution duration. Our results also show that gait analysis systems using only one sensor sampling below 200Hz, attached to the pelvis such as described in [24], close to the withers as in [25] or at the sternum level [26], would not be able to accurately classify terrain types but could benefit from the addition of a limb accelerometer sensor.

C. Limitations

This study is limited by the IMU locations tested. Many studies and wearable systems available nowadays include head, sternum, and/or hoof IMU [3], [4], [27], [28], which we did not evaluate here. These sensor locations are important in equine gait analysis because they enable the quantification of compensatory movement adaptations to lameness when horses try to unload the lame limb [3], [29] but also to characterize healthy movement patterns [26], [27]. Our study also evaluated the computational costs with a computer that has different characteristics (CPU, memory architecture, etc.) than an embedded system. Different tests would then need to be run on the developed system to confirm our results. Computational costs for FT were higher mostly because of the feature extraction step, which could be optimized and lowered when using fewer features (e.g. 20 instead of 50), without decreasing the classification performance which is already lower for these models than for the time-series-based ones. However, in our previous work [12], we showed that good classification accuracies were reached when using at least 50 features from a combination of all sensor locations (mainly limbs). Concerning model weight size, merging the vector pairs close to each other can be done for SVM models [30], thus reducing the final model size.

VI. Conclusion

This study shows that by using IMU mounted to the horse's body, it is possible to detect the type of terrain regardless of the gait. For this task, both feature- and time-series-based classifiers can be used and have sufficient F1 scores when using all IMU locations, at higher sampling frequencies. Overall, the CNN time-series classifiers yielded the most accurate and fastest classification process. CNNsss are able to identify and extract robust features in the signals automatically, with high generalization abilities, which is especially interesting in our dataset that includes different gait, direction, health status, and

individual locomotion patterns. When using only one front limb IMU, CNNs are also the best classifier and have excellent performances with sampling frequencies as low as 50Hz. The developed CNN models can thus potentially be embedded on the IMU board directly, optimizing the data processing task in the framework of (smarter) equine gait analysis systems.

ACKNOWLEDGMENT

The authors would like to thank the horses' owners and caretakers that were involved in the different data collection as well as the different colleagues present for the data collections. This project was partly funded by the EFRO OP-Oost (project Paardensprong) and the Dutch Arthritis Society (Centre of Excellence Grant LLP22). The Animal Ethics Committee of Utrecht University issued ethical permissions for the different data collections when required.

REFERENCES

[1] A. Egenvall, B. N. Bonnett, P. Olson, J. Penell, and U. Emanuelson, "Association between costly veterinary-care events and 5-year survival of Swedish insured warmblooded riding horses," *Preventive Veterinary Medicine*, vol. 77, no. 1-2, pp. 122–136, 2006.

[2] M. W. Ross, "Chapter 7 - Movement," in *Diagnosis and Management of Lameness in the Horse (Second Edition)*, second edition ed., M. W. Ross and S. J. Dyson, Eds. Saint Louis: W.B. Saunders, 2011, pp. 64–80.

[3] K. G. Keegan, J. Kramer, Y. Yonezawa, H. Maki, P. F. Pai, E. V. Dent, T. E. Kellerman, D. A. Wilson, and S. K. Reed, "Assessment of repeatability of a wireless, inertial sensor–based lameness evaluation system for horses," *American Journal of Veterinary Research*, vol. 72, no. 9, pp. 1156–1163, 2011.

[4] S. Bosch, F. M. Serra Bragança, M. Marin-Perianu, R. Marin-Perianu, B. J. van der Zwaag, J. Voskamp, W. Back, R. P. van Weeren, and P. Havinga, "EquiMoves: A wireless networked inertial measurement system for objective examination of horse gait," *Sensors*, vol. 18, no. 3, 2018.

[5] T. Pfau, E. Persson-Sjodin, H. Gardner, O. Orssten, E. Hernlund, and M. Rhodin, "Effect of speed and surface type on individual rein and combined left–right circle movement asymmetry in horses on the lunge," *Frontiers in Veterinary Science*, vol. 8, 2021.

[6] E. Marunova, K. Hoenecke, A. Fiske-Jackson, R. K. W. Smith, D. M. Bolt, M. Perrier, C. Gerdes, E. Hernlund, M. Rhodin, and T. Pfau, "Changes in head, withers, and pelvis movement asymmetry in lame horses as a function of diagnostic anesthesia outcome, surface and direction," *Journal of Equine Veterinary Science*, vol. 118, p. 104136, 2022.

[7] F. M. Serra Bragança, M. Rhodin, and P. R. van Weeren, "On the brink of daily clinical application of objective gait analysis: What evidence do we have so far from studies using an induced lameness model?" *The Veterinary Journal*, vol. 234, pp. 11–23, 2018.

[8] F. M. Serra Bragança, S. Broomé, M. Rhodin, S. Björnsdóttir, V. Gunnarsson, J. P. Voskamp, E. Persson-Sjodin, W. Back, G. Lindgren, M. Novoa-Bravo, C. Roepstorff, B. J. van der Zwaag, P. R. van Weeren, and E. Hernlund, "Improving gait classification in horses by using inertial measurement unit (IMU) generated data and machine learning," *Scientific Reports*, vol. 10, no. 1, pp. 2045–2322, 2020.

[9] A. M. Hardeman, P. R. Van Weeren, F. M. Serra Bragança, H. Warmerdam, and H. G. J. Bok, "A first exploration of perceived pros and cons of quantitative gait analysis in equine clinical practice," *Equine Veterinary Education*, vol. 34, no. 10, pp. e438–e444, 2022.

[10] N. Anantrasirichai, J. Burn, and D. Bull, "Terrain classification from body-mounted cameras during human locomotion," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2249–2260, 2015.

[11] P. C. Dixon, K. H. Schütte, B. Vanwanseele, J. V. Jacobs, J. T. Dennerlein, J. M. Schiffman, P.-A. Fournier, and B. Hu, "Machine learning algorithms can classify outdoor terrain types during running using accelerometry data," *Gait Posture*, vol. 74, pp. 176–181, 2019.

[12] J. I. M. Parmentier, F. M. Serra Bragança, and B. J. van der Zwaag, "Benchmarking feature selection algorithms for optimal classification and dataset comprehension: a biomechanical application – abstracts 47th congress of the society of biomechanics," vol. 25, no. sup1. Taylor Francis, 2022, pp. S245–247.

[13] M. Z. U. H. Hashmi, Q. Riaz, M. Hussain, and M. Shahzad, "What lies beneath one's feet? Terrain classification using inertial data of human walk." *Applied Sciences*, vol. 9, no. 15, 2019.

[14] B. Hu, S. Li, Y. Chen, R. Kavi, and S. Coppola, "Applying deep neural networks and inertial measurement unit in recognizing irregular walking differences in the real world," *Applied Ergonomics*, vol. 96, p. 103414, 2021.

[15] V. Shah, M. W. Flood, B. Grimm, and P. C. Dixon, "Generalizability of deep learning models for predicting outdoor irregular walking surfaces," *Journal of Biomechanics*, vol. 139, p. 111159, 2022.

[16] T. J. P. Spoormakers, L. St. George, I. H. Smit, S. J. Hobbs, H. Brommer, H. M. Clayton, S. H. Roy, J. Richards, and F. M. Serra Bragança, "Adaptations in equine axial movement and muscle activity occur during induced fore- and hindlimb lameness: A kinematic and electromyographic evaluation during in-hand trot," *Equine Veterinary Journal*, vol. n/a, no. n/a.

[17] H. Darbandi, C. Munsters, J. Parmentier, and P. Havinga, "Detecting fatigue of sport horses with biomechanical gait features using inertial sensors," *PLOS ONE*, vol. 18, no. 4, pp. 1–19, 04 2023.

[18] E. G. Valenti, I. Dryanovski, and J. Xiao, "Keeping a good attitude: A quaternion-based orientation filter for IMUs and MARGs," *Sensors*, vol. 15, no. 8, pp. 19 302–19 330, 2015.

[19] A. Abourachid, "A new way of analysing symmetrical and asymmetrical gaits in quadrupeds," *Comptes Rendus Biologies*, vol. 326, no. 7, pp. 625–630, 2003.

[20] C. Roepstorff, M. T. Dittmann, S. Arpagaus, F. M. Serra Bragança, A. Hardeman, E. Persson-Sjödin, L. Roepstorff, A. I. Gmel, and M. A. Weishaupt, "Reliable and clinically applicable gait event classification using upper body motion in walking and trotting horses," *Journal of Biomechanics*, vol. 114, p. 110146, 2021.

[21] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection." in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'11. Barcelona, Spain: AUAI Press, 2011, p. 266–273.

[22] A. M. Wilson, M. P. McGuigan, A. Su, and A. J. van den Bogert, "Horses damp the spring in their step," *Nature*, vol. 414, pp. 895–899, 2001.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2 2020.

[24] E. Marunova, L. Dod, S. Witte, and T. Pfau, "Smartphone-based pelvic movement asymmetry measures for clinical decision making in equine lameness assessment," *Animals*, vol. 11, no. 6, 2021.

[25] A. Schmutz, L. Chèze, J. Jacques, and P. Martin, "A method to estimate horse speed per stride from one IMU with a machine learning method," *Sensors*, vol. 20, no. 2, 2020.

[26] E. Barrey, M. Hermelin, J. L. Vaudelin, D. Poirel, and V. J.P., "Utilisation of an accelerometric device in equine gait analysis," *Equine Veterinary Journal*, vol. 26, no. S17, pp. 7–12, 1994.

[27] H. Darbandi, F. S. Bragança, B. J. van der Zwaag, and P. Havinga, "Accurate horse gait event estimation using an inertial sensor mounted on different body locations," in *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2022, pp. 329–335.

[28] M. Tijssen, E. Hernlund, M. Rhodin, S. Bosch, J. P. Voskamp, M. Nielen, and F. M. Serra Bragança, "Automatic hoof-on and -off detection in horses using hoof-mounted inertial measurement unit sensors," *PLoS ONE*, vol. 15, no. 6, 2020.

[29] M. Rhodin, E. Persson-Sjodin, A. Egenvall, F. M. Serra Bragança, T. Pfau, L. Roepstorff, M. A. Weishaupt, M. H. Thomsen, P. R. van Weeren, and E. Hernlund, "Vertical movement symmetry of the withers in horses with induced forelimb and hindlimb lameness at trot," *Equine Veterinary Journal*, vol. 50, no. 6, pp. 818–824, 2018.

[30] D. Nguyen and T. Ho, "An efficient method for simplifying support vector machines," in *Proceedings of the 22nd International Conference on Machine Learning*, ser. ICML '05. Bonn, Germany: Association for Computing Machinery, 2005, p. 617–624.

ANNEXES

TABLE IV
LINEAR MIXED MODEL RESULTS FOR CNN CLASSIFIER PERFORMANCES

IMU Set	Fs (Hz)	emmean	lower CI	upper CI
All	200	95.41	91.52	99.30
	100	94.67	90.78	98.57
	50	93.67	89.77	97.58
	20	90.38	86.44	94.31
	10	84.58	80.60	88.57
Withers-Pelvis	200	80.00	75.97	84.03
	100	81.60	77.59	85.61
	50	81.93	77.92	85.94
	20	80.22	76.20	84.25
	10	75.85	71.77	79.92
Pelvis	200	87.32	83.36	91.28
	100	83.44	79.45	87.44
	50	78.61	74.57	82.66
	20	74.80	70.71	78.89
	10	69.30	65.14	73.46
Pelvis-Front	200	94.74	90.85	98.64
	100	94.80	90.90	98.69
	50	91.85	87.93	95.77
	20	87.84	83.89	91.80
	10	82.25	78.24	86.25
Pelvis-Hind	200	92.24	88.33	96.16
	100	90.25	86.32	94.19
	50	90.26	86.33	94.19
	20	86.30	82.33	90.27
	10	81.11	77.09	85.13
Front-Hind	200	95.84	91.95	99.73
	100	94.62	90.72	98.52
	50	91.97	88.05	95.89
	20	88.20	84.25	92.15
	10	81.40	77.39	85.42
Front	200	95.43	91.54	99.32
	100	94.85	90.95	98.75
	50	92.68	88.76	96.59
	20	85.08	81.10	89.06
	10	80.20	76.17	84.23
Hind	200	92.19	88.27	96.11
	100	89.01	85.07	92.96
	50	86.62	82.66	90.59
	20	82.92	78.92	86.91
	10	75.75	71.67	79.82

Fs: sampling frequency; CI: Confidence Intervals

TABLE V
PAIRWISE COMPARISONS – IMU SETS FOR CNN CLASSIFIER PERFORMANCES

	200Hz			100Hz			50Hz		
IMU set	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value
All – Withers-Pelvis	15.41	1.14	***	13.07	1.13	***	11.75	1.13	***
All – Pelvis	8.09	1.10	***	11.23	1.12	***	15.06	1.15	***
All – Pelvis-Front	0.66	1.07	1.000	-0.12	1.08	1.000	1.82	1.09	1.000
All – Pelvis-Hind	3.17	1.08	0.588	4.42	1.09	*	3.41	1.10	0.436
All – Front-Hind	-0.43	1.07	1.000	0.05	1.08	1.000	1.70	1.09	1.000
All – Front	-0.02	1.07	1.000	-0.18	1.08	1.000	1.00	1.09	1.000
All – Hind	3.22	1.08	0.549	5.66	1.10	***	7.05	1.11	***
Withers-Pelvis – Pelvis	-7.32	1.16	***	-1.84	1.17	1.000	3.32	1.19	0.701
Withers-Pelvis – Pelvis-Front	-14.74	1.14	***	-13.20	1.13	***	-9.92	1.14	***
Withers-Pelvis – Pelvis-Hind	-12.24	1.15	***	-8.65	1.15	***	-8.33	1.14	***
Withers-Pelvis – Front-Hind	-15.84	1.13	***	-13.02	1.13	***	-10.04	1.14	***
Withers-Pelvis – Front	-15.43	1.14	***	-13.25	1.13	***	-10.75	1.14	***
Withers-Pelvis – Hind	-12.19	1.15	***	-7.41	1.15	***	-4.69	1.16	*
Pelvis – Pelvis-Front	-7.43	1.10	***	-11.36	1.12	***	-13.24	1.15	***
Pelvis – Pelvis-Hind	-4.92	1.11	*	-6.81	1.14	***	-11.65	1.16	***
Pelvis – Front-Hind	-8.52	1.10	***	-11.18	1.12	***	-13.36	1.15	***
Pelvis – Front	-8.11	1.10	***	-11.41	1.12	***	-14.06	1.15	***
Pelvis – Hind	-4.87	1.11	*	-5.57	1.14	**	-8.01	1.17	***
Pelvis-Front – Pelvis-Hind	2.50	1.09	0.951	4.54	1.09	*	1.59	1.10	1.000
Pelvis-Front – Front-Hind	-1.10	1.07	1.000	0.18	1.08	1.000	-0.12	1.10	1.000
Pelvis-Front – Front	-0.69	1.07	1.000	-0.05	1.08	1.000	-0.83	1.09	1.000
Pelvis-Front – Hind	2.55	1.09	0.937	5.78	1.10	***	5.23	1.12	**
Pelvis-Hind – Front-Hind	-3.60	1.08	0.286	-4.37	1.09	*	-1.71	1.10	1.000
Pelvis-Hind – Front	-3.19	1.08	0.571	-4.60	1.09	*	-2.42	1.10	0.974
Pelvis-Hind – Hind	0.05	1.09	1.000	1.24	1.11	1.000	3.64	1.12	0.343
Front-Hind – Front	0.41	1.07	1.000	-0.23	1.08	1.000	-0.71	1.09	1.000
Front-Hind – Hind	3.65	1.08	0.257	5.61	1.10	***	5.35	1.12	**
Front – Hind	3.24	1.08	0.532	5.84	1.10	***	6.06	1.12	***

p-value $<$ 0.001: *** ; p-value $<$ 0.01: ** ; p-value $<$ 0.05: *
SE: standard error

TABLE VI
PAIRWISE COMPARISONS – SAMPLING FREQUENCY (HZ) FOR CNN CLASSIFIER PERFORMANCES

IMU set	All			Withers-Pelvis			Pelvis			Pelvis-Front			Pelvis-Hind			Front-Hind			Front			Hind		
Sampling frequency	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value	estimate	SE	p-value
200 - 100	0.73	1.07	1.000	-1.60	1.19	1.000	3.88	1.15	0.256	-0.05	1.08	1.000	1.99	1.10	0.999	1.22	1.07	1.000	0.58	1.07	1.000	3.18	1.11	0.631
200 - 50	1.73	1.08	1.000	-1.93	1.19	1.000	8.71	1.17	***	2.89	1.09	0.785	1.98	1.10	0.999	3.87	1.08	0.158	2.75	1.08	0.855	5.57	1.12	**
200 - 20	5.03	1.09	**	-0.22	1.19	1.000	12.52	1.19	***	6.90	1.10	***	5.94	1.12	***	7.64	1.10	***	10.35	1.11	***	9.28	1.13	***
200 - 10	10.83	1.11	***	4.15	1.22	0.234	18.02	1.22	***	12.50	1.13	***	11.13	1.14	***	14.44	1.13	***	15.23	1.13	***	16.44	1.17	***
100 - 50	1.00	1.08	1.000	-0.33	1.18	1.000	4.83	1.19	*	2.95	1.09	0.751	-0.01	1.11	1.000	2.65	1.09	0.907	2.17	1.08	0.993	2.39	1.13	0.984
100 - 20	4.30	1.09	0.056	1.38	1.19	1.000	8.64	1.21	***	6.95	1.10	***	3.95	1.13	0.184	6.42	1.10	***	9.77	1.11	***	6.10	1.14	***
100 - 10	10.09	1.12	***	5.75	1.21	**	14.14	1.24	***	12.55	1.13	***	9.14	1.15	***	13.22	1.13	***	14.65	1.14	***	13.27	1.18	***
50 - 20	3.30	1.10	0.518	1.71	1.19	1.000	3.81	1.23	0.441	9.60	1.14	***	3.96	1.13	0.181	3.77	1.11	0.246	7.60	1.12	***	3.71	1.15	0.362
50 - 10	9.09	1.12	***	6.08	1.21	***	9.31	1.26	***	4.01	1.11	0.147	9.15	1.15	***	10.57	1.14	***	12.48	1.14	***	10.88	1.19	***
20 - 10	5.79	1.13	***	4.37	1.22	0.147	5.50	1.28	*	5.60	1.15	**	5.19	1.16	**	6.80	1.15	***	4.88	1.17	*	7.17	1.20	***

p-value $<$ 0.001: *** ; p-value $<$ 0.01: ** ; p-value $<$ 0.05: *
SE: standard error