Eurasia Specialized Veterinary Publication

International Journal of Veterinary Research and Allied Science

ISSN:3062-357X

2023, Volume 3, Issue 1, Page No: 139-156

Available online at: www.esvpub.com/

Smartphone-Based Gait Classification of Five-Gaited Horses Using Deep Learning: Efficient Data Acquisition and High-Accuracy Modeling

Alejandro Torres[1]*, Felipe Ramos[1]

[1]Department of Animal Epidemiology, Faculty of Veterinary and Animal Sciences, University of Chile, Santiago, Chile.

***E-mail** ✉ a.torres.epi@outlook.com

ABSTRACT

Traditionally, horse gait classification has depended on sensors attached to the horse itself. Mobile phones offer a more practical alternative, but the effectiveness of gait models based on such sensors has been underexplored. In this study, we applied deep learning to classify horse gaits using data from smartphones carried by riders. Seventeen horses and fourteen riders participated, with data collected simultaneously from the rider's phone and a four-sensor horse-mounted system. Using this approach to generate labeled data efficiently, we trained a Bi-LSTM model that relied solely on 50 Hz accelerometer and gyroscope signals aligned to the horse's frame of reference. The model successfully distinguished the five gaits of Icelandic horses with 94.4% accuracy. These results suggest that mobile phones can facilitate large-scale monitoring of horse movement. Future research should examine how factors such as rider style, equipment, and phone placement influence classification accuracy, which will further enhance understanding of equine gait and its applications in equestrian activities.

Keywords: Horse, Smartphone sensors, Gait analysis, Inertial measurement unit, Machine learning

Received: 19 January 2023
Revised: 28 April 2023
Accepted: 03 May 2023

How to Cite This Article: Torres A, Ramos F. Smartphone-Based Gait Classification of Five-Gaited Horses Using Deep Learning: Efficient Data Acquisition and High-Accuracy Modeling. Int J Vet Res Allied Sci. 2023;3(1):139-56.
https://doi.org/10.51847/F1h2kR5qRf

Introduction

Smartphones are now ubiquitous in daily life, and their capabilities are expanding rapidly as sensor technology improves. Beyond tracking human activities, these devices can also monitor animal movements. One practical application is equine gait recognition, as seen in apps like Equilab (https://equilab.horse, accessed 1 May 2022), which identifies four gaits: walk, trot, canter, and tölt. This study was conducted in partnership with Horseday ehf., aiming to extend gait recognition to the five-gaited Icelandic horse, including flying pace.

The scientific study of horse locomotion dates back to Hildebrand's influential 1965 work [1], which established a framework for analyzing limb movements and stride patterns. Since then, equine-specific motion sensors have been widely used to collect precise data on gait events. Such data have applications in predicting hoof contact timing [2,3], assessing lameness [4], monitoring performance in dressage and show jumping [5], and evaluating fatigue during training [6]. Machine learning techniques using horse-mounted sensors have achieved high classification accuracy, with some models reaching up to 97% [7,8]. Moreover, research has shown that smartphone sensors can produce measurements comparable to dedicated inertial measurement units when attached to horses [9].

Recent studies highlight that deep learning models, such as LSTM networks, can process raw sensor signals with minimal preprocessing and still approach the accuracy of models based on traditional gait variables [8,10]. Convolutional neural networks have also been applied successfully to raw accelerometer data from horse-mounted sensors [11]. Despite these advances, a key limitation remains: riders must attach multiple sensors to the horse, which can be cumbersome. This motivates exploring whether data collected from a smartphone carried by the rider could achieve reliable gait classification. Prior research on three-gaited horses (walk, trot, canter) suggests this approach is feasible [12,13], but its effectiveness for five-gaited Icelandic horses has yet to be established.

In this work, we examined the potential of using rider-carried smartphones to classify all five gaits of the Icelandic horse. Unlike most horse breeds, the Icelandic horse is capable of two additional gaits—tölt and flying pace—on top of walk, trot, and canter, due to a genetic variation [14]. To obtain reliable labeled data for model training, we utilized the TöltSense system (TS, https://toltsense.com, accessed 1 May 2022), which automatically identifies horse gaits. Traditional approaches to gait labeling have relied on video recordings under controlled conditions [8] or manually timing gait transitions with a stopwatch [13], both of which are labor-intensive and less adaptable to natural riding settings. TS simplifies this process by providing objective, automated labels, making data collection more scalable and suitable for real-world environments.

The study was designed to answer a central question: how accurately can the five gaits of Icelandic horses be classified using machine learning models trained on data from smartphones carried by riders? Our objectives were twofold: first, to assess the labeling accuracy of the TS system itself, and second, to evaluate the performance of models trained on rotated smartphone sensor data paired with TS labels. We hypothesized that this approach would allow the models to classify all five gaits with high reliability.

Materials and Methods

The experimental protocol for both the smartphone sensor models and the TS validation was reviewed by the local Ethics Committee (The Icelandic Food and Veterinary Authority and the Ethics Review Board at the Royal College of Veterinary Surgeons), which determined that formal ethical approval was not required. The studies were considered outside the scope of European Directive 2010/63/EU, as no procedures were performed that could cause pain, distress, or lasting harm to the horses. All procedures adhered to established guidelines and regulations. Written informed consent was obtained from all horse owners and riders as appropriate, including consent for publication from the rider depicted in **Figure 1**.

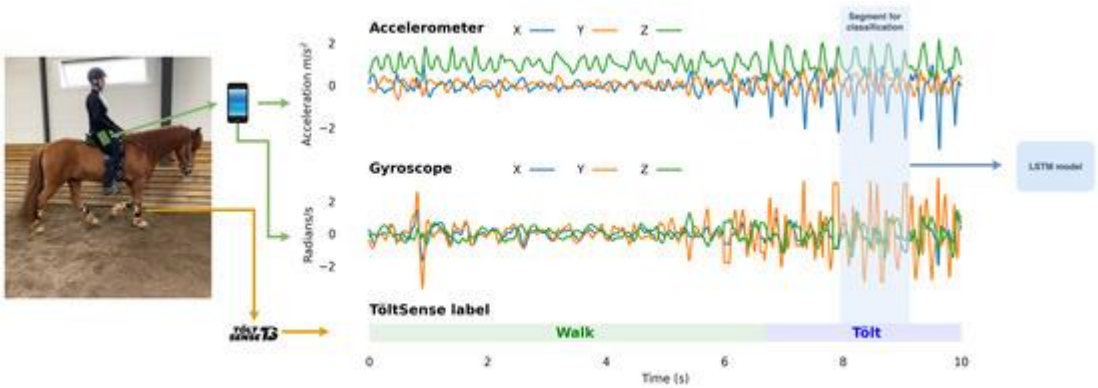

Figure 1. Gait labeling using the TöltSense system

The left panel illustrates the placement of the sensors: the TöltSense units are attached to the lateral sides of the horse's metacarpal and metatarsal bones, while the smartphone is carried in the rider's pocket. The right panel displays the smartphone sensor data after rotation into a defined world-frame coordinate system, alongside gait labels as the horse transitions from walk to tölt. The X and Y curves represent horizontal-plane variations in accelerometer and gyroscope readings, while the Z curve captures vertical-axis changes. The sign of each axis indicates the direction of the signal. The highlighted portion demonstrates a sample input used for training the Bi-LSTM model.

The töltsense system

For labeling our training dataset, we relied on the TöltSense system (TS), a tool developed to classify and assess Icelandic horse gaits, providing riders with real-time feedback. The system consists of four wireless motion sensors attached to the horse's lower limbs, synchronized to within 8 ms, and a cross-platform mobile application that processes the signals and outputs gait labels **(Figure 1)**. The TS can achieve labeling accuracy of up to 99.7% (see Results section). Unlike machine learning approaches, TS operates using a rule-based algorithm grounded in the gait definitions established by the International Federation of Icelandic Horse Associations (FEIF) [15]. By accurately measuring hoof-on and hoof-off timings, the system can determine gaits with high reliability.

Dataset and labeling—töltsense validation

We collected data from eight Icelandic horses of varying skill levels at a horse farm in the U.K. The horses were equipped with TS sensors and recorded during both training and warm-up sessions for an indoor oval track competition. Each session began with the rider pressing the TS app's "START" button **(Figure 2)**, which generated a log file containing gait classifications and timestamps. These timestamps provided a reference to synchronize video recordings with the TS data. All videos were trimmed to align precisely with the "START" time, ensuring correspondence between the video frames and the app's log entries.

Figure 2. TöltSense interface and an example video frame showing the START button press used to mark the beginning of data recording

A panel of four certified Icelandic sport judges independently assigned gait labels while reviewing the video footage using a custom web application **(Figure 3)**. For each video, a continuous observation segment of approximately 5 minutes was selected to capture multiple gait transitions, ensuring judges focused only on relevant sequences and were not required to watch unnecessary portions of the recordings. At no point were the TS system's classifications disclosed to the judges during this process.

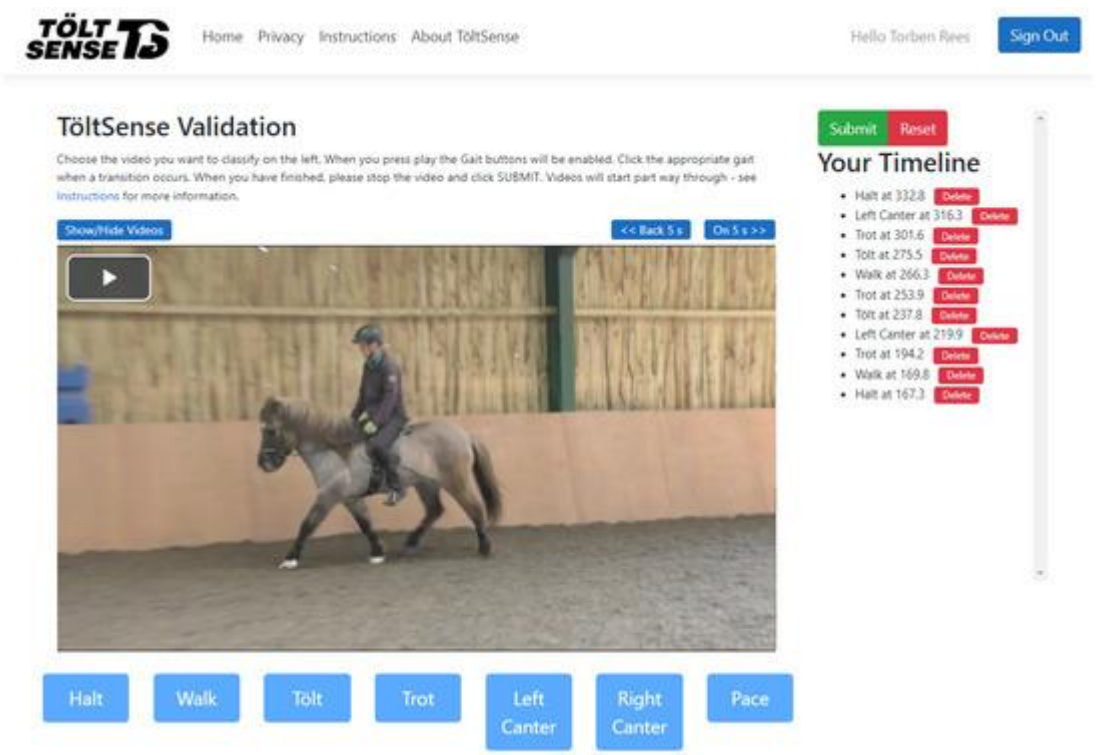

Figure 3. Web application used by judges to assign gait labels to video segments

Gaits were annotated at 250 ms intervals by selecting the label most frequently suggested by the judges within each interval. This majority-vote approach was applied consistently across the entire observation period, with a new data point generated whenever the prevailing gait changed. The resulting sequence forms a continuous time series representing the gait at every moment during the observation window, which we refer to as the **aggregate judge classifications**. An example of this time series for a single horse is presented in **Figure 4**, while **Figure 5** shows the horse's speed alongside the corresponding TS gait labels. The time series values correspond to gait categories as defined in **Table 1**.

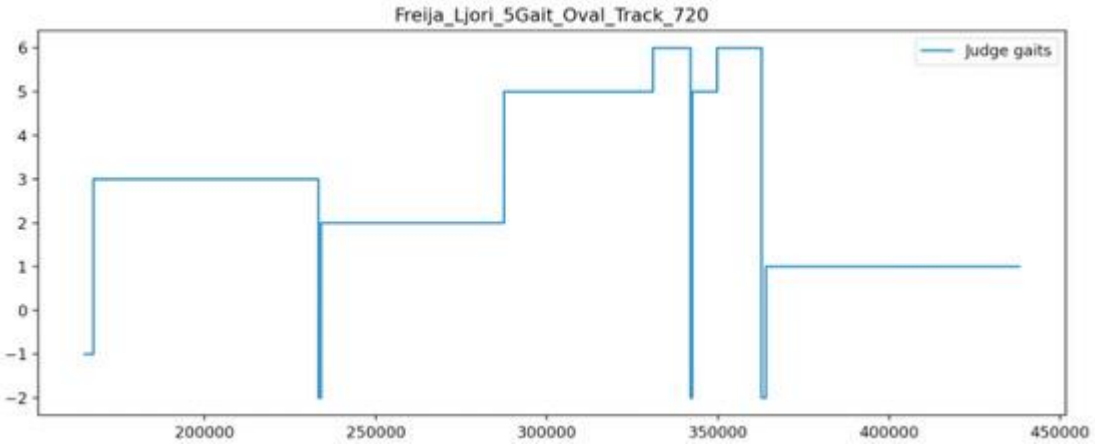

Figure 4. Plot of aggregate judge classifications against time, showing initial unclassified period (−1) and three periods of dispute (−2). The y-axis corresponds to the chosen label as defined in Table 1

Figure 5. Plot of speed vs. time colour coded by gait from the TöltSense session overview screen

Table 1. Gait labels used by judges. The table shows the gait labels from the TöltSense system that the judges used to label the video segments

TöltSense Label	Number
No majority/disputed	−2
Not classified	−1
Halt	0
Walk	1
Trot	2
Tölt	3
Left Canter	4
Right Canter	5
Pace	6

The label "No majority/disputed" was used to account for moments when the judges did not reach a consensus on the horse's gait, which mostly occurred briefly around gait transitions. Icelandic horses sometimes perform movements that are intermediate between gaits, such as a pacey tölt or a 4-beat trot, making it unclear which gait is being displayed. In such instances, where qualified observers could not agree, there is no definitive ground truth for comparison with the TS system, so periods labeled as −2 were excluded from the evaluation. The label "Not classified" appears only at the very start of each session, before any gait has been identified, and these periods were also excluded. All other gait labels are self-explanatory; note that left and right canter are treated as distinct categories to verify that TS can differentiate between them.

The TS generates a gait label each time a hoof-on event is detected from any leg, resulting in a variable labeling rate between approximately 4 Hz and 10 Hz, depending on the horse's activity. To create a uniform timeline for analysis, a sliding window of 1 second with a 0.5-second step was applied. Within each window, all gait labels were collected, and the most frequent label was assigned to the midpoint timestamp of the window. By subtracting the timestamp of the first log entry from all subsequent windows, we obtained a continuous timeline starting at 0 seconds with 0.5-second increments for the entire session.

For analysis, the TS-derived time series is treated as the predicted values, while the aggregated judge classifications serve as the ground truth. Each observation window contains both a TS prediction and the corresponding judge-generated label. At each predicted timestamp, the matching ground truth value was retrieved to form a prediction–truth pair, referred to as a "test case."

Test cases were discarded in two scenarios: if the aggregated judge classification was −2 or −1, or if the timestamp was too close to a gait transition identified by either TS or the judges. Excluding these transition-adjacent cases is necessary because delays are inherent in both human and sensor-based labeling. Judges may recognize a gait change slightly later than it occurs, and the process of clicking a button in the web application introduces further delay. Conversely, the TS system may register brief or subtle gait changes that judges cannot react to quickly, potentially generating false positives or negatives. Additionally, TS may require multiple strides to confirm slower gaits, such as walk, whereas judges can often identify the transition immediately.

To account for these timing discrepancies, a 1-second exclusion window was applied around all transitions identified by either the TS system or the judges. Any test cases occurring within these windows were removed from the analysis to ensure a fair comparison between predicted and true values.

Dataset and labeling—mobile phone-based gait classification

Between May and August 2021, we collected smartphone sensor data from riders of Icelandic horses, using the TöltSense system (TS, https://toltsense.com/, accessed 1 May 2022) to provide gait labels **(Figure 1)**. Recordings were conducted at a horse farm in southern England as well as several facilities in Iceland. The study included seventeen horses and fourteen riders. Each rider carried a smartphone in a pocket of their choice—either on clothing or jackets—introducing natural variability in sensor placement. Horses were ridden across diverse environments, including outdoor tracks, sandy arenas, and natural trails, resulting in a total of approximately 5.8 hours of labeled data and thousands of short gait segments.

The TS sensors were attached to the lateral sides of the horse's lower limbs (metacarpal or metatarsal regions) and operated at 125 Hz, with an acceleration range of ±16 g and angular velocity up to 2000 deg/s. The system ensured precise synchronization between sensors and handled both processing and logging of smartphone data. Gait labeling was based on calculated hoof-on and hoof-off events, with the TS algorithm extracting stride parameters to determine the gait. The system provides ten distinct gait labels **(Table 2)**, which were the only TS outputs used for this study. Labels were generated whenever a hoof contacted the ground, reaching up to ten events per second during fast tölt, roughly corresponding to four labels per stride at rates up to 2.5 strides per second.

Table 2. Mapping from the TöltSense labels to the ones used in this study.

TöltSense Label	Label
Standing	Not Used
Walk	Walk
Trot	Trot
Tölt	Tölt
L Canter	Canter
R Canter	Canter
L Cross Canter	Canter
R Cross Canter	Canter
Flying Pace	Flying Pace

The smartphones used in this study included various Samsung models and iPhones. iPhone data were recorded at a consistent 50 Hz sampling rate, while Samsung devices captured data at higher rates ranging from 100 to 1200 Hz depending on the specific model. The collected data consisted of accelerometer and gyroscope measurements, which were subsequently rotated into specific frames of reference. One rotation aligned the signals to the world-frame using a quaternion provided by the phone. This standard rotation method, recommended in prior equine gait studies [9,16], isolates vertical acceleration into a single dimension, making the signal easier to interpret. However, the world-frame does not fully represent lateral movements relative to the horse, so we also performed a rotation into the horse's frame of reference.

To define the horse's running direction in the x-y plane, we used the 1 Hz GPS signal by comparing consecutive longitude and latitude points, resulting in an angle within [0, 360) degrees. This directional signal was smoothed over a 1-second window to reduce noise. Circular averaging was applied by first unwrapping the signal—adjusting values with absolute differences greater than 180° to their period-complementary values—then wrapping the averaged signal back to the [0, 360) range. Horse direction and speed were obtained using the Android and iOS location APIs.

For model input, six dimensions were used: acceleration along the x-, y-, and z-axes, and angular velocity along the same axes, expressed in the chosen frame of reference. Additional experiments explored alternative inputs, such as including speed as an extra feature or using only accelerometer or gyroscope data.

Preprocessing of smartphone sensor data

Data preprocessing was performed using Python libraries including NumPy, Pandas, and PyTorch. Recordings sampled above 50 Hz were downsampled to 50 Hz, which previous studies indicate has minimal effect on gait classification performance [11] or vertical movement symmetry measures during trot [17].

To evaluate model generalization, the dataset was divided into training and test sets. The test set contained data from four horses not included in training, as well as a horse–rider combination unseen during training **(Table 3)**. The continuous recordings were split into overlapping segments, illustrated by the blue rectangles in **Figure 1**. Segments overlapped by 90% with the previous one, ensuring a proportional rather than absolute shift, which

helps reduce overfitting for longer time windows. Flying pace data were relatively scarce, so for this gait, segments were generated every 20 milliseconds regardless of window length to increase representation in the training data.

Table 3. Overview of the horses used for the study. The total time of labelled data sums up to 5.4 h. Note that only the horses in the 2nd, 10th, and 11th row are five-gaited

Horse No.	Location	Rides	Canter	Flying Pace	Walk	Trot	Tölt	Total Time
1	England	1	190 s	0 s	1026 s	212 s	376 s	30 min
2	England	2	39 s	85 s	1488 s	98 s	253 s	33 min
3	England	2	154 s	0 s	767 s	427 s	973 s	39 min
4	Iceland	3	42 s	0 s	470 s	416 s	935 s	31 min
5	Iceland	2	0 s	0 s	394 s	121 s	710 s	20 min
6	Iceland	4	88 s	0 s	1394 s	628 s	2275 s	73 min
7	Iceland	2	257 s	0 s	1319 s	712 s	1161 s	57 min
8	Iceland	1	63 s	0 s	354 s	192 s	496 s	18 min
9	Iceland	1	32 s	0 s	157 s	120 s	212 s	8 min
10	Iceland	2	0 s	28 s	0 s	0 s	0 s	0.5 min
11	Iceland	1	0 s	27 s	0 s	0 s	0 s	0.5 min
12	England	1	40 s	0 s	155 s	55 s	67 s	5 min
13	England	1	0 s	0 s	144 s	0 s	338 s	8 min
14	England	1	102 s	0 s	141 s	90 s	144 s	8 min
15	England	1	14 s	0 s	124 s	61 s	100 s	5 min
16	England	1	24 s	0 s	117 s	41 s	80 s	4 min
17	England	1	38 s	0 s	134 s	80 s	134 s	6 min

As with the TS validation, any segments occurring within 2 seconds of a gait transition—defined as a change in the TöltSense labels—were excluded from analysis. This approach removed periods where the horse was transitioning between gaits, which may not have reliable labels, and helped reduce transient labeling errors caused by slight delays in TS classification.

The ten labels generated by the TS system were consolidated into five categories corresponding to the Icelandic horse's five gaits **(Table 2)**. Cross canter, which typically results from rider error during transitions to canter or flying pace, was mapped to canter. Although pace horses sometimes wear protective over-reach boots due to increased risk of injury from cross canter, no cross-canter instances were present in the collected training data. The distribution of gaits within the training set is illustrated in **Figure 6**.

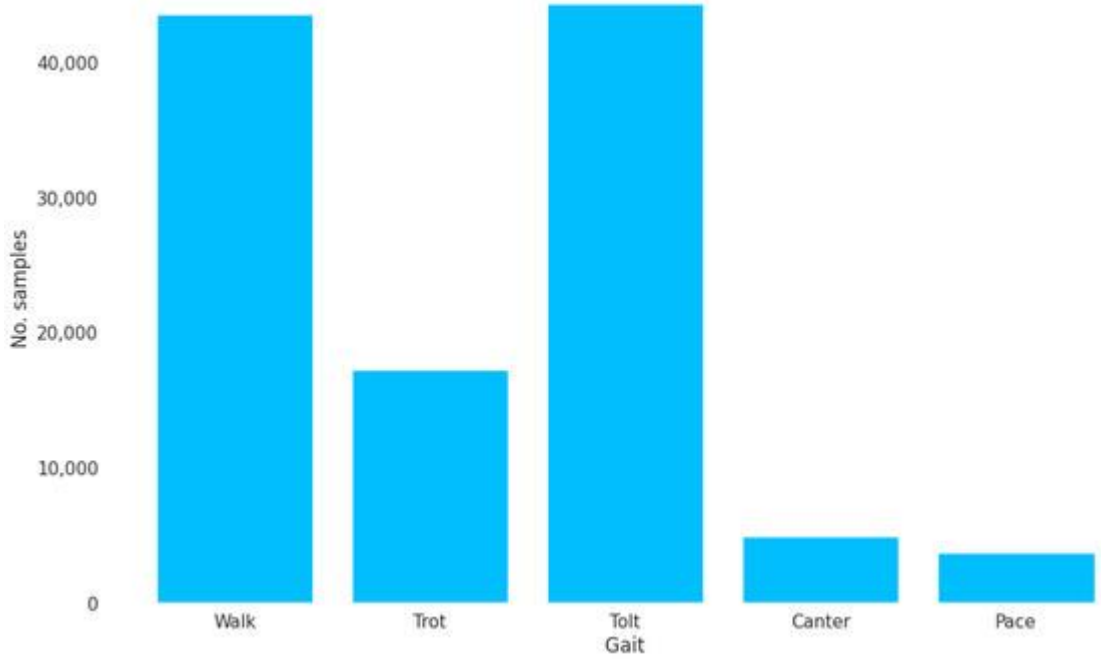

Figure 6. Proportion of each gait in the dataset based on 1.5-second segments

The dataset used for training was further divided into a training subset (85%) and a validation subset (15%), with the order of samples randomized during training. The leave-one-out test set consisted of rides from Horses 1, 2, 8, 9, and 11. Although rides from Subject 2 appeared in both training and test sets, these involved different riders and distinct smartphone models, ensuring independence between the sets.

Gait classification model

For this study, the primary approach to gait classification was a recurrent neural network (RNN), specifically a long short-term memory (LSTM) model, which is commonly used for analyzing sequential data [10]. Additional architectures were also explored, including gated recurrent units (GRU) [18], bidirectional LSTM (Bi-LSTM) [19], and a one-dimensional convolutional neural network (1D-CNN).

The LSTM model employed a single layer with 200 hidden units, while the Bi-LSTM used 400 units, and the GRU model used 200 units. The 1D-CNN consisted of four layers with kernel size 3 and dilation rates of 12, 8, 4, and 1. The number of feature maps in each layer increased progressively from ni (the number of input features) to 16, 32, 64, and finally 128. Each model concluded with a fully connected linear layer of 128 dimensions **(Figure 7)**.

All models were trained for 20 epochs with a batch size of 64 using the ADAM optimizer [20]. Training employed the cross-entropy loss function, and early stopping was applied based on the validation loss to prevent overfitting.

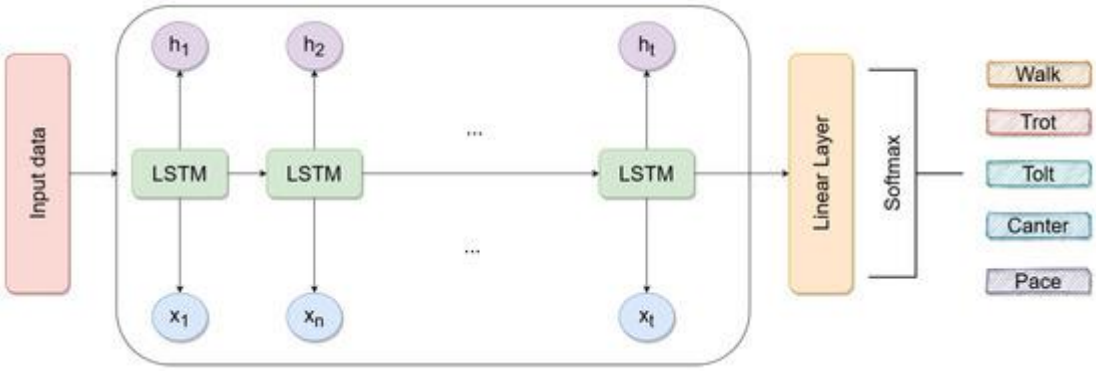

Figure 7. Architecture of the gait classification model used in this study

The model receives input from mobile sensors that have been rotated into a specific frame of reference. In this setup, the z-axis represents vertical motion, while the x- and y-axes represent horizontal motion. In the world-frame, the horizontal axes correspond to south–north and east–west directions, whereas in the horse's frame, they represent front–back and left–right movements. We also investigated including speed as an additional input feature, calculated from GPS coordinates. Consequently, the input consisted of six dimensions from the accelerometer and gyroscope (three each) and, optionally, one dimension for speed.

Smoothing the classifier output

Since the model predicts a gait label for each segment independently, occasional misclassifications may occur. For instance, the model might predict that a horse is in tölt, briefly in walk for 100 milliseconds, and then back to tölt—an implausible sequence given horse locomotion. To address this, two post-processing smoothing techniques were applied to the sequential outputs in the leave-one-out test sets. The first method utilizes the output probability vector from the model's final linear layer to adjust predictions over time, reducing brief inconsistencies in the predicted gait sequence.

The vector output of the softmax layer is updated with respect to the vectors preceding it in time using the exponential weight decay defined as:

$$z_0 = h_0, and z_t = \frac{z_{t-1} + h_t}{2} \quad (1)$$

The vector h denotes the output from the LSTM and z the refined vector after using exponential decay. This method ameliorates the problem to some extent, but does not fix it completely. To further refine the result, we

applied a majority vote window of size 7, which we slid over the gaits as determined by the zt vectors **(Figure 8)**. Concretely, we define

$$g_t = \arg max\ z_t \quad (2)$$

as the gait chosen at time t through the largest component of z_t. We then define g_t' as the most common gait in the set

$$\{g_{t-3}, g_{t-2}, g_{t-1}, g_t, g_{t+1}, g_{t+2}, g_{t+3}\}$$

where ties are broken by using g_t. The sequence g_t' represents the output of our method after the two post-processing steps.

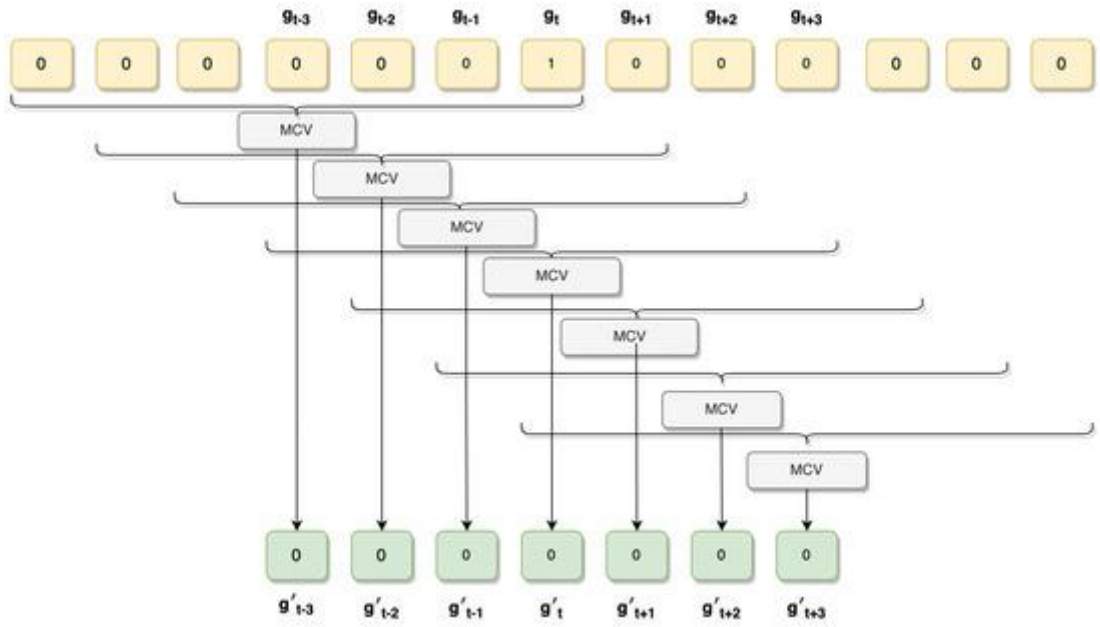

Figure 8. Sliding-window majority voting for smoothing predictions

In this illustration, gt represents the gait label assigned after the initial preprocessing step, while gt′ is the label determined after applying the majority-vote procedure. "MCV" stands for "Most-Common Value," and the curly brackets indicate the window over which the most frequent label is selected.
This post-processing approach smooths the predicted sequence by considering both previous and subsequent predictions within the window, reducing abrupt, short-lived changes in the model output.

Performance measures
To measure the performance of the model across all gaits, we used the micro-averaged classification accuracy defined as

$$\frac{\text{number of correctly classified examples}}{\text{total number of examples}}, \quad (3)$$

Accuracy is defined as the proportion of correct predictions, calculated by dividing the sum of the diagonal elements of the confusion matrix by the total number of entries. To evaluate the performance for individual gaits, we report **one-vs.-all accuracy**, treating each gait as a binary classification problem against all other gaits. Additionally, we present the **macro-averaged gait classification accuracy**, obtained by averaging the single-gait accuracies across all gaits.

Results

Validation of töltsense labels

From the TöltSense logs, 4,421 one-second windows were generated, encompassing 179 gait transitions. After excluding windows labeled −2 or −1, 4,371 valid test cases remained. Applying transition exclusion periods further reduced the number of usable test cases **(Table 4)**. TS label accuracy was computed as the number of correct predictions divided by the total number of valid cases.

Table 4 presents how the measured accuracy varied when different exclusion periods were applied for transitions identified by either the judges or the TS system.

Table 4. Gait classification accuracy for different sizes of exclusion windows.

TS Excl (ms)	Judge Excl (ms)	Test Cases	Accuracy
2000	2000	3358	99.73%
1000	1000	3757	98.86%
1000	0	3909	97.95%
0	1000	3990	96.62%
0	0	4371	93.89%

The length of the exclusion period was systematically varied from 0 to 2 seconds, revealing that longer exclusion windows corresponded to higher accuracy. When no exclusion periods were applied, the micro-averaged accuracy was 93.89%. Increasing the exclusion to 2 seconds for both judge- and TS-identified transitions raised the micro-averaged accuracy to 99.73%. The distribution of gaits within the dataset is illustrated in **Figure 9**. **Figure 10** presents the confusion matrix for a 1-second exclusion period applied to both TS and judge transitions, while **Figure 11** shows the confusion matrix when no exclusion period was used.

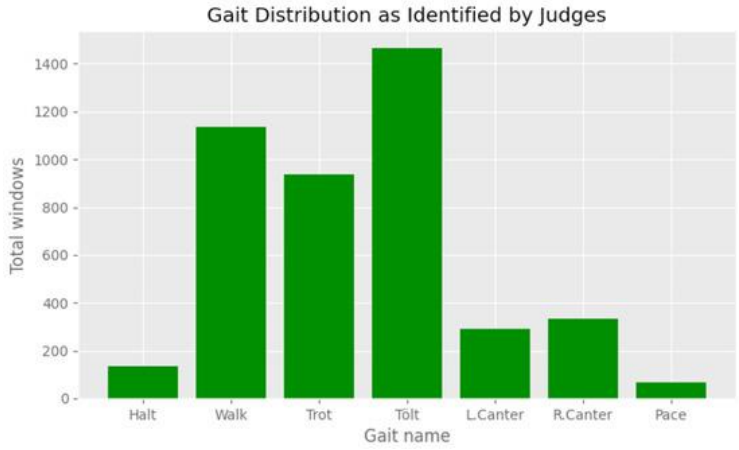

Figure 9. Distribution of gaits as identified by the judges. We note that pace is under-represented

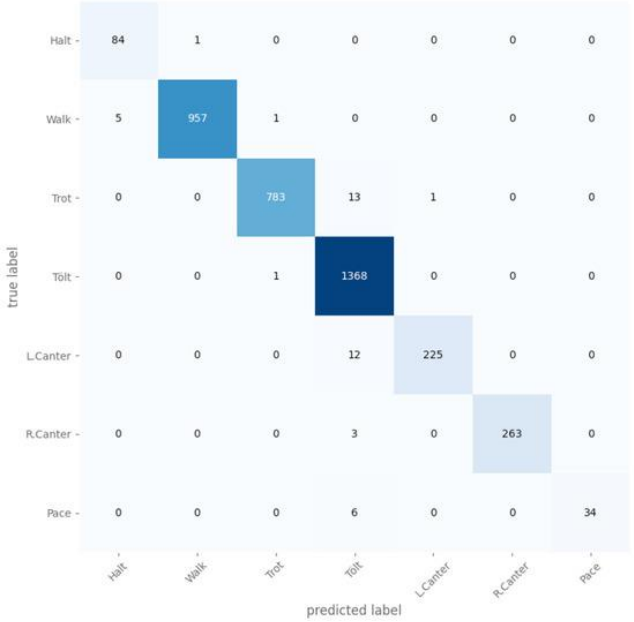

Figure 10. A confusion matrix for the case where both the TS and judge exclusion period was 1000 ms

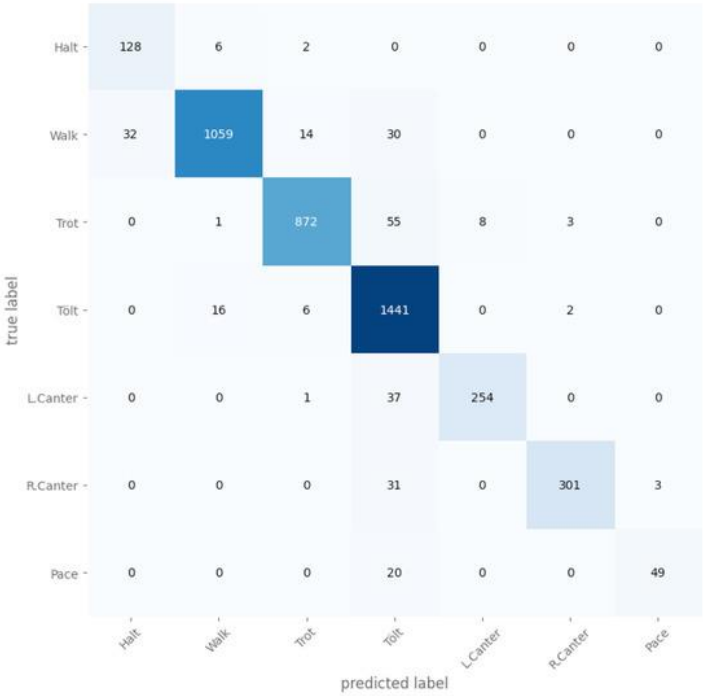

Figure 11. A confusion matrix for the case where both the TS and Judge exclusion period was 0 ms.

Evaluation of sequence models with mobile phone data

Data from 17 Icelandic horses (see Methods) were used to assess the performance of various sequential models for gait classification. To estimate generalization, a leave-one-horse-out cross-validation approach was applied: data from one horse were withheld during training and used for validation. For additional performance testing, a separate hold-out set excluded four horses entirely from the training data. Due to the limited number of horses capable of performing flying pace, one horse appeared in both training and test sets; however, the rider differed in each set to ensure independence.

Several neural network architectures were compared under this framework. The micro-averaged classification accuracy—calculated across all horses—is summarized in **Table 5**. All tested models exceeded 90% accuracy, with the bidirectional LSTM (Bi-LSTM) achieving the top performance of 94.4%.

Table 5. Micro-averaged gait classification accuracy for each horse, computed using leave-one-horse-out cross-validation. Each model was trained and evaluated five times, and the mean accuracy is reported

Model	Macro avg.
Bi-LSTM	94.4
GRU	91.2
LSTM	93.3
1D CNN	93.9

Detailed per-horse results for the Bi-LSTM model are presented in **Table 6**. The highest classification accuracy was observed for walk at 97%, followed by canter at 94%, flying pace at 93%, tölt at 89%, and trot at 82%. For four horses, the overall micro-averaged accuracy was lower (see Discussion). Overall, the model performed as expected, successfully distinguishing the two Icelandic-specific gaits, tölt and flying pace, without misclassification.

Table 6. Cross-validation results for each horse using the Bi-LSTM model, with each horse left out in turn. Evaluations were repeated with five different random initializations of the model, and the micro-average accuracy for each gait is reported. The macro-average accuracy across all gaits was 0.91. Training data per horse are listed in Table 3

Horse ID.	Canter	Flying Pace	Walk	Trot	Tölt	Micro avg.
1	1.0	-	1.0	1.0	1.0	1.0
2	1.0	0.87	0.98	0.72	0.82	0.95

3	1.0	-	1.0	0.97	0.99	0.99
4	0.61	-	0.83	0.82	0.96	0.89
5	-	-	0.99	0.68	0.94	0.94
6	0.69	-	0.99	0.97	0.99	0.98
7	0.99	-	1.0	0.88	0.92	0.95
8	1.0	-	0.95	0.92	1.0	0.97
9	1.0	-	0.93	0.88	1.0	0.96
10	-	0.95	-	-	-	0.95
11	-	0.9	-	-	-	0.9
12	0.96	-	0.98	0.44	0.74	0.87
13	-	-	0.97	0.0	0.98	0.97
14	1.0	-	1.0	1.0	1.0	1.0
15	1.0	-	1.0	0.57	0.61	0.82
16	1.0	-	0.97	0.66	0.5	0.74
17	0.99	-	1.0	0.95	0.93	0.94
Macro avg.	0.94	0.93	0.97	0.82	0.89	0.94

Improved gait classification using a horse-centered reference frame

Aligning sensor measurements to the world-frame isolates vertical motion along a single axis but provides limited information about lateral or longitudinal movements relative to the horse. By rotating the data into a horse-centered frame, acceleration along the front–back and side-to-side axes can be more clearly represented, which we expected to enhance classification accuracy.

To investigate this, we trained an LSTM network with 200 units on 1.5-second segments of accelerometer and gyroscope data from 17 Icelandic horses **(Table 3)**. Although the dataset was unbalanced across gaits, the collection method produced thousands of usable segments for each gait, as illustrated in **Figure 6**.

Performance was assessed on a hold-out set consisting of five rides. When using data rotated to the horse's frame, the model achieved an average accuracy of 96.1% **(Figure 12)**, while the same model trained on world-frame rotated data reached 93.9% **(Figure 13)**. These results demonstrate that referencing the horse's own orientation provides clearer motion patterns and improves gait classification.

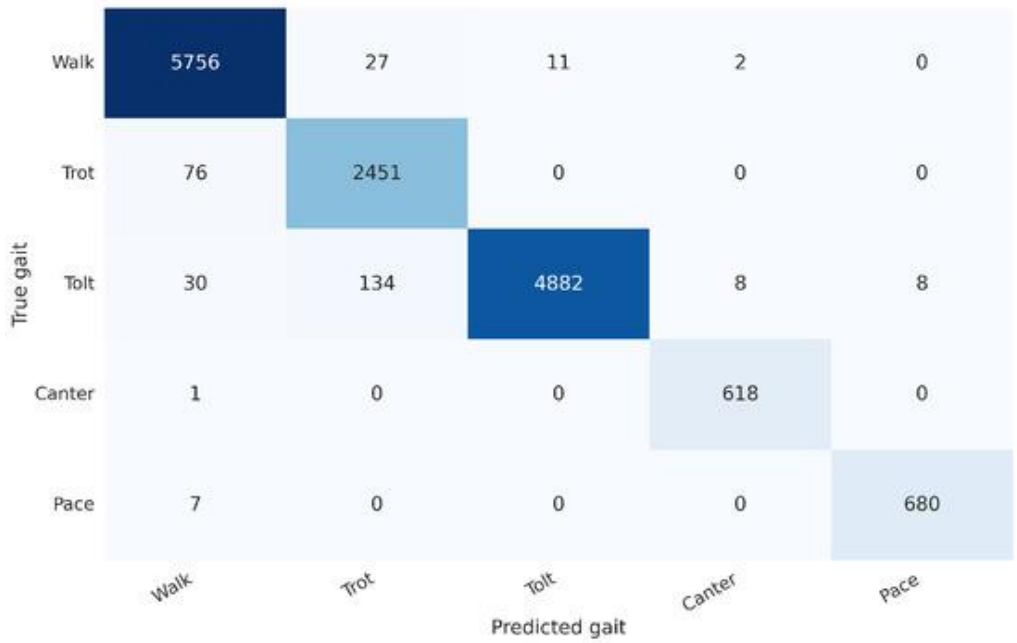

Figure 12. Confusion matrix for the gait classification model on the test set with 98.0% accuracy (best out of 9 random seeds with average accuracy at 96.1% and median accuracy at 95.8%). We use an input interval of length 1.5 s, where the input signal is aligned to the horse's frame of reference

Figure 13. Confusion matrix for the gait classification model on the test set

This matrix corresponds to the best-performing model across nine different random initializations, achieving 96.1% accuracy. The average accuracy across all runs was 93.9%, with a median of 92.7%. The model used 1.5-second input segments, with sensor signals aligned to the world-frame.

Robustness of the model to input interval length

Using shorter input segments can improve the responsiveness of gait predictions, which is particularly beneficial in interactive applications where quick reaction to gait changes is required. For recurrent models such as LSTMs, shorter segments also reduce computational demand, making them more suitable for deployment on mobile devices.

Our results indicate that model performance is relatively stable across input intervals ranging from 0.5 to 4 seconds **(Figure 14)**. The highest accuracy on the test set was observed with 3-second segments, while the mean accuracy was greatest with 4-second segments. For 1.5-second intervals, the model achieved an average accuracy of approximately 96% over nine runs, each initialized with a different random seed. Longer interval lengths showed greater variability in test accuracy, which can be partly attributed to the smaller number of segments available for training and testing, increasing the relative fluctuation in performance metrics.

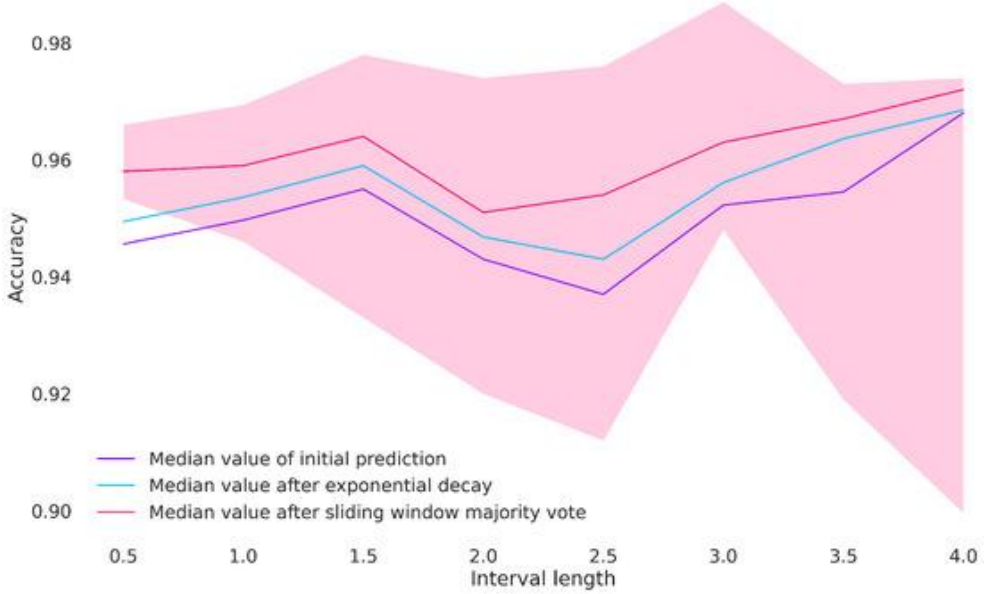

Figure 14. Accuracy of gait classification across different input interval lengths (seconds) using horse-frame rotation

The plot shows classification accuracy averaged over nine evaluations, each with a different random seed. Blue and red lines represent results after applying post-processing to smooth the predictions and improve overall accuracy. The shaded area indicates the range between the highest and lowest post-processed accuracies for each interval length.

Impact of Input Signal Variations on Classification Performance

We also investigated how different input signals and sampling rates affected model performance for 1.5-second segments aligned to the horse's frame of reference. These variations reflect practical differences between mobile devices, as some may lack either a gyroscope or accelerometer, and sampling rates can differ widely. Additionally, we evaluated the effect of including speed, derived from the phone's GPS data rather than accelerometer measurements, on classification accuracy.

Table 7 summarizes the results. The table reports median micro-averaged accuracy over nine runs with different random seeds, with the best-performing value for each sampling rate highlighted in bold. "G" indicates gyroscope-only input, "A" indicates accelerometer-only input, and "G+A" represents using both sensors.

Table 7.

	without Speed			with Speed		
Rate	**G**	**A**	**G+A**	**G**	**A**	**G+A**
10 Hz	80.8	**96.9**	92.2	76.7	95.7	91.1
15 Hz	91.4	**97.4**	96.8	92.4	96.1	96.0
25 Hz	92.3	**97.8**	97.1	92.5	97.1	96.6
50 Hz	92.6	96.8	96.4	90.7	**97.3**	95.8

The findings indicate that accelerometer data are the primary contributor to gait classification accuracy. From our analysis, including gyroscope readings or speed measurements did not consistently enhance model performance. Since acceleration inherently reflects changes in speed, the additional speed signal may offer little new information for the model. Moreover, GPS-derived speed may be unreliable or unavailable in indoor settings or areas with poor signal reception.

With respect to sampling rates, performance remained largely stable when using only accelerometer data at lower frequencies. However, when gyroscope data were included, accuracy dropped noticeably at 10 Hz and 15 Hz. Interestingly, higher sampling rates did not necessarily improve performance; the best results were observed at 25 Hz rather than 50 Hz.

Discussion

This study demonstrates the first smartphone-based classifier capable of identifying all five gaits of the Icelandic horse, including flying pace. Using raw accelerometer and gyroscope signals rotated to the horse-frame, derived from the magnetometer and GPS data, the Bi-LSTM model achieved up to 94.4% accuracy without the need for manual feature engineering.

A key innovation of our approach is the efficient labeling method: data were collected simultaneously from smartphones and the TöltSense (TS) system. TS generates gait labels based on four IMU sensors attached to the horse's limbs, capturing limb movement patterns that would normally require expert observation or detailed frame-by-frame video analysis. Manual labeling is not only resource-intensive but also constrained by environmental conditions, such as lighting, indoor arenas, or weather. In contrast, TS enables scalable and cost-effective collection of gait labels without such limitations.

Cross-validation showed that accuracy exceeded 0.9 for most gaits, with trot at 0.82 and tölt at 0.89. The primary confusion occurred between tölt and trot, likely due to limited training data for trot and minor overfitting, as training set performance exceeded test set performance.

Tölt is a four-beat gait with diagonal footfalls and partial suspension, whereas trot is a two-beat gait with full suspension. Tölt is intermediate between trot and flying pace, and variations in speed, quality, and style can make it difficult to distinguish from trot [15]. Even experienced observers may disagree when evaluating "trotty tölt" or "tölty trot," particularly at high speeds [21]. Furthermore, trot can be ridden using various techniques—sitting, rising, or two-point—which influence both rider and horse movement patterns [22]. These factors likely contribute

to model confusion, and future work could explore whether sensor data can differentiate between combined riding styles and gait variations.

On the test set, the model maintained a median accuracy of 96.9% with sampling rates down to 10 Hz. Notably, 25 Hz inputs outperformed 50 Hz, likely because shorter sequences are easier for LSTM models to process [23, 24]. Extremely low sampling rates reduce signal fidelity, explaining the performance drop at 10–15 Hz. Including speed as an additional input did not improve accuracy, consistent with previous work [8].

To assess TS reliability, we conducted a separate validation comparing TS labels to those of qualified sport judges. Agreement was high, exceeding 99% when excluding transition periods and around 94% without exclusions. While we used judge labels as "ground truth" to calculate accuracy, it is important to recognize that TS and human observers measure slightly different aspects of gait. Disagreements among judges are common—for example, when pacey tölt emerges from canter or at the boundaries between walk, tölt, and trot. Examining these disagreements could be a study in itself. A limitation of the validation study is the scarcity of flying pace segments. Although this study focused on Icelandic horses, the data collection methods described here could be adapted for a wide range of horse breeds in varied settings using different wearable sensors. Collecting additional test data would allow for a more robust evaluation of model generalization, particularly for flying pace, which had the fewest examples in our dataset [25].

It is well established that sensor placement can affect model performance [9,26], and it is likely that the location of the mobile phone contributed to the lower accuracy observed for some horse–rider pairs. Performance could potentially be enhanced by standardizing variables such as phone location, device model, riding surface, and rider clothing (tight versus loose-fitting). One solution might involve developing a model that explicitly accounts for phone placement, which could reduce variability and improve overall accuracy. Additionally, to improve usability in an app, the model could be designed to automatically detect the phone's position and potentially the riding surface, eliminating the need for manual input from the user.

Improved mobile device sensors could further enhance performance. For example, while we relied on a 1 Hz GPS signal to determine movement direction, higher-frequency GPS measurements could provide more detailed direction estimates and additional parameters relevant to gait analysis. Pfau *et al.* showed that essential stride parameters could be obtained using a 10 Hz GPS signal [27]. Another potential enhancement is to use multiple sensors simultaneously, such as a smartphone and a smartwatch, which has been shown to improve speed estimation in horse-mounted sensors [28] and could similarly benefit rider-carried devices.

Mobile phones provide an accessible platform for large-scale data collection in equine studies. They could facilitate volunteer-driven research to observe horse behavior, classify horses, define phenotypes, and even correlate behavioral data with genetic information. Furthermore, mobile sensors have been used to detect lameness [9,29], and our labeling approach could support more accurate detection. Existing horse-mounted sensors and commercial solutions (e.g., https://equisense.com/ accessed on 1 May 2022; https://equinosis.com/ accessed on 1 May 2022) have been applied for this purpose [30-32].

Historically, human observation has been considered the standard for gait analysis. However, Bragança *et al.* [8] highlighted that human assessment can be inconsistent. Studies on lameness detection have also reported low intraobserver reliability [33, 34], underscoring the limitations of relying on subjective evaluation. Models that exceed human performance in tasks such as lameness detection may provide more consistent and reliable labels. Approaches similar to ours—collecting sensor data alongside labels generated by a high-performing model—could be used to train models for lameness detection. Additionally, unsupervised methods, such as anomaly detection, could be applied to large-scale datasets to identify unusual patterns without manual labeling.

Conclusions

In this study, we examined whether mobile phone sensors can be effectively used to classify the gaits of Icelandic horses, which are capable of performing five distinct gaits: walk, trot, canter, tölt, and flying pace. Training data were labeled using the TöltSense (TS) system, and several machine learning models were evaluated. Our findings show that mobile phone sensors can provide accurate gait classification, with the best-performing model achieving 94.4% accuracy.

These results demonstrate the practicality of leveraging widely available mobile devices for equine gait analysis. Unlike traditional horse-mounted sensors, mobile phones are easy to carry and deploy, offering a low-cost, flexible

approach for collecting gait data. This approach could facilitate larger-scale studies and reduce logistical constraints associated with sensor attachment to the animal.

It should be noted that classification accuracy may vary depending on environmental conditions, sensor placement, and device quality. Additional research is required to explore these factors and to determine the effectiveness of mobile phone-based gait classification across different breeds, riding conditions, and sensor configurations.

Overall, this work confirms that mobile phone sensors represent a feasible and convenient tool for gait classification in Icelandic horses. By providing an accessible alternative to traditional methods, these findings pave the way for broader applications, including mobile-based systems for real-time gait monitoring and large-scale equine studies.

Abbreviations

The following abbreviations are used in this manuscript:

TS	TöltSense
RNN	Recurrent neural network
LSTM	Long short-term memory
GPS	Global positioning system
GRU	Gated recurrent unit
CNN	Convolutional neural network

Acknowledgments: We would like to thank Thilo Pfau for helpful comments on the manuscript. We would also like to thank Ann Winter, Nanco Lekkerkerker, Rebecca Hughes, and Harriet Vincent for labelling in the TöltSense validation study. We would further like to thank the Icelandic Directorate of Labour, the University of Iceland, and Horseday for funding H.B.D. in the summer of 2021.

Conflict of Interest: H.E. declares no conflict of interest. H.B.D.'s work in the summer of 2021 was partially funded by Horseday ehf., and he has been employed by Horseday ehf. since the fall of 2021. M.R.Ó. is a shareholder of Horseday ehf., and T.R. is the owner of Toltsense Ltd.

Financial Support: This research was funded by the Icelandic directorate of labour, the University of Iceland, and Horseday ehf.

Ethics Statement: None

References

1. Hildebrand M. Symmetrical gaits of horses. Science. 1965;150(3697):701–8.
2. Olsen E, Haubro Andersen P, Pfau T. Accuracy and precision of equine gait event detection during walking with limb- and trunk-mounted inertial sensors. Sensors. 2012;12(6):8145–56.
3. Holt D, George LBS, Clayton HM, Hobbs SJ. A simple method for equine kinematic gait event detection. Equine Vet J. 2017;49(5):688–91.
4. Bosch S, Serra Bragança F, Marin-Perianu M, Marin-Perianu R, Van der Zwaag BJ, Voskamp J, et al. EquiMoves: a wireless networked inertial measurement system for objective examination of horse gait. Sensors. 2018;18(3):850.
5. Eerdekens A, Deruyck M, Fontaine J, Damiaans B, Martens L, De Poorter E, et al. Horse jumping and dressage training activity detection using accelerometer data. Animals. 2021;11(10):2904.
6. Pasquiet B, Biau S, Trébot Q, Debril JF, Durand F, Fradet L. Detection of horse locomotion modifications due to training with inertial measurement units: a proof-of-concept. Sensors. 2022;22(13):4981.
7. Robilliard JJ, Pfau T, Wilson AM. Gait characterisation and classification in horses. J Exp Biol. 2007;210(Pt 2):187–97.

8. Serra Bragança FM, Broomé S, Rhodin M, Björnsdóttir S, Gunnarsson V, Voskamp JP, et al. Improving gait classification in horses using inertial measurement unit (IMU) data and machine learning. Sci Rep. 2020;10:17785.
9. Pfau T, Weller R. Comparison of a standalone consumer-grade smartphone with a specialist inertial measurement unit for quantification of movement symmetry in the trotting horse. Equine Vet J. 2017;49(1):124–9.
10. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
11. Eerdekens A, Deruyck M, Fontaine J, Martens L, De Poorter E. Automatic equine activity detection by convolutional neural networks using accelerometer data. Comput Electron Agric. 2020;168:105139.
12. Casella E, Khamesi A, Silvestri S. A framework for the recognition of horse gaits through wearable devices. Pervasive Mob Comput. 2020;67:101213.
13. Maga M, Björnsdotter S. Development of equine gait recognition algorithm [master's thesis]. Lund: Lund University; 2017.
14. Andersson LS, Larhammar M, Memic F, Wootz H, Schwochow D, Rubin CJ, et al. Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. Nature. 2012;488(7413):642–646.
15. Kristjánsson Þ, Reynisson G, Bárðarson S, Ævarsson S. The gaits of the Icelandic horse: basic definitions. Reykjavik; 2014.
16. Pfau T, Witte TH, Wilson AM. A method for deriving displacement data during cyclical movement using an inertial sensor. J Exp Biol. 2005;208(Pt 13):2503–14.
17. Pfau T, Reilly P. How low can we go? Influence of sample rate on equine pelvic displacement calculated from inertial sensor data. Equine Vet J. 2021;53(5):1075–81.
18. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint. 2014;arXiv:1412.3555.
19. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 2005;18(5–6):602–10.
20. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Proc 3rd Int Conf Learn Representations (ICLR); 2015.
21. Zips S, Peham C, Scheidl M, Licka T, Girtler D. Motion pattern of the tölt of Icelandic horses at different speeds. Equine Vet J. 2001;33(Suppl 33):109–111.
22. Persson-Sjodin E, Hernlund E, Pfau T, Haubro Andersen P, Rhodin M. Influence of seating styles on head and pelvic vertical movement symmetry in horses ridden at trot. PLoS One. 2018;13(4):e0195341.
23. Li S, Li W, Cook C, Zhu C, Gao Y. Independently recurrent neural network (IndRNN): building a longer and deeper RNN. In: Proc IEEE Conf Comput Vis Pattern Recognit; 2018. p. 5457–5466.
24. Khandelwal U, He H, Qi P, Jurafsky D. Sharp nearby, fuzzy far away: how neural language models use context. In: Proc 56th Annu Meet Assoc Comput Linguist; 2018. p. 284–294.
25. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput Surv. 2019;52(4):79.
26. Serra Bragança F, Rhodin M, Wiestner T, Hernlund E, Pfau T, Van Weeren P, et al. Quantification of the effect of instrumentation error in objective gait assessment in the horse on hindlimb symmetry parameters. Equine Vet J. 2018;50(3):370–6.
27. Pfau T, Bruce O, Edwards WB, Leguillette R. Stride frequency derived from GPS speed fluctuations in galloping horses. J Biomech. 2022;145:111364.
28. Darbandi H, Serra Bragança F, van der Zwaag BJ, Voskamp J, Gmel AI, Haraldsdóttir EH, et al. Using different combinations of body-mounted IMU sensors to estimate speed of horses: a machine learning approach. Sensors. 2021;21(3):798.
29. Marunova E, Dod L, Witte S, Pfau T. Smartphone-based pelvic movement asymmetry measures for clinical decision making in equine lameness assessment. Animals. 2021;11(6):1665.
30. Keegan KG, Kramer J, Yonezawa Y, Maki H, Pai PF, Dent EV, et al. Assessment of repeatability of a wireless inertial sensor-based lameness evaluation system for horses. Am J Vet Res. 2011;72(9):1156–63.
31. McCracken MJ, Kramer J, Keegan KG, Lopes M, Wilson DA, Reed SK, et al. Comparison of an inertial sensor system of lameness quantification with subjective lameness evaluation. Equine Vet J. 2012;44(6):652–6.

32. Yigit T, Han F, Rankins E, Yi J, McKeever K, Malinowski K. Wearable IMU-based early limb lameness detection for horses using multilayer classifiers. In: Proc IEEE Int Conf Autom Sci Eng (CASE); 2020. p. 955–60.
33. Keegan KG. Evidence-based lameness detection and quantification. Vet Clin North Am Equine Pract. 2007;23(2):403–23.
34. Arkell M, Archer R, Guitian F, May S. Evidence of bias affecting interpretation of local anaesthetic nerve blocks in equine lameness assessment. Vet Rec. 2006;159(11):346–48.