IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 April 2026, accepted 26 April 2026, date of publication 29 April 2026, date of current version 8 May 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3688918

RESEARCH ARTICLE

Semi-Supervised Learning for Time-Series Segmentation in Equine Gait Event Detection

MAHAUT GERARD [1,2,3,4], GUILLAUME DUBOIS [4], SANDRINE HANNE-POUJADE [4], CAMILLE HEBERT [4], HENRY CHATEAU [3,5], AND NEILA MEZGHANI [1,2], (Member, IEEE)

[1]Applied Artificial Intelligence Institute (I2A), TELUQ University, Montreal, QC H2S 3L5, Canada
[2]LIO, CHUM Research Center, Montreal, QC H2X 0A9, Canada
[3]ACAP3, CHUV-Eq, Ecole Nationale Vétérinaire d'Alfort, 14430 Goustranville, France
[4]LIM Group, 24300 Nontron, France
[5]Ecole Nationale Vétérinaire d'Alfort, 94700 Maisons-Alfort, France

Corresponding author: Mahaut Gerard (mahaut.gerard@vet-alfort.fr)

This work was supported in part by the Agence Nationale de la Recherche et de la Technologie (ANRT), France; in part by the Research Funding from the Région Nouvelle-Aquitaine, France; and in part by the Institutional Funding from French Ministry of Agriculture, France.

This work involved animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Clinical Research Ethics Committee under Application No. 2022-01-19.

ABSTRACT Locomotor injuries in horses impair both welfare and performance. Traditional visual assessments of equine locomotion by veterinarians are inherently subjective. To provide objective evaluation, systems using Inertial Measurement Units (IMUs) have been developed, enabling quantitative analysis by calculating trunk symmetry indices (head, withers, pelvis) during trot. Stance temporal parameters provide valuable information for locomotion analysis, but their accurate estimation requires precise detection of stance-related events. Gait events are typically identified using signal processing techniques applied to data from limb-mounted IMUs, often placed on the cannon bones. Despite their potential, these sensors are not well tolerated by all horses and the associated signal processing methods often fail to generalize across diverse locomotion patterns. A large amount of data is gathered by veterinarians in uncontrolled environments, opening the door for deep learning methods, but the annotation process needed for supervised learning remains time-consuming and costly. The aim of this study was to segment horse stances using trunk-mounted IMU data and to introduce two novel semi-supervised time-series segmentation methods, specifically tailored for contexts with limited strongly labeled data and abundant weakly labeled data. The best performing pipeline achieved errors of $-0.1\pm1.6\%$ of the stride duration for foot-on and $0.2\pm2.6\%$ for heel-off events of the right forelimb, and $0.2\pm1.5\%$ for foot-on and $-0.4\pm1.4\%$ for heel-off events of the right hindlimb. Beyond improving stance segmentation in real clinical settings, this work explores two semi-supervised methods for time-series segmentation.

INDEX TERMS Event detection, horse, segmentation, semi-supervised learning, time-series.

I. INTRODUCTION

Locomotor injuries in horses significantly affect animal welfare and athletic performance, with substancial economic consequences for stakeholders [1], [2], [3]. Evaluation of equine locomotion is commonly done by visual observation of a veterinarian. However, visual assessment of horses' lameness is known to lack reliability [4]. Recently, inertial measurement units (IMUs) based systems have been increasingly adopted to provide quantitative assessments of equine locomotion [5], [6], [7], offering complementary measurements that are both reliable [8], [9] and reproducible [10]. Most of these systems compute symmetry indices during trot, typically derived from differences in vertical displacement

The associate editor coordinating the review of this manuscript and approving it for publication was Patrizia Livreri.

of a trunk-mounted sensor between right and left limb phases within a stride. The stride itself is defined by two consecutive gait events of the same limb, emphasizing the fundamental role of trunk sensors and accurate stride event detection in equine locomotion analysis. In addition, temporal stride parameters provide valuable insights into locomotion analysis [11], [12] and can serve as indicators of fitness [13], fatigue [14], or compensatory mechanism associated with lameness [15], [16].

The stance phase of a limb is defined as the time interval between foot-on, when the hoof is fully in contact with the ground, and heel-off, when the hoof heels lift [17], [18]. The reference method for detecting these gait events relies on force plates; however, in uncontrolled environments, these events can be accurately measured using IMU mounted on the hoof [17], [18]. Specifically, the foot-on event coincides with the beginning of the hoof gyroscope norm plateau, while heel-off coincides with the end of the hoof accelerometer norm plateau [19]. However, hoof-mounted IMUs are impractical for routine veterinary locomotor examinations, as secure attachment is challenging, and the sensors are highly exposed to damage during exercise. As an alternative, several studies have used IMUs mounted on the cannon bone of each limb [19], [20], [21]. Darbandi et al. [22] further explored trunk-based sensor placements, reporting the feasibility of detecting forelimb and hindlimb stance events from withers- and sacrum-mounted IMUs, respectively, across different gaits. This later study suggested that trunk-mounted IMUs may be sufficient for stance phase detection, which is particularly relevant given their widespread use in equine locomotion analysis. Existing gait event detection studies have been conducted on horses trotting on treadmills [20], on soft ground [19], [21], [22] or on hard ground [19], [21]. Since ground surface is known to influence equine locomotion [23], developing gait event detection methods that are robust across different ground conditions remains a critical objective.

Deep learning models have demonstrated promising performance for gait events detection using IMU data [24]. In particular, Long-Short-Term Memory (LSTM) models have been successfully applied to recognize gait phases in humans [25], [26] and detect gait events in horses [22]. However, the performance of supervised Neural Network (NN) models is constrained by the need for large volumes of quality labeled data, whose collection and annotation are often time-consuming and costly. To mitigate this limitation, semi-supervised learning approaches have been proposed to leverage both labeled and unlabeled datasets [27], [28]. In this context, exploring semi-supervised learning for gait event detection aims to develop more efficient and robust models that better leverage the available data.

The objective of this study was to assess the feasibility of accurate and practical equine gait event detection in real clinical settings while reducing sensor complexity and annotation requirements. To this end, the study main contributions are:

- assess whether stance phases can be accurately segmented using trunk-mounted IMUs only in uncontrolled environments in a diversity of trot speed, ground, figure, and horses, while decreasing errors on foot-on and heel-off events compared to previous studies;
- propose two original semi-supervised time-series segmentation methods adapted to limited strongly annotated data and abundant partially annotated data (weakly labeled data), and suited for classes that are not well-separated;
- evaluate their potential and performance relative to other approaches built for equine gait event detection.

II. RELATED WORKS

Most of the semi-supervised learning research focused on classification [29]. Van Engelen and Hoos [29] divided semi-supervised classification models into three categories: wrapper methods, unsupervised preprocessing, and intrinsically semi-supervised models. Wrapper methods are based on supervised learners that are iteratively trained with the labeled data and the pseudo-labels generated from the predictions of the earlier iterations of the learners. Pseudo-labels chosen at each iteration typically come from confident predictions. These methods are widely known and among the oldest in semi-supervised learning [28], as they can be used with any supervised model. Unsupervised preprocessing approaches employ unlabeled and labeled data in two distinct phases. Most of the time, the unlabeled data are leveraged in an unsupervised manner for feature extraction, data clustering, or pre-training. Finally, intrinsically semi-supervised models utilize both labeled and unlabeled samples in the objective function. Depending on the assumptions that can be made on the data, this approach is typically based on maximum-margin, perturbation-based, manifold-based, or generative models.

In semi-supervised image segmentation, the same categories are found [30] but the methods are adapted to solve this more complex task. Jiao et al. [30] found that the wrapper method, which pseudo-labels images by choosing unlabeled pixels with predicted probability greater than a fixed threshold was the easiest way to go, though it is subject to noise. Other wrapper-based methods divide error into intra-class similarity and inter-class inconsistencies. Unsupervised pre-processing algorithms propagate labels in comparison with each class prototype, which has the advantage of avoiding confirmation bias as the labeling process does not result directly from the segmentation model. Unsupervised regularization for curriculum learning involves solving an easier task first to support the main learning objective. For image segmentation, Kervadec et al. [31] first trained a regressor to predict the size of the target region and then trained the segmentation network using both a supervised loss and an unsupervised loss based on the discrepancy between the predicted and the estimated region sizes. Unsupervised regularization for consistency learning enforces invariance of predictions under different perturbations [30], [32], yet invariant perturbations may be challenging to find for time-series.

To the best of our knowledge, few studies have specifically addressed semi-supervised time-series segmentation. Soares Junior et al. [33] segmented trajectories with a cost function designed to maximize homogeneity within consecutive points sharing the same label and similarity to labeled segments. They used aggregated features to represent each segment and employed Euclidean distance as a comparison metric. Using unsupervised pre-processing, Leodolter et al. [34] improved transport mode segmentation by refining weak segmentation, i.e. noisy labels. They searched for new breakpoints between two transport modes by locating local maxima in the difference between preceding and succeeding subsegments. The resulting segments were then clustered per label using Dynamic Time Warping (DTW) distance. At evaluation time, no breakpoints were assumed, instead sliding windows of one second duration were compared to the clusters' centroids. The DTW distance is widely used to compare time-series of unequal length [34], [35]. Although these two approaches enable time-series segmentation, they have primarily been applied to classification tasks involving well-separated classes in terms of both duration and amplitude, where each class segment encompasses multiple cycles. In contrast, equine gait event detection requires identifying events that occur in every cycle, independently of ground types, movement patterns, or trotting speed. Building on existing semi-supervised learning principles, the present study introduces two semi-supervised pseudo-labeling methods, compatible with any supervised model, to perform time-series segmentation for equine gait event detection using trunk-mounted IMUs. Their performance is subsequently compared with that of a standard supervised learning approach.

III. MATERIAL AND METHOD

A. DATA

1) COLLECTION

A total of 103 horses were examined either during standard veterinarian locomotor assessments at a clinic specialized in equine locomotion between 2021 and 2024 or during weekly check-ups in elite horses stables between 2023 and 2024. Horses were presented in-hand at the trot in various relative speed (slow, fast), in various figures (straight line, right and left circles, straight line after flexion tests), and on two types of ground-surfaces (hard and soft). The dataset gathers 156 conditions (triplet trot, figure, ground). Data acquisition was performed using the EQUISYM system (Arioneo, LIM France, Nouvelle-Aquitaine, France) consisting of seven IMUs placed on the four cannons, head, withers, and pelvis. A dedicated fixture system was used to ensure the correct positioning of the IMU at the attachment site, with trained operators handling its placement. Each IMU was sampled at 200 Hz, synchronized with the other six, and equipped with a three-dimensional accelerometer (measuring acceleration within a full-scale range of ±8 gravitational force equivalence) and gyroscope (measuring angular velocity within a range of ±1000 degrees per second). In addition, thirteen horses (52 conditions) were further equipped with 3 supplementary IMUs, one on each right forelimb and hindlimb hoof and one on the left forelimb cannon (at the same position as the EQUISYM IMU) to achieve synchronization of the two sensor sets.

2) STRONG LABELLING

For the 13 horses equipped with additional IMUs, hoof-on and heel-off events were manually labeled using signals from the hoof-mounted IMUs. Hoof-on events were identified at the beginning of the gyroscope norm plateau and heel-off events at the end of the accelerometer norm plateau of each stance, following definitions from previous studies [17], [18], [19]. This resulted in 736 seconds of IMU signals and 950 (resp. 952) stances for the right forelimb (RF) (resp. right hindlimb (RH)). Data were distributed into 71% outside-stance and 29% inside-stance labels for the RF and into 74% outside-stance and 26% inside-stance labels for the RH. Hereafter, we refer to the annotations of these 13 horses as strong labels.

3) WEAK LABELLING

For the remaining 90 horses, gait events were automatically labeled using signal processing techniques applied to IMU data from the canon bone, following the method described by Hatrisse et al. [19]. While the method based on peak detection yields good results, it does not always precisely detect hoof-on and heel-off events and is not robust for all horses. In the following, these annotations are denoted as weak labels, in accordance with the terminology established in the weakly supervised learning literature [36]. This weakly labeled dataset resulted in 13767 seconds of IMU signals and includes 16889 (resp. 15899) stances for the RF (resp. RH). Data were distributed into 72% outside-stance and 28% inside-stance labels for the RF and into 76% outside-stance and 24% inside-stance labels for the RH.

4) PREPROCESSING

From the time-series data, only trunk (head, withers, pelvis) sensors were used for gait events detection. It resulted in 18 dimensions (3 sensors, 2 measures, 3 axis) signals pre-processed as follows:

1) Standardization of each dimension independently, with mean and standard deviation from the training set;
2) Addition of norm signals. For each location L (head, withers, pelvis) and each sensor S (accelerometer, gyroscope), a new dimension was added to the time-series defined as a L2 norm: $\sqrt{LS_x^2 + LS_y^2 + LS_z^2}$;
3) Sequence creation. Overlapping windows of size l were taken every k timesteps;
4) Data augmentation with jittering (sigma=0.1), scaling (sigma=0.02), and rotation (axis_range=0.5,

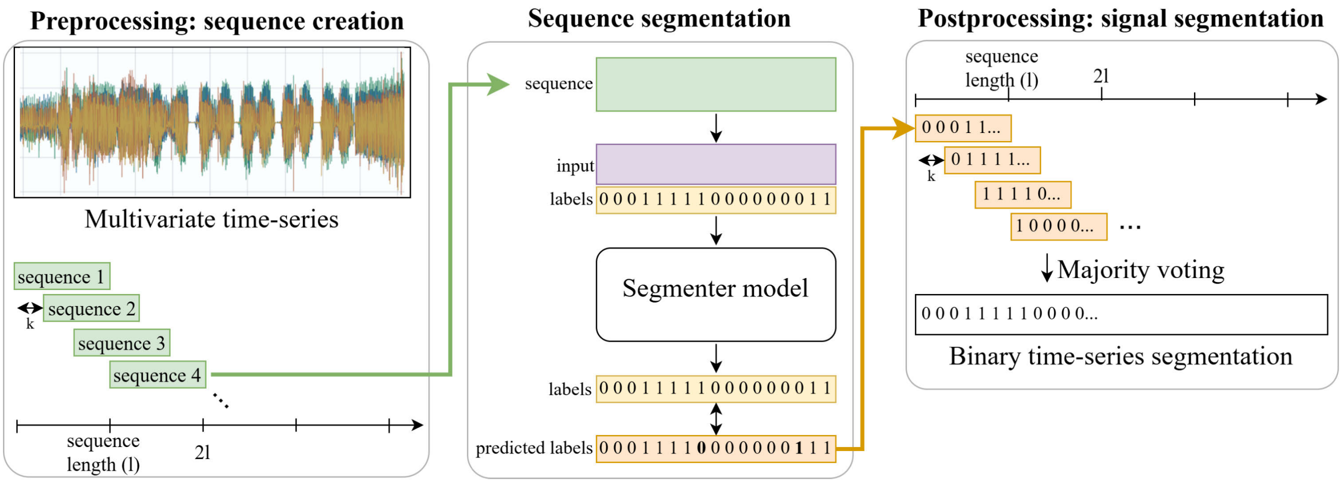


FIGURE 1. Time-series segmentation pipeline seen from the data perspective. During the preprocessing, raw signal is divided into sequences with an overlapping sliding window approach. Each timestamp from each sequence is then labelled during the sequence segmentation step. During postprocessing, the sequences labels are gathered by majority voting to finally segment the entire raw signal.

angle_range=$\pi/4$). The transformations were selected by fine-tuning among basic methods for time-series augmentation [37].

The time-series segmentation pipeline seen from a data perspective is described in Figure 1. At the end of the preprocessing steps, the input sequences were therefore of shape $(n, l, 24)$ and the output sequences were of shape $(n, l, 1)$, with n the number of sequences and l the sequence length.

B. MODELS

Models were built to perform a segmentation task based on labeled data. Gait event detection was restricted to the right limbs, based on the assumption of bilateral symmetry in quadrupeds trotting under sound conditions. To use strongly and weakly labeled data, two semi-supervised approaches achieving pseudo-labelling of weak labels were built. They were based on two models: a XGBoost and a CNN-LSTM. The XGBoost model was chosen because NN training efficiency may be limited given the reduced availability of strongly labeled sequences. The CNN-LSTM model was selected as it already showed promising performances when available data is reduced. This simple architecture, combining convolutional and LSTM layers, has been widely applied for equine locomotion analysis [22], [23], [38], [39]. Hyperparameters not detailed in the model descriptions are either described for each experiment (optimized hyperparameter depending on the settings) or defaulted to library values. Unless otherwise stated, hyperparameters were set by fine-tuning, a fixed seed and random weight initialization were used for all NN.

1) XGBOOST-BASED SEGMENTATION

XGBoost took aggregated features as input to predict one label per sequence. For each sequence and each dimension, the minimum, maximum, median, mean, standard deviation (std), and amplitude values were used as aggregation metrics. At prediction time, all timestamps in the sequence are labeled with the unique label predicted for each sequence.

2) NEURAL NETWORK SEGMENTER: CNN-LSTM

The base NN segmenter was a combination of a Convolutional Neural Network (CNN) and of a LSTM. The CNN-LSTM took raw time-series data as input and had the following architecture. First, one CNN layer in one dimension with batch normalization, ReLU activation, padding set to 'same', striding to 1, and output size 32. It was followed by one LSTM layer with batch normalization, tanh activation, return sequence parameter set to True and output size 16. Finally, the output layer was a CNN with sigmoid activation, padding set to 'same' and output size 1. Inside-stance points (label=1) and outside-stance points (label=0) are grouped together, meaning that transitions between the two classes (from 0 to 1 or from 1 to 0) are relatively infrequent compared to periods of class stability. Therefore, a zeros initialization is used for the output layer. It resulted in a network with 6385 trainable parameters.

3) SEMI-SUPERVISED LEARNING ITERATIVE APPROACH: SSL-IT

The semi-supervised iterative approach (SSL-IT) relies on the assumption that when two independent sources of weak labels (initial and predicted) coincide, then the associated labelling error is marginal. This pseudo-label filtering strategy is illustrated in Figure 2 and consists of the following steps:

1) A XGBoost-based pipeline was trained using only strongly labeled data;
2) The trained model was then used to predict labels for each timestamp in the weakly labeled dataset, generating what we refer to as pseudo-labels;

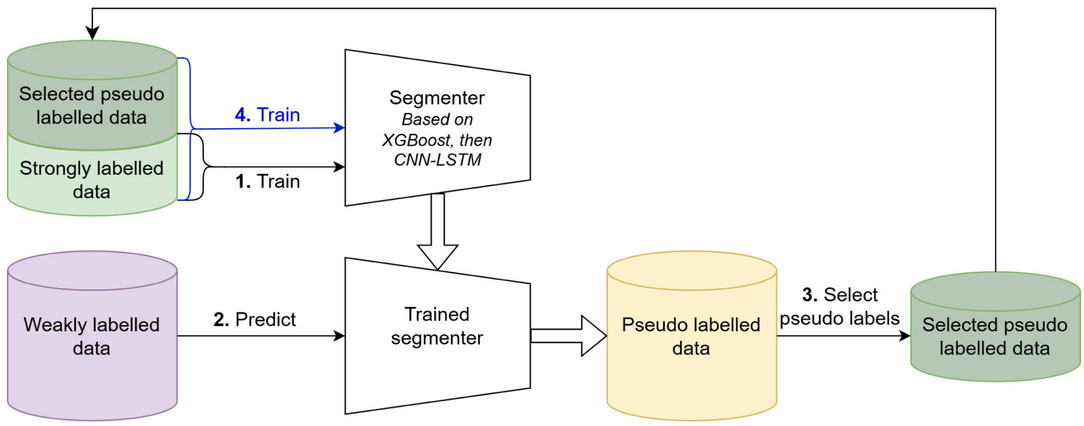


FIGURE 2. **Semi-supervised learning - iterative approach (SSL-IT) pipeline. The XGBoost segmenter is first trained using strongly labelled data (1), the trained segmenter is then used to compute pseudo-labels from weakly labelled data (2), the pseudo-labels that coincide with weakly labelled data are selected (3), the CNN-LSTM segmenter is finally trained with strong and pseudo-labels (4). Steps 2 to 4 are repeated iteratively.**

3) Pseudo-labels to retain were selected based on agreement with the original weak labels. To this end, pseudo-labeled foot-on and heel-off events were compared with the corresponding weakly-labeled events. All stances for which the differences between pseudo-labeled and weakly labeled foot-on and heel-off events were below the defined threshold *th* were selected;
4) The CNN-LSTM model was trained using strongly labeled and selected pseudo-labeled data.

Steps (2) to (4) were repeated iteratively until either the maximum number of iterations (*iter*) was reached or the validation accuracy began to decrease.

Two baselines were used to assess the efficiency of this approach: (1) the XGBoost trained during the first iteration on strongly labeled data to determine whether the subsequent of the SSL-IT pipeline yield a performance gain over the initial model; and (2) the CNN-LSTM trained once with all data without pseudo-labelling to evaluate whether the iterative selection of weak labels throughout the pipeline is effective compared to using all weak labels indiscriminately.

4) SEMI-SUPERVISED LEARNING CLUSTER-BASED APPROACH: SSL-CL

The semi-supervised cluster-based approach (SSl-CL) adjusts weak labels using a stance prototype built from strong labels. This method was divided into three main parts, as shown in Figure 3.

1) Stance signal parts from the strongly labeled dataset were isolated and clustered using the K-means algorithm with one cluster. The resulting cluster centroid served as a prototype representing the typical strongly labeled stance pattern;
2) For each stance in the weakly labeled dataset, potential stances were isolated. Potential stances were all possible combinations of potential foot-on and heel-off events. A potential foot-on (resp. heel-off) can be spaced by $\pm t_{cl}$ by steps of k_{cl} of the original event. Only potential stances longer than the shorter stance in the strongly labeled dataset were retained. Consequently, for each stance, a maximum of $|footon_{potential}| * |heeloff_{potential}| = (\frac{2t_{cl}}{k_{cl}})^2$ potential stances could exist. For each stance, the selected stance (pseudo-label) among potential stances was the one minimizing the DTW distance to the cluster's centroid;
3) The CNN-LSTM was then trained using both the strongly labeled and pseudo-labeled data.

One baseline was used to assess the efficiency of this approach: the CNN-LSTM trained with all data before pseudo-labelling to evaluate whether the contrastive pseudo-labeling strategy brings additional benefit over simply training on all available data.

C. POSTPROCESSING

Postprocessing steps were applied to convert overlapping sequence segmentation outputs into continuous signal segmentation, producing one label every 5 milliseconds (ms). At prediction time, each timestamp may be associated with multiple predicted labels, due to overlapping sequences. Therefore, the final class assigned to each timestamp was determined by majority voting, following the approach used in previous studies [40], [41].

D. PIPELINES TRAINING AND EVALUATION

Given the relatively small number of strongly labeled individuals (n = 13), models were trained using the Leave-One-Out (LOO) validation process to assess the stability of the pipeline across individuals, with each horse serving as the test subject in turn. For each LOO iteration, we computed the mean and std of both training and testing accuracies, as well as the mean and std of the error in detecting foot-on and heel-off events. Errors were tracked in ms and in stride percentage (s%), to allow comparison with existing studies which may

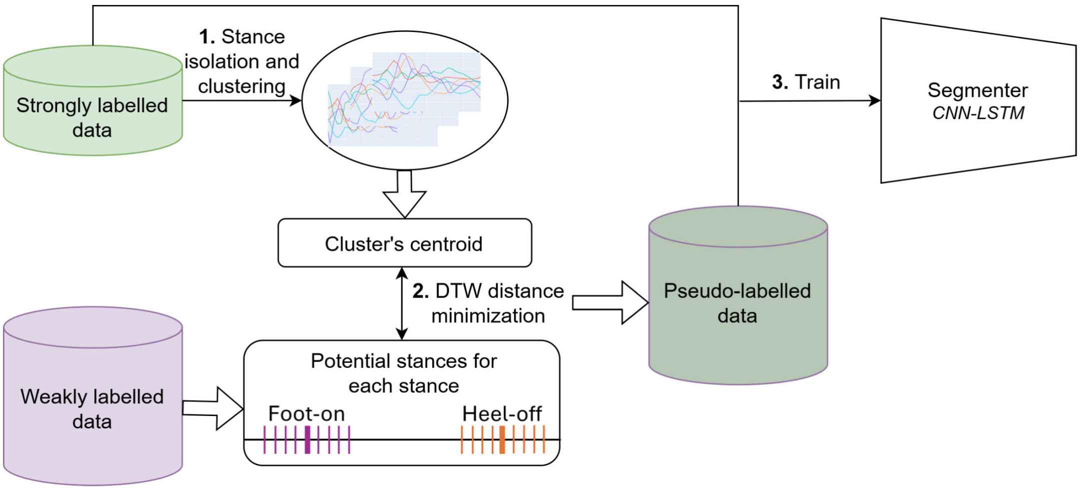


FIGURE 3. Semi-supervised learning - clustering appoach (SSL-CL) pipeline. Stances from strongly labelled data are isolated and clustered (1), the cluster's centroid representing the stance prototype; each stance from the weakly labelled data are adjusted by DTW distance minimization compared to the cluster's centroid (2); the CNN-LSTM segmenter is finally trained using strong and pseudo-labels (3).

use either unit.The LOO validation process therefore served as an estimation of the pipelines robustness, although the limited number of strongly labeled samples restricted the statistical power.

Finally and independently of the LOO protocol, to improve the clarity of the results presentation, the 13 strongly labelled horses were randomly split into a training set (n=8) and a testing set (n=5) to further evaluate the pipelines. Accuracy and event detection errors were computed for this split as well.

IV. RESULTS

Unless otherwise stated, time-series inputs to the models consisted of sequences of length $l = 512$, sampled every $k = 20$. They were set following prior works on equine locomotion ([22], [23], [41]). The number of training sequences built from the strongly labeled data (<6000) being smaller than the number of the trainable parameters of the CNN-LSTM, the CNN-LSTM baseline was trained with strongly and weakly labeled data together. The accuracies obtained through the LOO validation are presented in Table 1. The results for each model on the test set are displayed in Table 2. The CNN-LSTM was trained using an Adam optimizer and a batch size of 1024 on 100 epochs with a Mean Absolute Error (MAE) loss and a learning rate of 1e-4. To account for class imbalance, a class weight was used and set as $0 : nt/(2 * n0)$, $1 : nt/(2 * n1)$, with nt the total number of timestamps, $n0$ the number of timestamps with label 0, and $n1$ the number of timestamps with label 1.

A. SSL-IT

The threshold was set to $th = 0.02s = 20ms$. This value was chosen to be lower than observed errors std (27.2ms, 17.2ms, 24.3ms, 24.1ms) of the XGBoost baseline on strongly labelled samples, and lower than the errors std

TABLE 1. Accuracy (Acc) ± standard deviation (std) at training (train) and testing (test), with LOO evaluation, for the right forelimb (RF) and right hindlimb (RH). Bold values highlight for each column the best accuracy.

	RF		RH	
Model	Acc train	Acc test	Acc train	Acc test
CNN-LSTM	95.4±0.1	95.3±1.8	95.1±0.0	96.2±0.7
XGBoost	92.7±0.3	92.4±2.7	92.6±0.3	91.9±2.5
SSL-IT	**98.8±0.2**	**95.8±2.2**	**99.0±0.1**	**96.9±0.6**
SSL-CL	95.2±0.1	94.9±2.1	94.1±0.1	94.7±1.1

reported by Darbandi et al. [22] for gait event detection using trunk-mounted IMUs (21.0ms, 17.0ms, 29.0ms, 25.0ms). The maximum number of iterations was $iter = 5$ and was chosen because no substantial change (number of sequences selected, accuracies) was observed after five iterations. The sequence length used for the XGBoost was $l = 36$, which is the mean stance duration at the trot. When shifting to the CNN-LSTM, the sequence length was changed to $l = 128$ to contain at least one trot stride.

1) LOO ITERATIONS

The details of LOO iterations were gathered in Tables 3, 4, and 5. No outlier iteration was observed, which confirms the robustness of the pipeline for each individual horse used as test. As displayed on Table 1, the train accuracy was stable with a std of 0.2% (resp. 0.1%) and high with a mean of 98.8% (resp. 99.0%) for the RF (resp. RH). The mean test accuracy was 3 points (resp. 2.1 pts) smaller than the mean train accuracy, the std was increased to 2.2% (0.6%) for the RF (resp. RH). Compared to the CNN-LSTM baseline, the train (+3.4 pts) and test (+0.5 pt) accuracies for the RF, and the train (+3.9 pts) and test (+0.7 pt) accuracies for the RH were higher with the SSL-IT pipeline. They were also higher than the XGBoost baseline train (+6.1%) and test (+3.4%)

TABLE 2. Accuracy (%) and gait event (in milliseconds (ms) and stride percentage (s%)) errors with standard deviation (std) for the right forelimb (RF) and the right hindlimb (RH) on the test set for each pipeline. Bold values highlight the best values for each metric of each limb.

	Model	Accuracy train	Accuracy test		Error±std foot-on	Error±std heel-off
RF	CNN-LSTM	95.5	93.6	ms	-3.8±19.4	-19.4±**12.9**
				s%	-0.4±2.4	-2.5±**1.8**
	XGBoost	93.6	92.0	ms	1.8±27.2	6.9±17.2
				s%	0.5±2.7	0.8±2.1
	SSL-IT	**99.0**	**97.2**	ms	**-0.4±16.2**	**1.3**±19.1
				s%	**-0.1±1.6**	**0.2**±2.6
	SSL-CL	95.4	93.3	ms	-2.2±19.9	3.4±15.9
				s%	-0.3±2.6	0.4±2.1
RH	CNN-LSTM	95.3	96.3	ms	-9.8±32.5	-15.6±17.3
				s%	-0.3±2.5	-2.0±1.9
	XGBoost	93.5	92.4	ms	**1.8**±24.3	-9.3±24.1
				s%	**0.1**±2.8	-1.1±2.5
	SSL-IT	**99.2**	**98.0**	ms	1.9±**12.8**	-4.1±**11.4**
				s%	0.2±**1.5**	-0.4±**1.4**
	SSL-CL	94.0	93.0	ms	-9.1±25.2	**-0.7**±17.9
				s%	-0.3±2.3	**-0.2**±1.8

accuracies for the RF, and train (+6.4%) and test (+5.0%) accuracies for the RH.

2) EVALUATION ON THE TEST SET

When evaluated on the 5 horses in the test set (Table 2), the approach achieved a training accuracy of 99.0% for the RF and 99.2% for the RH, and a test accuracy of 97.2% for the RF and 98.0% for RH. The error biases were close to 0 with −0.1 s% (resp. 0.2 s%) for the RF (resp. RH) foot-on and 0.2 s% (resp. −0.4 s%) for heel-off. Overall, the errors were similar for the RF and the RH foot-on and heel-off events. Compared to the CNN-LSTM baseline, all metrics showed an improved performance with the SSL-IT pipeline. For the RF, the SSL-IT pipeline had higher train (+3.5 pts) and test (+3.6 pts) accuracies, smaller absolute foot-on bias (−0.3s%) and heel-off bias (-2.3s%), and equivalent gait event error std. For the RH, the SSL-IT pipeline had higher train (+3.9 pts) and test (+1.7 pts) accuracies, smaller absolute foot-on bias (−0.1s%), heel-off bias (−1.6s%), foot-on error std (−1s%), heel-off error std (−0.5s%). Similarly, the SSL-IT pipeline performed better than the XGBoost baseline.

B. SSL-CL

The parameters were set as follows: $t_{cl} = 0.04s = 40ms$ to be above error standard deviation observed for the CNN-LSTM baseline (maximum observed std of 32.5 ms), and $k_{cl} = 0.005s = 5ms$ by consistency with the 200 Hz sampling rate.

1) PSEUDO-LABELLING

During the pseudo-labelling process of weakly labelled data, 4 RF and 4 RH stances were randomly chosen. For each potential stance, the DTW distances to the cluster centroid were computed and are shown in Figure 4. The results show that DTW distance values vary across stances. For instance, stance S1 of the RF (resp. RH) ranges from 8 to 9 (resp. 7 to 10) whereas stance S2 ranges from 14 to 19 (resp. 13 to 16). The minimum for one stance can be greater than the maximum observed for another stance. Furthermore, we observe qualitatively that the distance used is consistent, as all lines of each plot follow the same tendency. Therefore, close potential stances (i.e. close points) are close in the computed distance.

2) LOO ITERATIONS

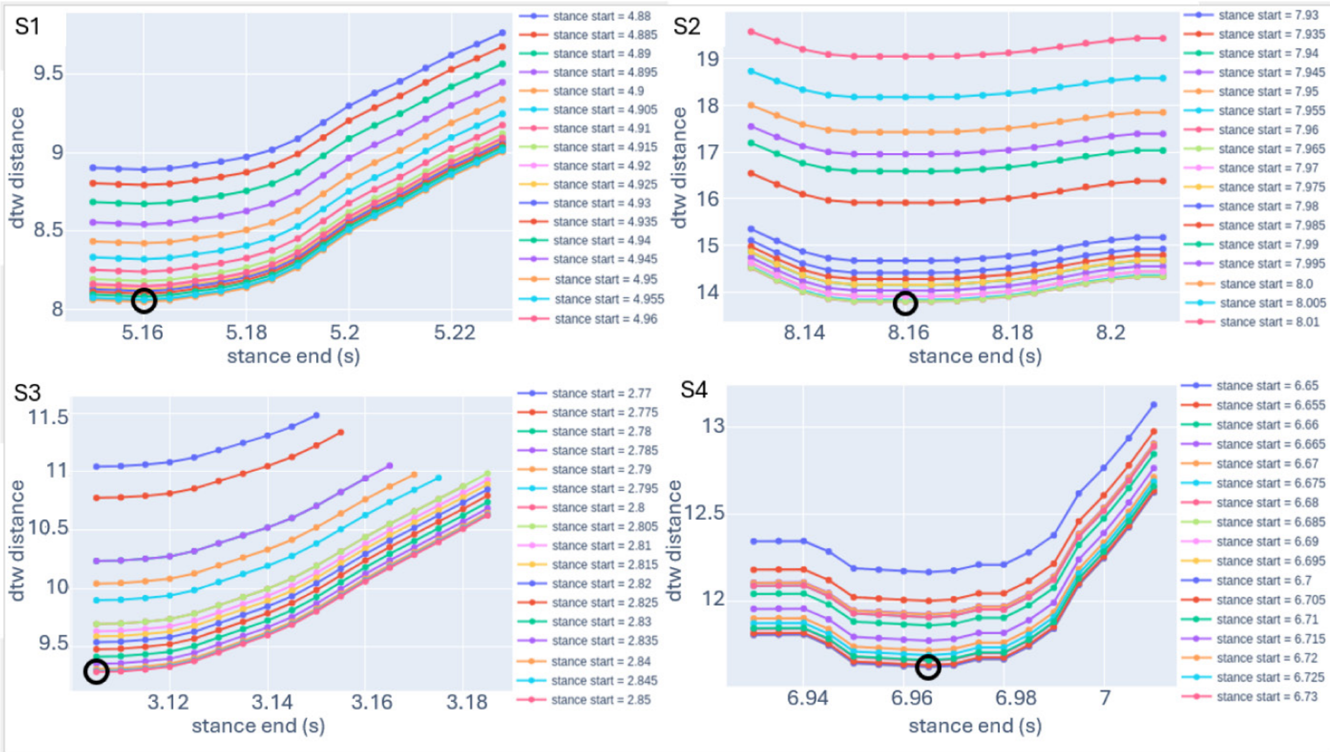
The details of LOO iterations were gathered in Table 5. No outlier iteration was observed, which confirms the robustness of the pipeline for different training and testing sets. As displayed on Table 1, the train accuracy was stable with a bias of 0.1% for the RF and the RH, and satisfactory with a mean of 94.9% for the RF and 94.7% for the RH. No overfitting was observed, highlighting a generalizable training. Compared to the CNN-LSTM baseline, the train and test accuracies for the RF were within std intervals and train (−1 pt) and test (−1.5 pts) accuracies for the RH were reduced in the SSL-CL pipeline. Compared to the XGBoost baseline, the train (+2.5 pts) and test (+2.5 pts) accuracies for the RF, and the train (+1.5 pts) and test (+2.8 pts) accuracies for the RH were higher with the SSL-CL.

3) EVALUATION ON THE TEST SET

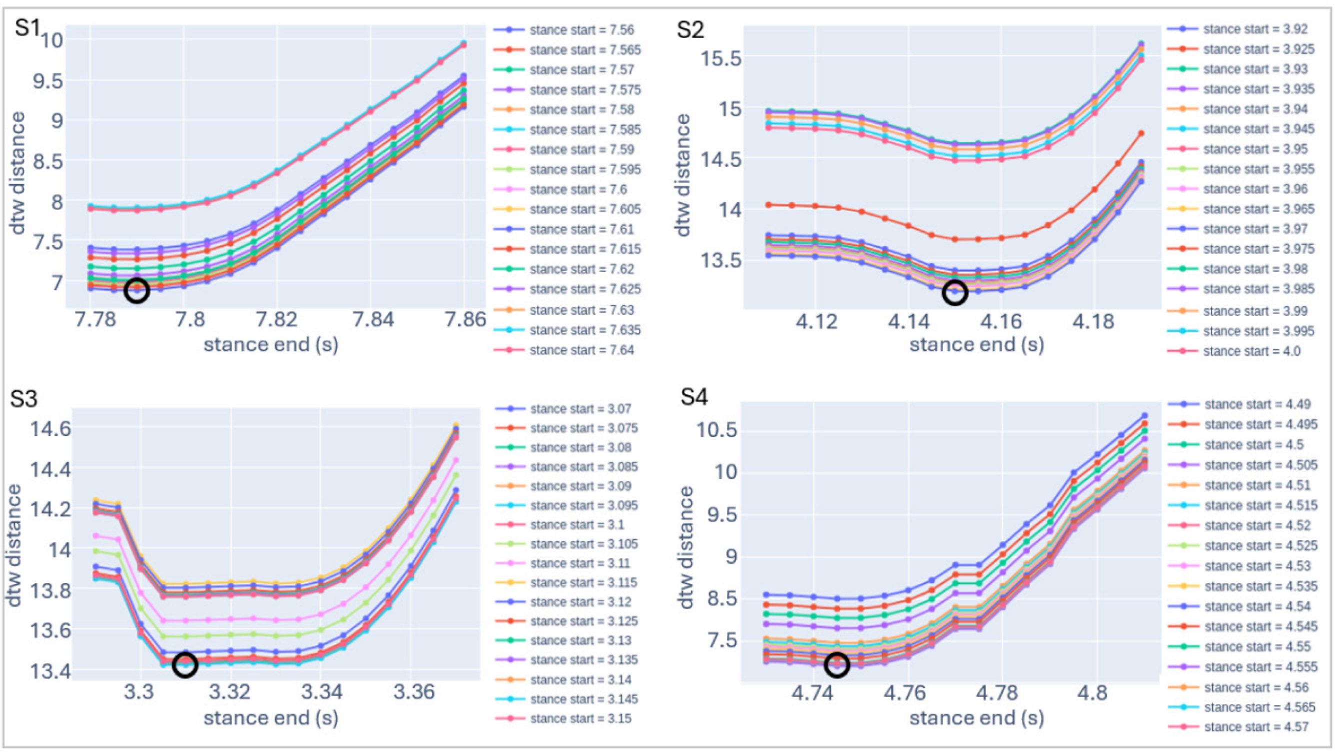
When evaluated on the 5 horses in the test set (Table 2), the approach achieved a train accuracy of 95.4% for the RF and 94.0% for the RH, with test accuracies of 93.3% for the RF and 93.0% for the RH. The error biases in event detection were close to 0 with −0.3 s% (resp. −0.3 s%) for the RF (resp. RH) foot-on and 0.4 s% (resp. −0.2 s%) for heel-off. Overall, the errors magnitudes and trends were similar between RF and RH for both foot-on and heel-off events. Compared to the CNN-LSTM baseline, the SSL-CL pipeline reduced the absolute bias for the RF heel-off events by -2.1s% and the RH heel-off events by −1.8s%, whereas other metrics remained comparable.

V. DISCUSSION

This study presented two original semi-supervised approaches for time-series segmentation. They achieve equine gait events



(a) RF



(b) RH

FIGURE 4. Potential stance DTW distance to cluster's centroid for 4 right forelimb (RF) stances chosen randomly (a), and 4 right hindlimb (RH) stances chosen randomly (b), each numbered from S1 to S4. The selected stance (black circle) on each plot minimizes the DTW distance (y-axis).

detection using trunk-mounted IMU data only. The LOO iterations demonstrated that the proposed pipelines are stable across different sets of training and testing sequences. This stability supports the validity of the independent test set and allows for a meaningful comparison of performance metrics, which is discussed in the following sections.

TABLE 3. LOO iterations, with each row for the test of one horse, for the CNN-LSTM, for the right forelimb (RF) and hindlimb (RH).

Horse	Size		Test stance number	Accuracy		Error (ms)		Error (%)	
	train	test		train	test	Foot-on	Heel-off	Foot-on	Heel-off
RF									
1	125599	260	81	95.4	94.1	-10.2±41.7	-16.3±17.3	-1.5±3.4	-2.1±2.3
2	125584	272	70	95.4	89.1	-6.8±26.7	-21.3±18.0	-1.6±3.3	-3.0±2.7
3	125232	456	96	95.4	93.8	0.8±14.5	-34.0±13.0	0.1±1.8	-3.9±2.0
4	125789	171	64	95.5	94.2	-18.1±13.3	-26.1±10.5	-2.3±1.8	-3.2±0.6
5	125831	145	56	95.4	95.2	-5.3±12.9	-28.2±9.4	-0.6±1.6	-3.5±0.9
6	126013	56	44	95.3	93.5	-11.6±22.6	-25.0±13.0	-2.1±2.8	-3.4±1.5
7	125777	164	58	95.4	95.1	-6.4±14.6	-24.1±10.7	-1.1±2.0	-3.4±1.4
8	125742	183	68	95.4	91.7	-6.8±15.5	-42.8±13.5	-1.4±1.9	-6.2±1.8
9	125619	245	85	95.4	93.9	-11.1±18.5	-31.2±15.1	-1.7±1.8	-4.5±1.8
10	125768	168	70	95.4	94.0	-8.4±25.0	-35.0±10.0	-1.0±2.9	-4.2±1.3
11	125588	260	89	95.6	95.9	3.0±12.6	-17.8±10.6	0.3±1.9	-2.7±1.2
12	125779	116	57	95.4	95.0	-18.4±11.5	-23.4±12.8	-2.6±1.4	-2.6±1.7
13	125425	288	112	95.4	94.7	-8.5±23.8	-27.5±9.8	-1.3±3.1	-3.8±1.3
RH									
1	125599	260	82	95.1	95.5	-1.5±35.4	-10.5±10.5	1.1±2.5	-1.3±1.4
2	125584	272	78	95.1	97.8	-1.2±9.2	0.8±37.8	-0.1±1.2	-0.9±2.2
3	125232	456	95	95.1	95.4	-3.5±26.0	-25.6±7.2	0.4±1.4	-3.1±1.1
4	125789	171	63	95.1	96.5	-7.6±31.8	-21.8±7.7	0.5±0.9	-2.7±1.0
5	125831	145	55	95.1	95.3	-6.4±37.2	-25.3±5.3	1.0±1.4	-3.0±1.1
6	126013	56	43	95.1	96.2	-8.7±31.9	-26.8±6.5	0.3±1.4	-3.3±0.7
7	125777	164	58	95.1	96.7	-7.1±34.2	-18.4±7.4	0.7±1.2	-2.2±1.1
8	125742	183	68	95.1	95.5	-10.3±29.0	-23.8±18.9	-0.1±1.8	-3.1±2.4
9	125619	245	85	95.1	95.8	-13.6±31.5	-23.8±7.4	-0.3±1.3	-2.9±1.2
10	125768	168	72	95.1	96.8	-3.3±30.4	-6.6±30.6	0.9±1.6	-1.3±1.1
11	125588	260	88	95.1	97.1	-4.0±18.1	-9.0±8.0	-0.3±1.5	-1.3±1.1
12	125779	116	55	95.1	96.2	-23.7±42.0	-17.2±15.4	-0.8±1.5	-2.1±1.9
13	125425	288	110	95.1	96.2	-2.6±20.7	-12.3±33.3	0.1±1.1	-2.2±2.8

TABLE 4. LOO iterations, with each row for the test of one horse, for the SSL-IT pipeline.

Horse	Size		Test stance number	Accuracy		Error (ms)		Error (%)	
	train	test		train	test	Foot-on	Heel-off	Foot-on	Heel-off
1	45796	1066	81	98.9	96.0	0.2±16.6	6.9±18.7	0.0±1.9	0.5±3.2
2	32240	1019	70	99.2	88.4	-2.3±31.1	0.0±19.7	0.7±2.3	-0.1±2.8
3	42914	1349	96	99.0	96.5	6.5±17.1	-5.4±20.9	0.8±2.0	-0.6±2.7
4	45320	794	64	98.7	97.1	-4.5±14.1	1.1±16.9	-0.5±1.4	0.2±2.6
5	32050	711	56	99.0	97.0	-6.1±15.5	4.1±30.8	-0.8±1.9	1.0±4.0
6	34553	504	44	98.9	96.4	-14.5±31.2	7.8±30.1	-1.4±2.5	1.0±4.9
7	46042	710	58	98.5	97.1	11.1±15.4	4.3±9.8	1.1±1.8	0.6±1.2
8	44018	183	68	98.9	96.1	-1.0±16.4	-12.5±16.2	-0.2±2.1	-1.4±1.5
9	32692	1097	85	98.7	96.7	1.3±23.6	-2.0±15.2	0.0±1.8	-0.3±2.7
10	53417	954	70	98.6	97.1	5.7±15.0	2.1±20.7	0.5±1.5	0.3±2.7
11	31737	1010	89	98.9	95.8	10.3±16.1	16.1±25.8	1.5±2.2	2.2±3.7
12	50326	667	57	98.7	95.8	-4.9±18.1	13.9±33.7	-0.4±1.9	1.7±3.9
13	53364	1346	112	98.8	95.9	-18.1±22.5	-7.9±12.0	-1.4±1.8	-0.9±1.6
RH									
1	72071	1066	82	99.0	94.9	10.2±23.9	1.9±17.7	1.2±2.7	0.2±1.9
2	48332	1019	78	99	97.2	-3.3±12.4	1.4±12.1	-0.5±1.4	0.3±1.4
3	46962	1349	95	99.0	97.0	-2.2±12.6	-9.7±11.4	-0.3±1.5	-1.0±1.0
4	47067	794	63	99.0	97.0	0.8±16.4	-8.0±15.9	0.3±1.5	-0.8±1.6
5	46859	711	55	98.9	96.9	12.6±12.7	-6.0±11.9	1.4±1.5	-0.7±1.0
6	50074	504	43	99.0	96.7	4.4±13.1	-9.0±10.2	0.4±1.5	-0.8±1.1
7	81063	710	58	98.7	97.5	5.0±12.6	-2.7±8.4	0.6±1.5	-0.2±1.0
8	74924	849	68	98.9	97.0	4.9±13.1	-9.5±11.9	0.5±1.6	-1.2±1.4
9	87042	1097	85	99.1	96.9	-4.5±18.6	-10.2±10.6	-0.2±1.1	-1.2±1.2
10	94148	954	72	99.1	97.4	4.4±11.6	6.2±16.2	0.5±1.1	0.5±1.3
11	82459	1010	88	99.0	97.2	-1.7±13.7	-3.0±10.0	-0.2±1.7	-0.4±1.4
12	76691	667	55	99.0	96.8	-9.6±13.1	-9.3±11.6	-1.1±1.6	-0.8±1.3
13	81736	1346	110	99.0	97.2	-6.7±10.5	5.7±20.7	-0.8±1.2	0.5±2.6

TABLE 5. LOO iterations, with each row for the test of one horse, for the SSL-CL, for the right forelimb (RF) and hindlimb (RH).

Horse	Size		Test stance number	Accuracy		Error (ms)		Error (%)	
	train	test		train	test	Foot-on	Heel-off	Foot-on	Heel-off
RF									
1	121312	260	81	95.2	93.1	2.1±16.6	7.9±13.8	0.3±2.1	0.9±2.0
2	121294	272	70	95.2	88.5	-6.0±19.1	8.2±25.4	-0.6±0.9±3.4	
3	120939	456	96	95.2	93.5	4.7±15.3	0.6±15.0	0.7±2.0	0.0±1.9
4	121499	171	64	95.2	96.1	-16.6±18.3	0.4±8.6	-2.4±2.1	0.2±1.1
5	121558	145	56	95.2	95.8	-6.2±12.5	-6.9±9.1	-0.8±1.6	-0.9±1.2
6	121732	56	44	95.1	96.2	-15.6±20.2	-1.6±18.4	-1.7±2.7	-0.8±1.4
7	121514	164	58	95.3	97.5	2.2±15.2	-0.4±9.3	0.2±2.0	-0.1±1.2
8	121487	183	68	95.3	93.5	-2.4±18.4	-16.7±16.1	-0.5±2.6	-2.2±2.2
9	121348	245	85	95.2	96.2	-27.6±15.9	5.7±12.7	-3.5±2.3	0.6±1.5
10	121499	168	70	95.2	95.4	-4.5±23.7	1.8±18.8	-0.4±2.8	0.0±2.7
11	121323	260	89	95.2	93.9	6.9±12.3	4.1±11.6	0.9±1.7	0.5±1.7
RH									
1	121938	260	82	94.1	94.6	2.1±30.1	-14.3±19.4	1.3±2.2	-1.6±2.3
2	121920	272	78	94.2	96.3	3.3±23.9	-2.0±39.5	0.8±2.5	-1.3±2.6
3	121558	456	95	94.1	93.7	-4.9±34.2	-32.2±7.1	0.1±3.2	-1.0±1.2
4	122119	171	63	94.2	95.6	-7.8±35.7	-19.5±11.2	0.7±1.4	-2.6±1.4
5	122185	145	55	94.1	95.3	-11.6±35.5	-18.1±8.6	0.2±1.5	-2.1±1.2
6	122355	56	43	94.0	93.9	-12.6±31.2	-26.6±10.0	-0.3±1.7	-3.5±1.1
7	122133	164	58	94.2	96.3	-15.4±32.3	-23.1±9.7	-0.5±1.7	-2.8±1.6
8	122101	183	68	94.1	95.4	-3.9±32.4	-6.0±30.4	0.7±2.1	-1.3±1.9
9	121966	245	85	94.2	93.8	-5.5±43.3	-17.1±27.2	0.4±2.2	-2.3±2.4
10	122129	168	72	94.1	93.0	7.6±38.3	-1.3±12.5	2.2±2.6	-0.3±1.3
11	121951	260	88	94.2	93.3	-25.1±31.7	-21.4±10.6	-2.6±2.4	-3.2±1.4
12	122226	116	55	94.1	94.6	-51.2±31.3	-3.7±15.6	-5.0±2.0	-0.4±2.0
13	121876	288	110	94.1	94.7	-21.6±20.0	-5.0±27.5	-2.6±2.5	-1.1±3.1

A. METHODS COMPARISON

A slight overfitting was observed in the SSL-IT pipeline, with training accuracy exceeding test accuracy by 1.8% (resp. 1.2%) for the RF (resp. RH). This is likely due to the pseudo-labelling mechanism, which favors at each iteration, predicted stance segments closely matching the initial weak labels, resulting in a training dataset composed of similar samples. This process can ultimately reduce the model generalization and explain the overfitting observed. No such overfitting is seen in other pipelines.

Among the evaluated methods, the SSL-IT pipeline achieved the highest accuracies and lowest event detection errors. The method achieved similar performance across all contexts (figures: straight line, right circle, left circle; grounds: hard, soft) as illustrated in Figure 5. This indicates promising robustness across different conditions.

In contrast, the SSL-CL pipeline did not consistently outperform the baseline CNN-LSTM, although it remains promising for future developments. Notably, pseudo-labels in the SSL-CL pipeline are created only once, and can be reused with any supervised segmenter, without increasing computational complexity. Conversely, the SSL-IT pipeline is tightly coupled to the segmenter used, requiring iterative re-labelling.

B. COMPARISON WITH OTHER GAIT EVENT DETECTION METHODS

Focusing on gait phases detection in humans, Kreuzer et al. [42] obtained between 93% and 98% test accuracy when classifying the signal into 5 classes (initial contact, loading response, mid stance, terminal stance, swing phase) using a CNN-LSTM. Our test accuracies, in the range 93%-99%, are comparable.

Focusing on gait event detection in horses, two related works enable hoof-on and heel-off errors comparison. The SSL-IT pipeline achieved smaller biases than previous methods using cannon bone IMUs [19], with foot-on errors decreased by −1.7 s% (resp. −3.2 s%) for the RF (resp. RH) and heel-off errors decreased by −1.3 s% (resp. −1.5 s%). It also outperformed withers-mounted IMU for the RF events detection and sacrum-mounted IMU for the RH event detection [22], with foot-on errors decreased by −0.1 ms (resp. −8.5 ms) for the RF (resp. RH) and heel-off errors decreased by −2.6 ms (resp. −7.5 ms).

The standard deviations of the SSL-IT detection errors are comparable to those reported by Darbandi et al. [22] for the RF events, whereas the std values for the RH events are lower by −16.2 ms for foot-on and by −13.6 ms for heel-off. Overall, this study achieves similar performance for both forelimb and hindlimb stances, whereas prior studies typically reported higher errors for hindlimbs. A key factor contributing to this improvement is the combined use of three trunk-mounted IMUs (head, withers, pelvis), which are influenced by both forelimbs and hindlimbs movements. This likely explains the improved accuracy in RH gait events detection compared to models relying on a single IMU location.

Our methods achieve binary timestamp-level segmentation to detect stances. Hoof-on and heel-off events are extracted from each detected stance as the start and end timestamps, respectively. In contrast, Darbandi et al. [22] directly detected hoof-on and heel-off events. We chose timestamp-level stance detection to promote temporal consistency of stances, enable easy removal of artifacts, improve model robustness to noise, and produce a more balanced target distribution. However, our approaches do not directly minimize foot-on and heel-off detection errors. A custom loss function based on derivatives could be developed to better reflect these key metrics, since hoof-on and heel-off correspond to transition points in the predicted signal. Alternatively, direct event detection - as in Darbandi et al. [22] - are based on event-focused losses that explicitly penalize timing errors that can, in addition, intrinsically model different event types.

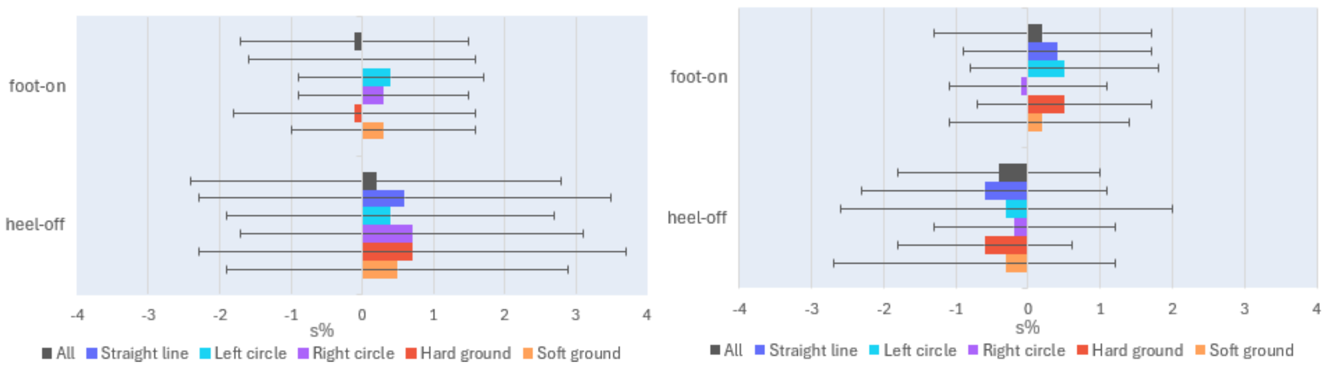
In addition, our data were collected in routine veterinary practice, across both hard and soft surfaces. The SSL-IT pipeline demonstrated robustness to ground type, offering a practical advantage over previous studies limited to soft ground conditions [22].

Importantly, the bias errors obtained with SSL-IT fall below the 5 ms resolution of the 200 Hz sampling rate, placing them within the limits of annotation reliability. Therefore, no further improvements in bias can be expected without increasing the sampling rate.

C. LIMITATIONS AND FUTURE WORK

This study would benefit from a larger set of manually (strongly) labeled data for testing. Although the stability of the pipelines was assessed using LOO iterations, the final testing was conducted on 5 horses only, which remains a small sample size given the inter-individual variability in equine locomotion patterns. In particular, the LOO iterations on the SSL-IT pipeline highlighted that the number of selected pseudo-labels varied across folds, and not all folds required the same number of iterations. Adding just a few individuals could create a validation set, which would enable an early stopping strategy and an exhaustive fine-tuning strategy, to further improve the approach performance and generalization. In addition, in this work, we assess pseudo-labelling effectiveness only indirectly through performance analysis. Future works will directly compare the final pseudo-labels with initial weak labels to quantify correction versus error propagation.

Given the limited number of strongly labelled data, a lightweight model architecture (CNN-LSTM) was chosen as a baseline to evaluate the proposed semi-supervised pipelines. However, this model does not reflect the current state-of-the-art segmentation architectures. More recent models, such as Temporal Convolutional Networks (TCN) [43] or Transformer-based models [44] could be



(a) RF (b) RH

FIGURE 5. Gait events (foot-on and heel-off) bias (mean error) and variability (error standard deviation) in the SSL-IT pipeline compared to the ground truth of the test set, expressed in stride percentage (s%), for the right forelimb (a) and the right hindlimb (b). Results are stratified by the context: figure (straight line, right circle, left circle) and ground (hard, soft). Across all conditions, bias remains close to zero, while variability reflects moderate dispersion depending on context. No systematic degradation is observed across figures or ground types, supporting the robustness of the method in heterogeneous field conditions.

explored in future work, under high-capacity architectural settings.

Although the associated works [19], [22] were also restricted to right limbs and that this assumption is commonly accepted, extending the approach to left limbs and to other gaits (e.g., walk or gallop) would be necessary to broaden the applicability of the method and to assess its robustness beyond trot-specific symmetry assumptions. Moreover, knowing the lameness profile of the included horses would be necessary to confirm the generalisability of the method to horses with pathological gait.

For the SSL-CL pipeline, the absence of an explicit distance threshold in the DTW-based pseudo-label selection implies that stance selection remains relative to each individual stance. Further improvements could set a threshold based on strong stance distance to cluster distribution and keep only pseudo-labels below this threshold. An iterative approach could also be considered to continue the pseudo-labelling in an iterative manner for stances farther to the cluster than the threshold. In addition, the current implementation relies on a single cluster centroid to represent the stance pattern, prior experiments concluded that the use of multiple centroids didn't improve the performance. The current dataset may be limiting the relevance of multiple prototypes construction. Though this choice may oversimplify gait variability across horses, it was motivated by the need to avoid the risk of overfitting to cluster-specific patterns in a setting where strongly labeled data is scarce. More data could enable alternative configurations using multiple centroids that could better capture inter-individual and contextual differences in stance dynamics. Future work could use multiple centroids per-limb or even per-limb per-context (figure, ground, speed), with each weak stance compared to all centroids. The final selected stance would be the stance that minimizes the distance stance-centroid all prototypes put together.

Finally, our two original semi-supervised approaches addressed the time-series segmentation with classes that are not well differentiated in terms of duration and amplitude, with events occurring every cycle. Future work could leverage the periodicity observed in the detected events to improve segmentation.

VI. CONCLUSION

This study proposes a first step toward addressing the challenge of horse stance segmentation using trunk-mounted IMU data, through the introduction of two novel semi-supervised time-series segmentation methods designed for settings with limited strongly labeled data and abundant weakly labeled data. It tended to demonstrate that trunk-mounted IMUs alone are sufficient to accurately detect foot-on and heel-off events in horses, even in uncontrolled environments characterized by heterogeneous ground surfaces and movement patterns. Two original semi-supervised methods for time-series segmentation were introduced, enabling the effective use of deep learning models in scenarios where strongly labeled data are scarce. Among the evaluated approaches, the iterative semi-supervised method (SSL-IT) achieved the best performance in terms of accuracy and event detection errors. The method proved robust across individuals and comparable, or superior, to previously reported approaches relying on limb- or withers/sacrum-mounted sensors. Beyond the specific application to horse gait analysis, the proposed semi-supervised strategies are broadly applicable to other time-series segmentation tasks involving recurrent events and limited annotated data. In particular, the SSL-CL method offers a reusable pseudo-labeling strategy compatible with any supervised segmenter, while SSL-IT provides higher accuracy at the expense of tighter model integration. By simplifying sensor configurations and reducing reliance on extensive manual annotation, this work contributes to making gait analysis systems more practical and scalable for veterinary practice and research. Future developments may focus on extending the approach to other gaits, refining loss functions, and exploiting gait periodicity to further improve segmentation performance.

ACKNOWLEDGMENT

The authors sincerely thank the CIRALE Team, including veterinarians, researchers, and technicians, for their crucial contribution to data collection.

REFERENCES

[1] S. Dyson, "Lameness and poor performance in the sport horse: Dressage, show jumping and horse trials," *J. Equine Veterinary Sci.*, vol. 22, no. 4, pp. 145–150, Apr. 2002.

[2] A. Egenvall, C. A. Tranquille, A. C. Lönnell, C. Bitschnau, A. Oomen, E. Hernlund, S. Montavon, M. A. Franko, R. C. Murray, M. A. Weishaupt, V. R. Weeren, and L. Roepstorff, "Days-lost to training and competition in relation to workload in 263 elite show-jumping horses in four European countries," *Preventive Veterinary Med.*, vol. 112, no. 3, pp. 387–400, Nov. 2013.

[3] M. M. Sloet van Oldruitenborgh-Oosterbaan, W. Genzel, and P. R. Van Weeren, "A pilot study on factors influencing the career of Dutch sport horses," *Equine Veterinary J.*, vol. 42, no. 38, pp. 28–32, Nov. 2010.

[4] K. G. Keegan, "Reliability of equine visual lameness classification," *Veterinary Rec.*, vol. 184, no. 2, pp. 60–62, Jan. 2019.

[5] J.-M. Denoix, "A look at lameness through the eyes of functional anatomy (and biomechanics)," Tech. Rep., Vol. 67, 2021.

[6] P. R. van Weeren, T. Pfau, M. Rhodin, L. Roepstorff, F. Serra Braganca, and M. A. Weishaupt, "Do we have to redefine lameness in the era of quantitative gait analysis?" *Equine Veterinary J.*, vol. 49, no. 5, pp. 567–569, 2017.

[7] T. Pfau, A. Fiske-Jackson, and M. Rhodin, "Quantitative assessment of gait parameters in horses: Useful for aiding clinical decision making?" *Equine Veterinary Educ.*, vol. 28, no. 4, pp. 209–215, 2016.

[8] T. Pfau, K. Landsbergen, B. L. Davis, O. Kenny, N. Kernot, N. Rochard, M. Porte-Proust, H. Sparks, Y. Takahashi, K. Toth, and W. M. Scott, "Comparing inertial measurement units to markerless video analysis for movement symmetry in quarter horses," *Sensors*, vol. 23, no. 20, p. 8414, Oct. 2023.

[9] F. J. Lawin, A. Byström, C. Roepstorff, M. Rhodin, M. Almlöf, M. Silva, P. H. Andersen, H. Kjellström, and E. Hernlund, "Is markerless more or less? Comparing a smartphone computer vision method for equine lameness assessment to multi-camera motion capture," *Animals*, vol. 13, no. 3, p. 390, Jan. 2023.

[10] I. Timmerman, C. Macaire, S. Hanne-Poujade, L. Bertoni, P. Martin, F. Marin, and H. Chateau, "A pilot study on the inter-operator reproducibility of a wireless sensors-based system for quantifying gait asymmetries in horses," *Sensors*, vol. 22, no. 23, p. 9533, Dec. 2022.

[11] T. H. Witte, K. Knill, and A. M. Wilson, "Determination of peak vertical ground reaction force from duty factor in the horse (Equus caballus)," *J. Experim. Biol.*, vol. 207, no. 21, pp. 3639–3648, Oct. 2004.

[12] S. J. Hobbs, L. St George, J. Reed, R. Stockley, C. Thetford, J. Sinclair, J. Williams, K. Nankervis, and H. M. Clayton, "A scoping review of determinants of performance in dressage," *PeerJ*, vol. 8, p. e9022, Apr. 2020.

[13] R. S. V. Parkes, R. Weller, T. Pfau, and T. H. Witte, "The effect of training on stride duration in a cohort of two-year-old and three-year-old thoroughbred racehorses," *Animals*, vol. 9, no. 7, p. 466, Jul. 2019.

[14] S. J. Wickler, H. M. Greene, K. Egan, A. Astudillo, D. J. Dutto, and D. F. Hoyt, "Stride parameters and hindlimb length in horses fatigued on a treadmill and at an endurance ride," *Equine Veterinary J.*, vol. 38, no. S36, pp. 60–64, Aug. 2006.

[15] H. H. F. Buchnefp, H. H. C. M. Savelberg, H. C. Schamhardt, and A. Barneveld, "Temporal stride patterns in horses with experimentally induced fore-or hindlimb lameness," *Equine Veterinary J.*, vol. 27, no. S18, pp. 161–165, May 1995.

[16] K. G. Keegan, D. A. Wilson, B. K. Smith, and D. J. Wilson, "Changes in kinematic variables observed during pressure-induced forelimb lameness in adult horses trotting on a treadmill," *Amer. J. Veterinary Res.*, vol. 61, no. 6, pp. 612–619, Jun. 2000.

[17] M. Tijssen, E. Hernlund, M. Rhodin, S. Bosch, J. P. Voskamp, M. Nielen, and F. M. S. Braganca, "Automatic hoof-on and -off detection in horses using hoof-mounted inertial measurement unit sensors," *PLoS ONE*, vol. 15, no. 6, Jun. 2020, Art. no. e0233266.

[18] M. Tijssen, E. Hernlund, M. Rhodin, S. Bosch, J. P. Voskamp, M. Nielen, and F. M. S. Braganca, "Automatic detection of break-over phase onset in horses using hoof-mounted inertial measurement unit sensors," *PLoS ONE*, vol. 15, no. 5, May 2020, Art. no. e0233649.

[19] C. Hatrisse, C. Macaire, M. Sapone, C. Hebert, S. Hanne-Poujade, E. De Azevedo, F. Marin, P. Martin, and H. Chateau, "Stance phase detection by inertial measurement unit placed on the metacarpus of horses trotting on hard and soft straight lines and circles," *Sensors*, vol. 22, no. 3, p. 703, Jan. 2022.

[20] M. Sapone, P. Martin, K. Ben Mansour, H. Chateau, and F. Marin, "Comparison of trotting stance detection methods from an inertial measurement unit mounted on the Horse's limb," *Sensors*, vol. 20, no. 10, p. 2983, May 2020.

[21] E. V. Briggs and C. Mazza, "Automatic methods of hoof-on and -off detection in horses using wearable inertial sensors during walk and trot on asphalt, sand and grass," *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0254813.

[22] 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 *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2022, pp. 329–335.

[23] J. I. M. Parmentier, F. M. S. Bragança, E. Hernlund, and B. J. van der Zwaag, "Terrain type detection for smart equine gait analysis systems using inertial sensors and machine learning," in *Proc. 19th Int. Conf. Distrib. Comput. Smart Syst. Internet Things (DCOSS-IoT)*, Jun. 2023, pp. 103–111.

[24] H. T. T. Vu, D. Dong, H.-L. Cao, T. Verstraten, D. Lefeber, B. Vanderborght, and J. Geeroms, "A review of gait phase detection algorithms for lower limb prostheses," *Sensors*, vol. 20, no. 14, p. 3972, Jul. 2020.

[25] T. Zhen, L. Yan, and P. Yuan, "Walking gait phase detection based on acceleration signals using LSTM-DNN algorithm," *Algorithms*, vol. 12, no. 12, p. 253, Nov. 2019.

[26] H. X. Tan, N. N. Aung, J. Tian, M. C. H. Chua, and Y. O. Yang, "Time series classification using a modified LSTM approach from accelerometer-based data: A comparative study for gait cycle detection," *Gait Posture*, vol. 74, pp. 128–134, Oct. 2019.

[27] O. Chapelle, S. Bernhard, and A. Zien, "Semi-supervised learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, p. 542, Mar. 2009.

[28] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. sciences, Univ. Wiconsin, Madison, Tech. Rep. 1530, 2005.

[29] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[30] R. Jiao, Y. Zhang, L. Ding, B. Xue, J. Zhang, R. Cai, and C. Jin, "Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation," *Comput. Biol. Med.*, vol. 169, Feb. 2024, Art. no. 107840.

[31] H. Kervadec, J. Dolz, É. Granger, and I. B. Ayed, "Curriculum semi-supervised segmentation," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, 2019, pp. 568–576.

[32] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12671–12681.

[33] A. Soares Junior, V. Cesario Times, C. Renso, S. Matwin, and L. A. F. Cabral, "A semi-supervised approach for the semantic segmentation of trajectories," in *Proc. 19th IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2018, pp. 145–154.

[34] M. Leodolter, P. Widhalm, C. Plant, and N. Brandle, "Semi-supervised segmentation of accelerometer time series for transport mode classification," in *Proc. 5th IEEE Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, Jun. 2017, pp. 663–668.

[35] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, ''Dynamic time warping,'' in *Discrete-Time Processing of Speech Signals*. Berlin, Germany: Springer, 2007, pp. 623–676.

[36] Z.-H. Zhou, ''A brief introduction to weakly supervised learning,'' *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[37] G. Iglesias, E. Talavera, Á. González-Prieto, A. Mozo, and S. Gómez-Canaval, ''Data augmentation techniques in time series domain: A survey and taxonomy,'' *Neural Comput. Appl.*, vol. 35, no. 14, pp. 10123–10145, May 2023.

[38] J. Schampheleer, A. Eerdekens, W. Joseph, L. Martens, and M. Deruyck, ''Detecting equine gaits through rider-worn accelerometers,'' *Animals*, vol. 15, no. 8, p. 1080, Apr. 2025.

[39] B. Savoini, J. Bertolaccini, S. Montavon, and M. Deriaz, ''Convolutional neural network for early detection of lameness and irregularity in horses using an IMU sensor,'' in *Proc. Int. Conf. Adv. Mach. Learn. Data Sci.*, Jul. 2025, pp. 318–324.

[40] M. Dallel, V. Havard, Y. Dupuis, and D. Baudry, ''A sliding window based approach with majority voting for online human action recognition using spatial temporal graph convolutional neural networks,'' in *Proc. 7th Int. Conf. Mach. Learn. Technol. (ICMLT)*, Mar. 2022, pp. 155–163.

[41] M. Gérard, S. Hanne-Poujade, G. Dubois, H. Chateau, and N. Mezghani, ''End-to-end horse gait classification in uncontrolled environments using inertial sensors,'' *IEEE Access*, vol. 13, pp. 80668–80679, 2025.

[42] D. Kreuzer and M. Munz, ''Deep convolutional and LSTM networks on multi-channel time series data for gait phase recognition,'' *Sensors*, vol. 21, no. 3, p. 789, Jan. 2021.

[43] J. Yan, L. Mu, L. Wang, R. Ranjan, and A. Y. Zomaya, ''Temporal convolutional networks for the advance prediction of ENSO,'' *Sci. Rep.*, vol. 10, no. 1, p. 8055, May 2020.

[44] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, ''Transformers in time series: A survey,'' 2023, *arXiv:2202.07125*.