6 February 2026

JAB_2025-0337.R1

Ear-worn inertial sensors can predict gait metrics and reconstruct vertical ground reaction force curves during running

Jake Stuchbury-Wass[1], Mathias Ciliberto[1], Kayla-Jade Butkow[1], Qiang Yang[1], Yang Liu[1], Ezio Preatoni[2], Dong Ma[3] and Cecilia Mascolo[1]

[1]Department of Computer Science and Technology, University of Cambridge, UK

[2]Department for Health, University of Bath, UK

[3]School of Computing and Information Systems, Singapore Management University

Conflict of Interest Disclosure: None to declare.

Correspondence Address:

Jake Stuchbury-Wass

15 JJ Thomson Avenue

Cambridge

CB3 0FD

UK

Abstract

Ear-worn wearables (aka: earbuds, hearables, or earables) are commonly used by runners for entertainment, and many modern devices also include inertial sensors for user interaction. We propose harnessing the technology embedded in earbuds to capture fundamental aspects of running mechanics and make them available to the wider community of users, outside a lab setting. While other wearables such as insoles, or ankle/sacrum mounted IMUs have already been presented, ear-worn devices may have a better potential for adoption and therefore offer an optimal compromise between validity of running gait analysis and usability. Thirty healthy participants (18 male, 12 female) ran on an instrumented treadmill (54000 gait cycles) and floor-mounted force plates (2800 gait cycles) at a variety of speeds. Building on the information brought about by the vibrations transmitted to, and motion of the head, we devised a gait event detection algorithm and a regression model to predict vertical ground reaction force (vGRF) waveforms. The validation of outcomes against quantities from force plates shows an average MAPE of 4.8 % on temporal metrics and 9.0 % on scalar GRF derived metrics. Additionally, the model tracks the full vGRF curve well, achieving an NRMSE of 11.1 % on unseen participants. Overall, we show comparable accuracy from an ear-worn consumer device in temporal and kinetic gait parameter estimation to specialist devices, paving the way for accessible running gait monitoring.

Keywords: **Running gait, kinetics, motion capture, machine learning, wearables**

Introduction

Running is one of the most popular sporting activities worldwide, with an estimated 621 million people participating globally [13]. During running, measuring gait events and ground reaction forces are essential for supporting performance analysis and injury prevention [5]. Direct links to injuries such as tibial stress fractures from ground reaction forces are debated in the literature e.g. [22], however, ground reaction forces are necessary to derive downstream measures, such as joint loading and moments, which have been indicated as biomechanical factors associated with injury risk [5]. Measurements of gait and ground reaction force metrics are typically confined to motion capture laboratories equipped with specialised equipment such as force plates, instrumented walkways, treadmills, and systems for kinematic analysis. Laboratory-based testing is costly, requires trained personnel, and provides only a limited, non-continuous snapshot of an individual's behaviour. Wearable technologies have emerged as a promising alternative for in-field gait analysis. However, to be truly effective and trusted, these devices must undergo rigorous validation [26].

Research aimed at predicting running gait events or ground reaction forces (GRF) via wearables emerged with works using the motion markers in a rigid body model to approximate gait kinetics [3] or how parameters of mass-spring model may be optimised to better capture movement kinetics [9]. Modern approaches typically employ pressure insoles e.g. [4] or inertial measurement units (IMUs) mounted on single locations (e.g., sacrum, tibia, foot) e.g. [1, 24] or multiple locations [11, 18, 27]. These studies have utilised dynamical models [24], signal processing-based approaches such as the Kalman filter [20] as well as machine learning approaches including traditional regression models e.g. [27] or deep learning methods such as CNNs [18], RNNs [1], LSTMs [4] and Transformers [34].. There are existing consumer devices that show demand from runners for tracking gait parameters and loading, examples include the Garmin Running Dynamics Pod and some Garmin heart rate monitors (Garmin Ltd., KS, USA)

which track a selection of spatiotemporal gait metrics, additionally, Arion Pro (ATO-Gear BV, Eindhoven, Netherlands), NURVV Run (NURVV Ltd., London, UK) and Striv (Striv, MA, USA) smart insoles can track loading using force-sensitive resistors and gait parameters with IMUs. However, all these solutions rely on purpose-built devices that lack widespread adoption or acceptance amongst runners [8]. It has also been demonstrated that ubiquitous smartwatches suffer from large motion artefacts during walking gait assessment [19], and this issue is likely to be exacerbated during running, limiting the suitability for in-depth biomechanical analysis.

Ear-worn wearables—commonly referred to as earbuds, hearables, or earables—represent a valuable resource for capturing and monitoring gait information whilst running. They are positioned in a relatively stable location on the body and are owned and regularly used by a majority of runners, with a survey reporting that 72% of runners listen to media whilst running [14]. Modern consumer-market devices, such as Apple AirPods Pro (Apple Inc., Cupertino, CA, USA), already incorporate inertial sensors for user interactions, such as recognising taps for playing and pausing audio and spatial audio. Consumer-market smart earbuds are currently utilised for spatial audio and actions such as pausing music. Earables could also address challenges in user feedback from wearables by enabling direct audio communication with users rather than relying on screens or vibration [33].

Ear-worn transducers are located at the body part farthest from the point of application of ground reaction forces, which might appear counterintuitive to how they can be used to estimate GRFs. However, these transducers receive vibrations generated by foot impact that propagate through the body, and their movement can provide insights into overall centre of mass kinematics. Foot-to-centre-of-mass motion is often modelled using a damped mass-spring system [7, 9, 20, 24]; this could be extended to include the head with an additional damped mass-spring component in the absence of independent head rotations. A machine learning model may be capable of learning these dynamics in this more complex scenario, as

demonstrated with foot-to-centre-of-mass relationships e.g. in [1, 34] for the torso and legs predicting vGRFs. We also experimentally observe strong correlations in signal morphology from the head IMU to both sacrum IMU and GRFs showing clear links between the signals.

Therefore, we propose using ear-worn wearables, or earables, to monitor gait information whilst running. Whilst related research exists on using earables for gait tracking, the majority of studies focus on walking and clinical applications. For example, Jarchi et al. demonstrated detection of gait events for tracking temporal gait parameters such as cadence and double and single support times during walking [16]. This work was further developed in subsequent research that improved the signal processing algorithm, achieving better accuracy [10, 29]. Later, Atallah et al. also correlated vGRF peak force to ear-worn acceleration peaks; however, R^2 values were approximately 0.3 [2] for mixed user data, and error reporting was not included. A recent study, WalkEar [31], demonstrated how earables can estimate temporal and vGRF parameters during walking with improved correlation results of 0.6-0.9 compared to prior work and reported accuracies comparable to other inertial wearables. This research shows that earables have promise for ground reaction force monitoring, but currently no research on earables for running analysis is available in the literature.

We developed and validated a novel algorithm that represents the first to use IMU signals from earables (i.e., the open-source OpenEarable [28]) to enable biomechanical assessment of running. Specifically, we estimate running gait events (e.g., cadence, flight and stance times) and predict individual vertical ground reaction force waveforms and their scalar features (e.g., peak force, loading rate). We devised a new signal processing algorithm and a machine learning regression model with the objective of achieving performance similar to devices worn on the ankle or sacrum while being in a commonly owned form-factor.

Methods

Thirty healthy participants (18 male, 12 female) took part in the study, which was approved by the University of Cambridge Department of Computer Science and Technology ethics committee under application #2134. All participants signed informed consent forms before participating in any experiments and had to be free from current injury as part of the inclusion criteria. The overall demographics and anthropometrics (mean ± standard deviation) of the participants were an age of 28.2 ± 9.2 years, height of 176.2 ± 8.8 cm and mass of 73.8 ± 13.2 kg. Each participant wore their own trainers, sports clothing, and a pair of open-source OpenEarable v1.4 [28], with firmware customised for this project to enable the 50 Hz sampling rate. The newer OpenEarable v2.0 includes this feature in the base firmware among other improvements so our firmware is not available publicly. Each device weighed 12 g and contained a Bosch BMX160 9-axis IMU, the accelerometer and gyroscope were sampled at 50 Hz with a maximum range of ±8G and ±2000 degrees/second, respectively; this data was saved to an SD card on the device. The choice of 50 Hz was to replicate the typical limitations of consumer devices, which sample the IMU at lower sampling rates than research-grade devices due to battery and bandwidth constraints, we also performed a pre-study on 6 participants with sample rates of 10 – 1000 Hz with 50 Hz showing good accuracy on peak force reconstruction. The choice of ±8 G was informed by an experiment looking at maximum acceleration experienced by the ear-worn IMU during running, this rarely exceeded 4 G, so 8 G was chosen as the range. The ear-worn device has a form factor similar to a hearing aid with a larger part containing the IMU, computational and radio modules behind the ear and an earpiece fitted with a yellow foam eartip (Figure 1.c). Whilst the device reliably sits in the ear, it was secured with an elastic headband for safety (Figure 1.a).

After a self-directed warm-up, participants were asked to run at different speeds both on an instrumented treadmill (Bertec, OH, USA, 1000 Hz) and overground. Treadmill running was carried out with no belt incline and included bouts of 4 minutes running,

142 interspersed by at least 2 minutes recovery, at 10 km/h (common to all participants) and 2
143 other self-selected speeds (e.g., 11-12 km/h). In the overground condition, runners looped
144 around the lab for 3 minutes at a self-selected speed and, without targeting, stepped on four
145 900 × 600 mm Kistler force plates (mod 9667A0906, Kistler, Winterthur, Switzerland, 1000
146 Hz) when transitioning in the central part of the room (Figure 1.b). This condition was used
147 to create a higher variance of gait parameters in the dataset. At the start of each recording,
148 participants performed a series of foot stamps to allow the manual synchronisation of force
149 plates and ear-worn sensors by matching the initial series of peaks, to allow matching stance
150 phases from the different data streams.

151 Treadmill GRF was pre-processed with a 15 Hz Butterworth 5th-order low-pass filter,
152 with a forward and backward pass to avoid phase distortion but was kept at the original
153 sampling frequency (1000 Hz) after filtering. This filter was used to eliminate noise from
154 mechanical vibrations of the treadmill belt and structure as the gradient changing model can
155 suffer from vibrations [4]. Overground GRF was pre-processed with a 100 Hz low-pass filter
156 as they were fixed models, the higher sampling rate can assess generalisation of the
157 algorithms. Thresholds of 40 and 20 N were used for treadmill and overground conditions,
158 respectively, to identify individual contacts and the temporal metrics derived thereof (i.e.,
159 cadence, stride time, stance time, flight time).

160 An algorithm was developed to use accelerometer and gyroscope data from both ear-
161 worn devices (left and right ears) to: (1) detect heel strike and toe-off from each gait cycle;
162 (2) segment the IMU data into individual flight and stance periods; (3) calculate spatio-
163 temporal features, such as cadence, stride time, stance time, and flight time; and (4) predict
164 vGRF waveforms and scalar metrics, which included peak values, impulse, and loading rate,
165 as defined by the gradient between the point where the vGRF curve first exceeded 200 N to
166 90% of the impact peak, as used previously in running studies [21]. The ground truth peak

force was always taken as the active peak; this was detected by finding the highest force in the stance region which was active peak over the impact peak in every example that was checked.

The principle of the segmentation algorithm was finding heel strike and toe-off gait events. Both gait events had corresponding patterns in the head acceleration signals (Figure 2). The heel strike was detected from the onset of the impact peak, which was captured in the second derivative of the vertical acceleration signal as the first of two peaks that occurred during the loading phase. To improve the accuracy of the measurement, the left and right earbud channels were summed, which strengthened the trend in the acceleration signal, giving more prominent peaks. The toe-off was detected by first applying a low-pass filter with a 10 Hz cutoff to the acceleration signal, this was then differentiated to give the jerk signal; upon inspection of the jerk signal, an inflection point could be seen that corresponded with the toe-off (Figure 2). This could be detected by looking for a local minimum in the derivative of the jerk signal. These events were combined with a system that checked realistic timing, i.e., a flight phase that was not over 0.5 s or under 5 ms. While not all gait events were correctly detected, no manual edits were made to this process and the reported results are made only from the described algorithm.

vGRF estimation used machine learning regression models. First, discrete scalar parameters—peak force, loading rate, and impulse—were estimated with predictions made by the regression model in body weight units (BW), as the training data was normalised by the user's body weight. This was rationalised by the user only needing to input their weight into the system to obtain predictions in Newtons if desired. The scalars were estimated using Gaussian process regression using time and frequency domain features calculated over the segmented IMU data for that stance phase.

The scalar measures were then combined with the raw IMU signals to regress the vGRF curve. The model for this sequence-to-sequence regression was a fixed number of boosted tree models, one for each vGRF sample. Each model saw every IMU sample from the stance phase and the estimated scalars to predict a single output point; this was then repeated for each output time point. Finally, the predicted vGRF waveform was post-processed with the same low-pass filter used on the ground truth, i.e., 15 Hz for treadmill data. The number of output time samples must be the same for each stance phase due to the fixed size of the machine learning model; this results in uneven sampling rates as the stance period can vary in length between steps and participants. This length was configurable depending on the granularity of predicted vGRF desired, but we presented results for 100 time samples, which corresponded to between 400-500 Hz depending on the exact stance time. This was chosen as it was above the Nyquist criterion for the 15 Hz and 100 Hz low-pass filter cut-off frequencies used on the treadmill and force plate data, respectively, so would still capture all trends in the pre-processed ground truth. The overall information flow described here is shown in Figure 3, and the model presented here was derived from the model developed in [31] used for walking vGRF estimation from earables. Data from the treadmill or force plates is not included here as it was used for validation of the proposed system.

The treadmill allowed a large volume (~54,000) of step samples to be collected for model training, whilst the force data from plates with relatively fewer samples (~2,800) was used as ground truth to validate the algorithm in the more common overground running scenario.

We compared predicted quantities and ground truth using a selection of accuracy metrics and validation schemes; this allowed for comparison with a greater amount of related works. Scalar quantities were assessed in terms of Mean Absolute Error (MAE), Mean

Absolute Percentage Error (MAPE), and Pearson Correlation Coefficient (Corr). A Bland-Altman analysis with Reproducibility Coefficient (RPC) and 95% Limits of Agreement (LOA) was also presented. Time series estimation was analysed through Root Mean Square Error (RMSE), Normalised Root Mean Square Error (NRMSE), normalised as a percentage to the ground truth value on a sample-by-sample basis, and the Pearson Correlation Coefficient of each point.

The regression evaluation used three validation schemes; these schemes were only relevant for the kinetic scalar and vGRF curve results, as temporal parameters were estimated with signal processing techniques requiring no training data. The first scheme was Unseen 4:1, using a 4:1 train:test data split ratio. The test set was comprised of only unseen users; this ratio had 24 users in the training set and 6 users in the test set. The second was Leave One Subject Out (LOSO) validation, which trained on 29 users whilst leaving out a single user. Finally, we also presented a Mixed (4:1) validation scheme, which used a mixed batch of data in a 4:1 train:test ratio. This did not repeat gait cycles, but users could be repeated in the test and train sets. This was included to allow comparison with works that did the same. The mixed and unseen were validated on 5 randomly sampled 4:1 sets and the LOSO on all 30 possible combinations. For the overground evaluation, training data was in the Unseen 4:1 ratio, where training data was taken from the 24 training participants on both the treadmill and force plates, and the testing data was from the 6 unseen participants' force plate data.

Results

Errors consistently lower than 20 ms were found for temporal measures between the treadmill and ear-worn device, with a high correlation of predicted temporal gait metrics compared to the treadmill data, providing an IMU segmentation that tracked the stance phase for the kinetic parameter stage (Table 1). A Bland-Altman analysis showed low RPC for stride

time with higher percentage RPC for stance and flight times, indicating that toe-off detection was less accurate than heel strike detection. The Bland-Altman analysis also showed the 95% LOA had a similar trend per gait metric as seen for the RPC. A bias in stance and flight times of +0.01 and -0.01, respectively, was observed (Figure 4).

The kinetic scalar gait parameters showed MAPEs below 10% for peak force and impulse and below 20% for loading rate between ground truth and predictions (Table 2). The correlation was higher for the 4:1 unseen group and lower for the LOSO group. The RPC was low for both peak force and impulse but much higher for loading rate at 60% (Figure 4). Similarly, the LOA followed the same trend across the kinetic metrics (Figure 4).

vGRF estimation in the mixed evaluation scenario had an NRMSE of 7.6%, followed by the more realistic unseen scenario with an NRMSE of 11.1% and LOSO with 13.1% (Table 2). Different examples for individual vGRF curves showed differences in the tracking accuracy for different participants given for the lowest (Figure 5a), average (Figure 5b and c), and highest (Figure 5d) RMSE participants relative to the dataset. For the two participants with average RMSE, it could be seen that one example captured (Figure 5b) the impact peak but had poor tracking during and after the active peak, whilst the other (Figure 5c) missed the impact peak but had good tracking around and after the active peak. The comparison between the lowest and highest RMSE participants both showed a forefoot strike example where tracking was relatively good for the lowest (Figure 5a) or poor for the highest (Figure 5d).

The per-user error across the participants for temporal parameters varied between 3%-8% (Figure 6a). For kinetic scalars, most participants had a constant MAPE around 5% in the LOSO validation, but a few participants had much higher values near 15% (Figure 6b), the higher error was dominated by the loading rate error. The vGRF showed a similar pattern to the kinetic scalar with variations in NRMSE of 8%–19% across the participants in the LOSO validation (Figure 6c).

In the overground experiments, the errors increased on all predicted gait parameters compared to the treadmill scenario; this was a small increase for temporal parameters of below 1%, but for kinetic scalars and vGRF estimation this was up to 4% higher. This might have been caused by the higher variance in running style in the untargeted overground experiments compared to the training data taken from the more controlled treadmill experiments or could be due to a relative lack of model training samples with only 2800 overground steps compared to 54000 treadmill steps.

Table 1: Results for temporal gait parameters evaluated for all 30 participants on the treadmill data, totalling around 54000 step samples.

Metric	**Cadence**	**Stride Time**	**Stance Time**	**Flight Time**
MAE	3.5 step/min	8.2 ms	11.9 ms	11.9 ms
MAPE	2.2%	1.1%	4.49%	11.1%
Correlation	0.872	0.960	0.883	0.885

Table 2: Results for kinetic scalar parameters and vGRF under the three different validation conditions, units and evaluation metrics are specified in the second row. Evaluated for all 30 participants on the treadmill data, totalling around 54000 step samples.

Validation Scheme	**Peak Force**			**Loading Rate**			**Impulse**			**vGRF**		
	MAE BW	MAPE%	CORR	MAE BW/s	MAPE%	CORR	MAE BW.s	MAPE%	CORR	RMSE BW	NRMSE %	
Unseen 4:1	0.143	6.03%	0.890	0.028	18%	0.693	0.015	3.8%	0.639	0.154	11.14	

LOSO 29:1	0.145	6.38%	0.417	0.032	22%	0.323	0.016	4.2%	0.324	0.182	13.09	
Mixed 4:1	0.100	4.10%	0.901	0.025	14%	0.774	0.013	3.5%	0.716	0.112	7.56	

Discussion

In this study, we aimed to develop and test a system that harnessed the potential of inertial sensors embedded in consumer-oriented earbuds and fed their data into a predictive model that estimated key biomechanical quantities for running assessment. From the assessment we carried out through treadmill and overground running on 30 participants, the system demonstrated an overall accuracy of 4.8% on temporal metrics and 11.1% on vGRF curve estimation, which was comparable to what was achieved by purpose-made gait assessment systems, whilst offering advantages in terms of being deployed on a comfortable, commonly owned device that could also be used for entertainment.

The presented algorithm was lightweight, with signal processing and machine learning models shown to run in real time on an iPhone device [31]. Additionally, the model used for kinetic scalars in this study could have been simplified even further, as the Gaussian Process model had a longer inference period than techniques such as support vector regression.

In this study, we used a relatively low sampling rate of 50 Hz compared to many studies that conducted running gait assessment with inertial sensors, i.e., 1000 Hz [11] and 2000 Hz [1]. This was done to demonstrate the performance on typical consumer devices (i.e., Apple AirPods Pro (Apple Inc., Cupertino, CA, USA) 25 Hz) or modified Samsung Galaxy Buds [15] that would be used in real scenarios with the algorithms developed. Furthermore, we believe that the integration of running metrics into devices such as heart rate straps (Garmin Ltd., KS, USA) showcases a prior viable route for running gait analysis integration into non-purpose made device form factors.

Related works used purpose-made or research devices that might have required attachment to shoes or the torso, which could have represented additional cost to potential users

compared to earables that were commonly owned and served additional purposes. In terms of accuracy of these systems for temporal parameters, shoe-mounted IMU methods had achieved MAEs of 8 ms with signal processing [6] and 15 ms with machine learning [12] compared to our MAE of 12 ms. For a sacrum-mounted IMU using machine learning, MAPEs of 9.3% and 3.6% [25] were obtained for flight and stance times compared to MAPEs of 11.1% and 4.5% for our system. Both comparisons showed that an earable device obtained comparable temporal metric accuracy against devices that would have been purpose-made for gait assessment.

For vGRF curve prediction, we reported an NRMSE of 11%. For related works using IMUs on the sacrum and shoes, vGRF curve prediction was reported as NRMSEs of 6.5% [1] and 6.1–9.1% [11], and insole-based systems reported between 0.8 and 8.8% [4], all using machine learning models. For kinetic scalars, a study with a sacrum-mounted device tested the peak force, loading rate, and impulse [1], obtaining MAPEs of 8.8%, 27.6%, and 6.4%, respectively, compared to our results of 6.0%, 18%, and 3.8%. This showed that for vGRF information, the earable device was also close to the purpose-made devices despite using more lightweight techniques and having a more complex dynamic relationship to capture due to the neck and head. It was likely that through a combination of more powerful modelling approaches, such as an LSTM [1, 4], and a larger dataset size, the results reported here could have been improved upon. However, system constraints on power, computational power, and latency would have had to be considered based upon the application and processing devices available if a larger model were to have been used.

A bias could be seen in flight and stance time Bland-Altman plots (Figure 4), which might have been caused by the threshold used to segment the treadmill ground truth. The threshold could lead to an underestimate of the stance time, which would explain the overestimate in the earable predictions. This could be seen visually in an example in Figure 2 as well as in the trend in Bland-Altman plots (Figure 4), so it was likely that the temporal metric

329 performance was better than reported due to the described error in the treadmill timings. This
330 could be investigated in future work using other validation systems such as motion capture or
331 hand labelling of the force plate data.

332 The reported error for temporal gait metrics was close to the best case for a sampling
333 rate of 50 Hz. This "best case" could be estimated by assuming exact gait events were randomly
334 distributed uniformly within the sensor sampling interval. Given that each interval was 20 ms
335 apart and that the algorithm always predicted the closest sample, the highest and lowest errors
336 were 10 ms and 0 ms, with an average error of 5 ms. Thus, on average, the MAE could be at a
337 minimum of 5 ms using a 50 Hz sensor with no interpolation or upsampling, assuming a perfect
338 gait event detection algorithm. Given that 5 ms was the best achievable result for gait event
339 detection with a 50 Hz sampling rate, the reported MAEs (Table 1) showed that the earable
340 gait event detection algorithm was close to this value, demonstrating accurate detection of gait
341 events.

342 To assess the participants who had lower accuracy in the kinetic scalar estimation
343 (Figure 6), it was found that participants 6 and 11 had the largest peak force values in the
344 dataset, over 3 BW. When the model was trained in the LOSO setting, it only saw a small
345 amount of data of this magnitude and was therefore less likely to make accurate predictions.
346 Participants 18 and 27 had a similar case but with the smallest peak forces, typically below 2
347 BW. The increase in peak force also affected the loading rate and impulse values. For the vGRF
348 prediction, the same participants had higher errors compared to the kinetic scalars, where the
349 same reasoning applied. Scaling up to more participants might have improved the performance
350 on users who were outliers in the current participant population. Another cause of cross
351 participant variation may be caused by decoupling of a device from the head. In most cases the
352 left and right earable IMU signals were similar in morphology. However, in a few experiments
353 one device showed more high frequency energy which may indicate a looser coupling with the

head in those experiments. This may result in lower accuracy predictions as there is high noise content in the signal.

In the kinetic scalar estimation, the loading rate had the worst RPC and highest MAPE, showing it was the most poorly estimated parameter. This could have been because the features used were less relevant to predicting the loading rate compared to the other parameters, as the features were calculated over the whole stance phase rather than just the loading period. The correlations were higher in the unseen validation compared to the LOSO validation across all three estimated parameters (Table 2). This could have implied that the magnitude of the parameters was estimated well, but the method was less sensitive to the trends within a single experiment or user.

A limitation of this study was head motions, such as rotating the head to talk to a friend; these would register high accelerometer values on the earables that were not related to gait, potentially affecting gait event detection or regression predictions. Whilst this was not tested for in this study, participants were not instructed to keep their head still, so the reported results included any errors incurred through head motions made during the data collection process. In other work, this was mitigated by the gyroscope signal, which was large in magnitude during head rotations compared to gait due to the orders of magnitude difference in radius of rotation. This allowed easy detection of head motion to remove these segments of the data entirely [31]; however, a more advanced compensation scheme could have been developed using the gyroscope to remove the head motion-related acceleration components.

A further limitation of this study is that the footstrike type for the participants was not systematically recorded. However, we believe that the model has learnt to be agnostic to footstrike type given the correct vGRF morphologies shown in Figure 5. There are also prior studies that show IMU data can be used to classify footstrike type with high accuracy, e.g. [30],

which gives further confidence that a model can learn important information about footstrike from IMU data.

Future development on this work could explore estimation of further biomechanical parameters in running from the head, examples include tibial bone force [23] and ground reaction moments [18], both of which have been estimated with wearable devices. Another future direction could incorporate the algorithms developed here onto fully mobile devices such as the earbuds themselves or an earbud – phone pairing to test the performance in the field.

Conclusion

Overall, we demonstrated for the first time the promise of using consumer ear-worn wearables to perform gait timing and loading analysis whilst running with algorithms that are lightweight enough to operate on a mobile phone and possibly the earbuds themselves with further development. We presented algorithms to estimate temporal gait metrics, which were then used to segment and estimate kinetic parameters and the vGRF curve, showing comparable accuracy to purpose-made gait assessment devices that are already popular with consumers. This presents a valuable tool for a range of parties, from amateur runners requiring fewer devices to track key running metrics, to coaches for continuously monitoring performance and elite athletes for real-time feedback on running gait without distractions from screens. This work opens up future research directions on long-term monitoring of running gait to enable research to assess causal relationships with performance and injuries, as well as for tracking athlete rehabilitation and mobile system-based challenges on optimising the system performance for earable-based running gait assessment.

Acknowledgements

This work was supported by EPSRC grants EP/Z53447X/1 and EP/S023046/1 for the EPSRC CDT in Sensor Technologies and Applications as well as EVIDEN/New Atos Life Sciences Centre of Excellence.

References

[1] Alcantara RS, Edwards WB, Millet GY, Grabowski AM. Predicting continuous ground reaction forces from accelerometers during uphill and downhill running: a recurrent neural network solution. PeerJ. 2022;10:e12752. doi:10.7717/peerj.12752

[2] Atallah L, Wiik A, Jones GG, et al. Validation of an ear-worn sensor for gait monitoring using a force-plate instrumented treadmill. Gait & Posture. 2012;35(4):674-676. doi:10.1016/j.gaitpost.2011.11.021

[3] Bobbert, M. F., Schamhardt, H. C., & Nigg, B. M. (1991). Calculation of vertical ground reaction force estimates during running from positional data. Journal of biomechanics, 24(12), 1095-1105.

[4] Carter J, Chen X, Cazzola D, Trewartha G, Preatoni E. Consumer-priced wearable sensors combined with deep learning can be used to accurately predict ground reaction forces during various treadmill running conditions. PeerJ. 2024;12:e17896. doi:10.7717/peerj.17896

[5] Ceyssens L, Vanelderen R, Barton C, Malliaras P, Dingenen B. Biomechanical Risk Factors Associated with Running-Related Injuries: A Systematic Review. Sports Med. 2019;49(7):1095-1115. doi:10.1007/s40279-019-01110-z

[6] Chew DK, Ngoh KJH, Gouwanda D, Gopalai AA. Estimating running spatial and temporal parameters using an inertial sensor. Sports Eng. 2018;21(2):115-122. doi:10.1007/s12283-017-0255-9

[7] Clark, K. P., Ryan, L. J., & Weyand, P. G. (2017). A general relationship links gait mechanics and running ground reaction forces. Journal of Experimental Biology, 220(2), 247-258.

[8] Clermont CA, Duffett-Leger L, Hettinga BA, Ferber R. Runners' Perspectives on 'Smart' Wearable Technology and Its Use for Preventing Injury. International Journal of Human–Computer Interaction. 2020;36(1):31-40. doi:10.1080/10447318.2019.1597575

[9] Derrick, T. R., Caldwell, G. E., & Hamill, J. (2000). Modeling the stiffness characteristics of the human body while running with various stride lengths. Journal of Applied Biomechanics, 16(1), 36-51.

[10] Diao Y, Ma Y, Xu D, Chen W, Wang Y. A novel gait parameter estimation method for healthy adults and postoperative patients with an ear-worn sensor. Physiol Meas. 2020;41(5):05NT01. doi:10.1088/1361-6579/ab87b5

[11] Dorschky E, Nitschke M, Martindale CF, van den Bogert AJ, Koelewijn AD, Eskofier BM. CNN-Based Estimation of Sagittal Plane Walking and Running Biomechanics From

Measured and Simulated Inertial Sensor Data. Front Bioeng Biotechnol. 2020;8. doi:10.3389/fbioe.2020.00604

[12] Falbriard M, Meyer F, Mariani B, Millet GP, Aminian K. Accurate Estimation of Running Temporal Parameters Using Foot-Worn Inertial Sensors. Front Physiol. 2018;9. doi:10.3389/fphys.2018.00610

[13] Globally, how many people practice running (as a sport), and what is the global market size for this sport? Wonder. Accessed June 7, 2025. https://askwonder.com/research/globally-people-practice-running-as-sport-global-market-size-sport-dba00dqvc

[14] How often do you run in comparison to other Runner's World readers? And how does your "easy run" pace compare? Runner's World. September 20, 2023. Accessed May 29, 2025. https://www.runnersworld.com/uk/news/a45036050/runners-world-running-survey-2023/

[15] Islam, Md Saiful, et al. "BallistoBud: Heart Rate Variability Monitoring using Earbud Accelerometry for Stress Assessment." Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025.

[16] Jarchi D, Wong C, Kwasnicki RM, Heller B, Tew GA, Yang GZ. Gait Parameter Estimation From a Miniaturized Ear-Worn Sensor Using Singular Spectrum Analysis and Longest Common Subsequence. IEEE Transactions on Biomedical Engineering. 2014;61(4):1261-1273. doi:10.1109/TBME.2014.2299772

[17] Jiang X, Napier C, Hannigan B, Eng JJ, Menon C. Estimating Vertical Ground Reaction Force during Walking Using a Single Inertial Sensor. Sensors. 2020;20(15):4345. doi:10.3390/s20154345

[18] Johnson, W. R., Mian, A., Robinson, M. A., Verheul, J., Lloyd, D. G., & Alderson, J. A. (2020). Multidimensional ground reaction forces and moments from wearable sensor accelerations via deep learning. IEEE Transactions on Biomedical Engineering, 68(1), 289-297.

[19] Kluge F, Brand YE, Micó-Amigo ME, et al. Real-World Gait Detection Using a Wrist-Worn Inertial Sensor: Validation Study. JMIR Formative Research. 2024;8(1):e50035. doi:10.2196/50035

[20] LeBlanc B, Hernandez EM, McGinnis RS, Gurchiek RD. Continuous Estimation of Ground Reaction Force During Long Distance Running Within a Fatigue Monitoring Framework: A Kalman Filter-Based Model-Data Fusion Approach. Journal of Biomechanics. 2021;115:110130. doi:10.1016/j.jbiomech.2020.110130

[21] Lieberman DE, Venkadesan M, Werbel WA, et al. Foot strike patterns and collision forces in habitually barefoot versus shod runners. Nature. 2010;463(7280):531-535. doi:10.1038/nature08723

[22] Matijevich, E. S., Branscombe, L. M., Scott, L. R., & Zelik, K. E. (2019). Ground reaction force metrics are not strongly correlated with tibial bone load when running across

speeds and slopes: Implications for science, sport and wearable tech. PloS one, 14(1), e0210000.

[23] Matijevich, Emily S., et al. "Combining wearable sensor signals, machine learning and biomechanics to estimate tibial bone force and damage during running." Human movement science 74 (2020): 102690.

[24] Nedergaard, N. J., Verheul, J., Drust, B., Etchells, T., Lisboa, P., Robinson, M. A., & Vanrenterghem, J. (2018). The feasibility of predicting ground reaction forces during running from a trunk accelerometry driven mass-spring-damper model. PeerJ, 6, e6105.

[25] Patoz A, Lussiana ,Thibault, Breine ,Bastiaan, Gindre ,Cyrille, and Malatesta D. Comparison of different machine learning models to enhance sacral acceleration-based estimations of running stride temporal variables and peak vertical ground reaction force. Sports Biomechanics. 0(0):1-17. doi:10.1080/14763141.2022.2159870

[26] Preatoni E, Bergamini E, Fantozzi S, et al. The Use of Wearable Sensors for Preventing, Assessing, and Informing Recovery from Sport-Related Musculoskeletal Injuries: A Systematic Scoping Review. Sensors. 2022;22(9):3225. doi:10.3390/s22093225

[27] Provot T, Choupani S, Bourgain M, Valdes-Tamayo L, Chadefaux D. Using Wearable Accelerometers to Develop a Vertical Ground Reaction Force Prediction Model during Running: A Sensitivity Study. Vibration. 2023;6(3):680-694. doi:10.3390/vibration6030042

[28] Röddiger T, Stuchbury-Wass J, Ciliberto M, Lepold P, Beigl M. OpenEarable 1.4: Dual Microphones Earpiece to Capture In-Ear and Outer-Ear Audio Signals. In: Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. UbiComp '24. Association for Computing Machinery; 2024:930-933. doi:10.1145/3675094.3678483

[29] Seifer AK, Dorschky E, Küderle A, Moradi H, Hannemann R, Eskofier BM. EarGait: Estimation of Temporal Gait Parameters from Hearing Aid Integrated Inertial Sensors. Sensors. 2023;23(14):6565. doi:10.3390/s23146565

[30] Shiang, T. Y., Hsieh, T. Y., Lee, Y. S., Wu, C. C., Yu, M. C., Mei, C. H., & Tai, I. H. (2016). Determine the foot strike pattern using inertial sensors. Journal of Sensors, 2016(1), 4759626.

[31] Stuchbury-Wass J, Liu Y, Butkow KJ, et al. WalkEar: Holistic Gait Monitoring using Earables. In: 2025 IEEE International Conference on Pervasive Computing and Communications (PerCom). 2025:200-209. doi:

[32] van der Worp H, Vrielink JW, Bredeweg SW. Do runners who suffer injuries have higher vertical ground reaction forces than those who remain injury-free? A systematic review and meta-analysis. Br J Sports Med. 2016;50(8):450-457. doi:10.1136/bjsports-2015-094924

[33] Van Hooren B, Goudsmit J, Restrepo J, Vos S. Real-time feedback by wearables in running: Current approaches, challenges and suggestions for improvements. Journal of Sports Sciences. 2020;38(2):214-230. doi:10.1080/02640414.2019.1690960

[34] Zhu Y, Xia D, Zhang H. Using Wearable Sensors to Estimate Vertical Ground Reaction Force Based on a Transformer. Applied Sciences. 2023;13(4):2136. doi:10.3390/app13042136

Figure Captions

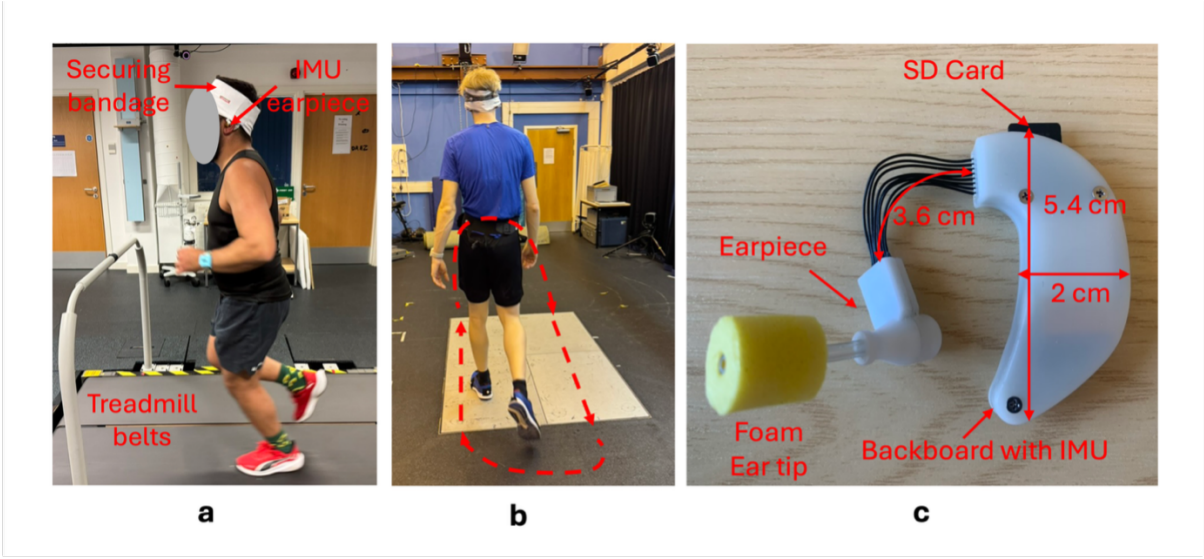

Figure 1 – Experimental setup showing (a) a participant running on a single instrumented treadmill whilst wearing the OpenEarable IMU earpiece secured with a bandage. (b) a participant walking over the footplates with a rough travel path annotated. (c) the OpenEarable 1.4 device with key dimensions annotated.

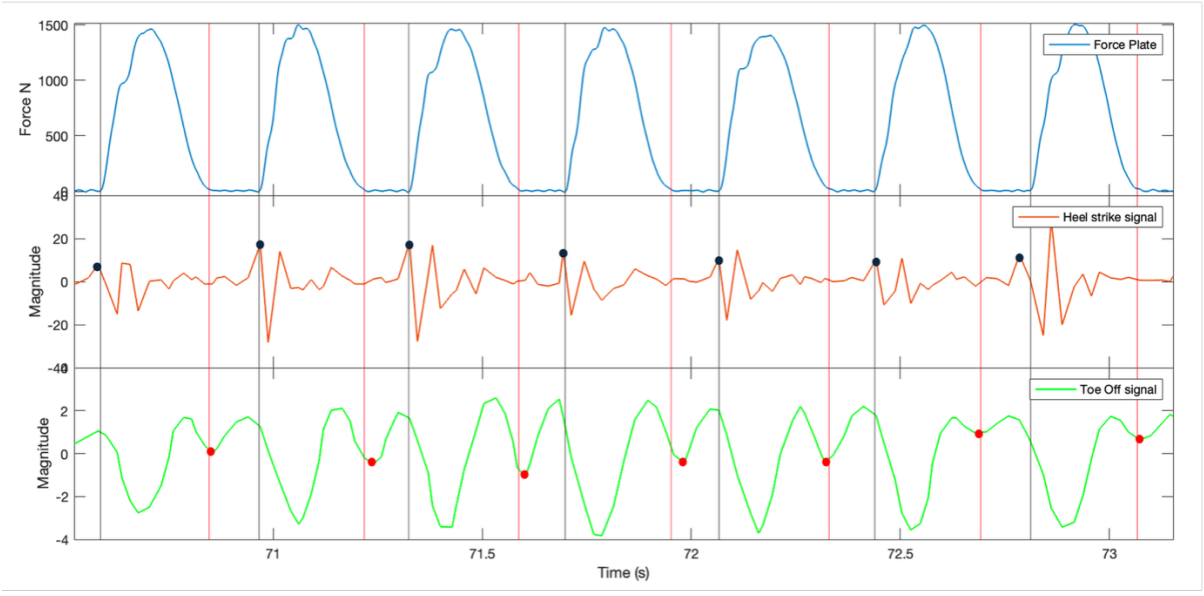

Figure 2 – Example of the head accelerometer gait event detection algorithm with a comparison to the treadmill data (top). Heel strikes are shown in a black line as ground truth

and a black dot as the estimation, both can be seen with the IMU second derivative signal (middle). The toe-off is shown as a red line for ground truth and a red dot for estimation results, both can be seen in the filtered IMU second derivative signal (bottom).

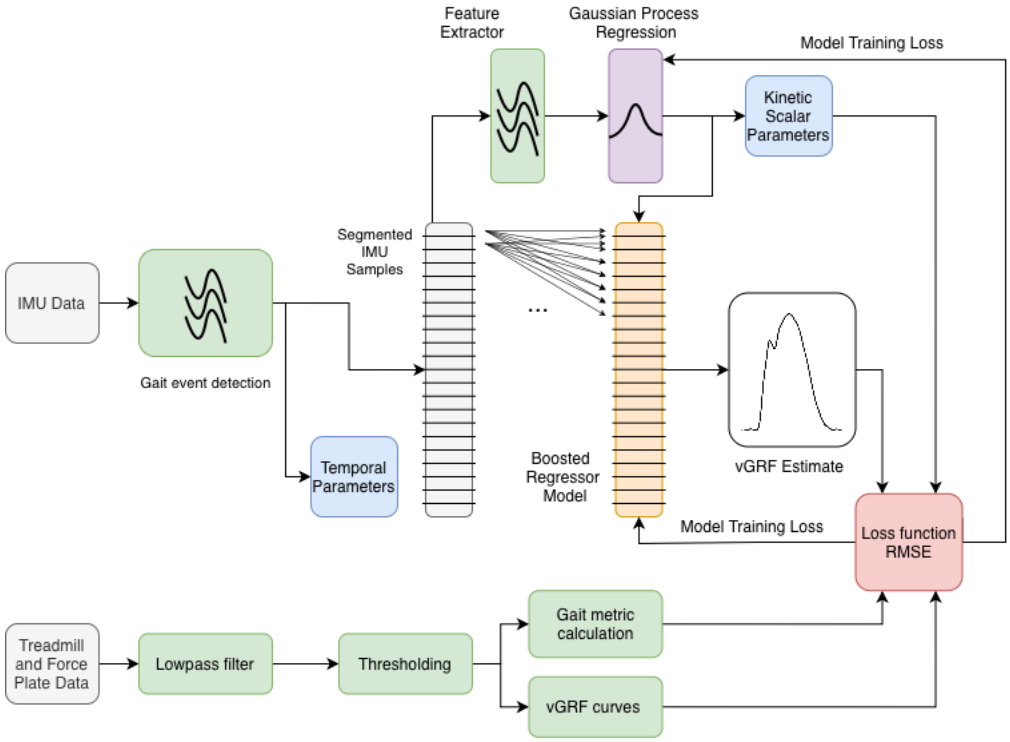

Figure 3 – Block diagram showing the flow of information from the ear-worn IMU to predict the gait parameters and reconstruct the vGRF curve. Each of the three kinetic scalars is regressed using the segmented IMU data then fed into the boosted regressor models which use both the kinetic scalars and IMU data as an input to regress the vGRF curve. The output is 100 time samples over the stance period making up the vGRF estimate. The ground truth data from the treadmill and force plates are shown with preprocessing steps as well as input to a loss function to train the Gaussian process and vGRF regression models.

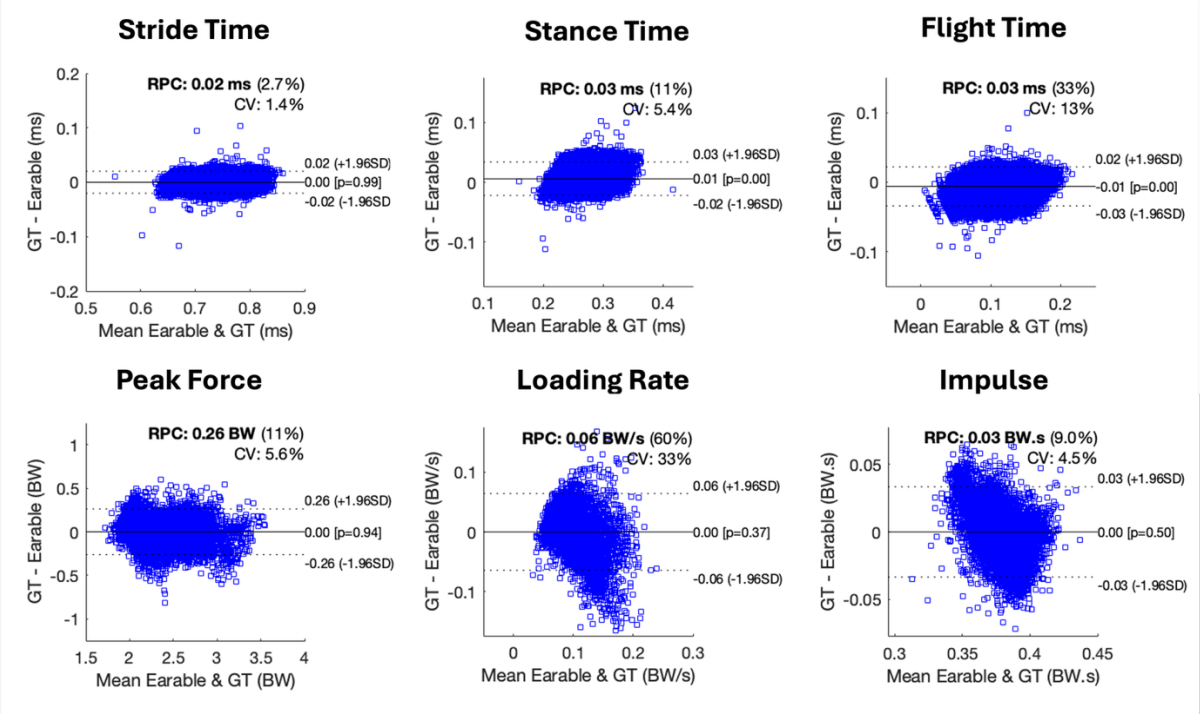

Figure 4 – Six Bland-Altman diagrams of temporal (top) and kinetic (bottom) gait parameters. Showing ±1.96 (95%) Limits of Agreement (LOA) and the mean difference with a p value for a t-test to indicate significance. The diagrams also show Reproducibility Coefficient (RPC) and its percentage value of the quantity as well as the Coefficient of Variation (CV), the standard deviation of mean values is shown. The data shown includes all 30 participants' treadmill data amounting to around 54000 samples for temporal parameters and 10800 samples (one of 5 folds with 6 unseen participants) for kinetic parameters.

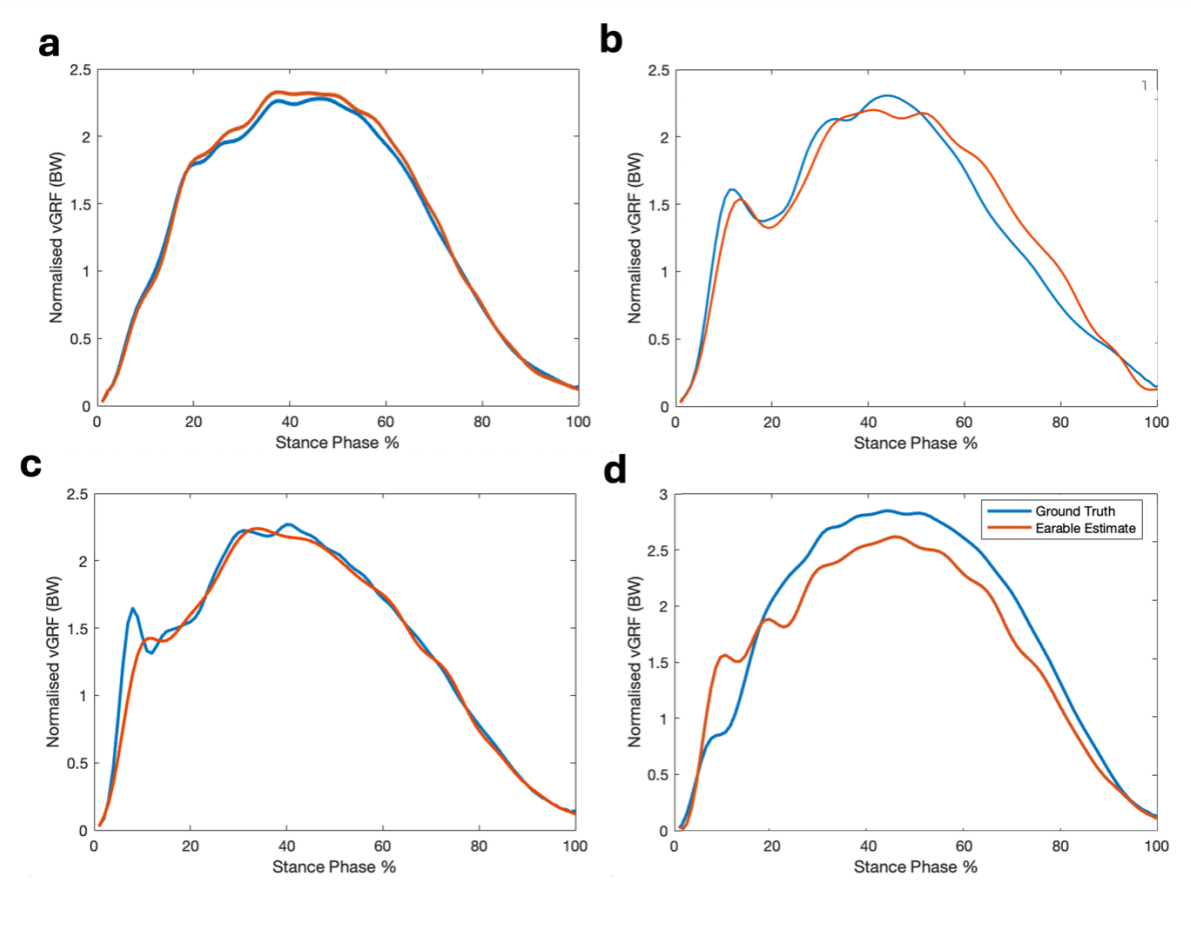

Figure 5 – Four example vGRF curve predictions (orange) with treadmill ground truth shown (blue). Showing a closely tracked forefoot strike from a low error participant who was an experienced runner (a). A correctly predicted impact peak but poor tracking of the peak and unloading phase from an average error participant (b). A missed impact peak in a prediction from an average error participant (c). A poor tracking example from the highest error participant (d).

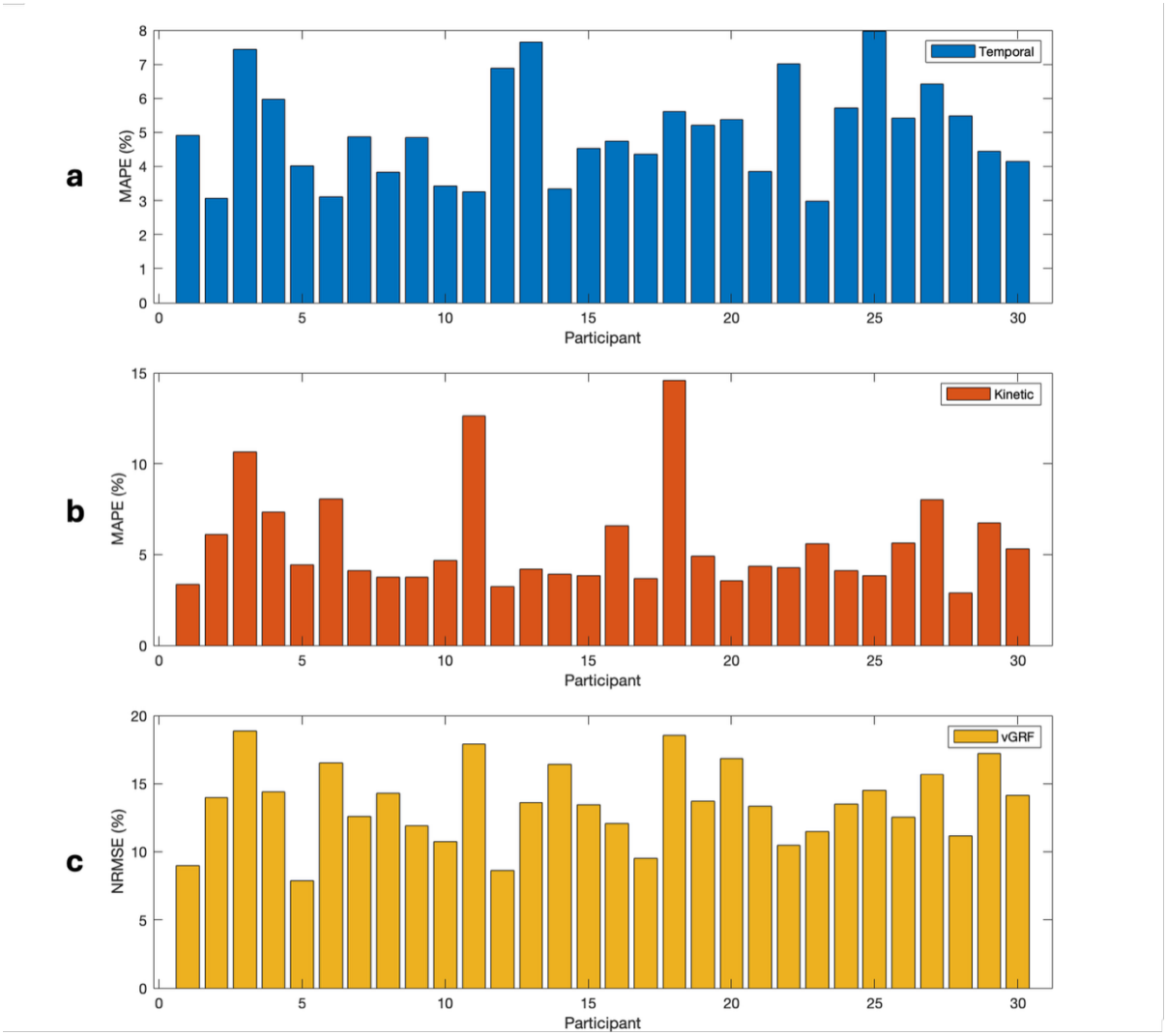

Figure 6 – Per Participant error bar chart showing MAPE for temporal (a), kinetic scalars (b) and NRMSE for vGRF curves (c).