

MSc Biomedical Engineering
Thesis

**No Extra Sensors Needed:
Estimating Vertical Ground
Reaction Force Using
Wearables Runners Already
Use**

Marijn Binnekamp

Supervisors:

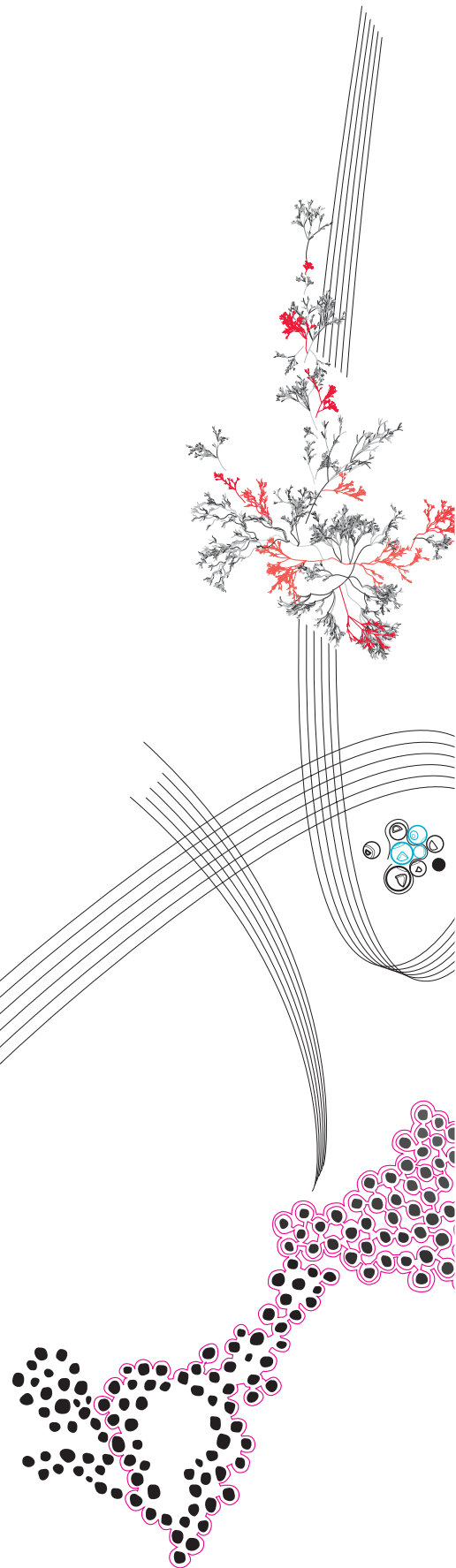
dr. J. Reenalda

dr.ir. F.J. Wouda

dr. E.H.F. van Asseldonk

June, 2025

Department of Biomedical Signals and Systems
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente



No Extra Sensors Needed: Estimating Vertical Ground Reaction Force Using Wearables Runners Already Use

M. Binnekamp

June 17, 2025

Abstract

Understanding the factors contributing to running-related injuries is of high importance, as injuries are common among runners. Ground reaction forces (GRFs) are potentially useful in quantifying the biomechanical load during running, but can only be measured reliably in a gait laboratory using force plates. To study GRF continuously outdoors, wearable sensors such as inertial measurement units (IMUs) can be used. These IMUs are also present in smartwatches and advanced heart rate monitor belts. This study aimed to predict vertical GRF (vGRF) using IMUs positioned at the wrist and sternum, mimicking smartwatch and heart rate strap placement, combined with smartwatch-derived variables. Twelve rearfoot strike runners ran on a forceplate instrumented treadmill at four speeds (8, 10, 12, 14 km/h) and three cadences (preferred $\pm 10\%$) across twelve trials of 90 seconds. Participants wore IMUs on the wrist and sternum (capturing 3D accelerations and angular velocities) and pressure insoles, providing vGRF estimates, which were used for outdoor validation. Following indoor trials, participants performed four rounds on an athletics track at the same four speeds. A long short-term memory neural network was trained to predict vGRF using IMU signals and smartwatch-derived variables. Feature importance analysis revealed that the sternum IMU was the most informative predictor. Leave-one-subject-out cross-validation resulted in an RMSE of 0.10 ± 0.024 Body Weight. Outdoor validation using pressure insoles did not show significant differences in model performance between indoor and outdoor trials. These findings suggest that runners may not require additional sensors to continuously gain meaningful insights into their vertical ground reaction forces in an outdoor setting.

1 Introduction

Running-related injuries (RRI) are a widespread problem among both recreational and professional runners [1, 2]. Despite extensive research, the underlying biomechanical factors contributing to RRI remain inconclusive [3], making injury prediction and thus prevention challenging. Among these factors, the ground reaction force (GRF) has been studied extensively as it represents the force exchanged between the runner and the ground. Of the three GRF components (vertical, anterior-posterior and medial-lateral), the vertical component (vGRF) is most commonly analysed due to its relatively large magnitude and biomechanical relevance during running. The vGRF is also assumed to be closely linked to internal loading of the musculoskeletal system [4, 5]. Although only low to moderate evidence is presented for higher vGRF (metrics) in relation to RRI [3, 6, 7], most research on vGRF characteristics has not been conducted longitudinally and in an outdoor setting. Consequently, our understanding of how vGRF characteristics relate to injury risk under real-world, long-term running conditions remains limited, highlighting an important need for longitudinal outdoor studies.

A longitudinal approach to GRF measurement is crucial to understanding the relationship between cumulative biomechanical load and RRI [8]. Edwards et al. modelled overuse injuries as a mechanical fatigue phenomenon, where repetitive loads on biological tissue show consistent results with mechanical fatigue phenomena such as damage accumulation and structural failure [9]. By weighting the loading magnitude, differences between the loading magnitude and the loading rate can be accounted for. This way, the cumulative load of one or multiple training sessions can be calculated. This model was tested in a study that implemented it for nine collegiate track athletes using a wearable sensor fixed to their right lateral iliac crest [10]. Estimations of the peak vGRF were used as load variable, demonstrating significantly different loading profiles between injured and non-injured runners. Although this study provided valuable insights, it was limited by its focus on a highly homogeneous group of participants and by considering only peak vGRF. Expanding on this by estimating full GRF waveforms could offer a better understanding of the biomechanical load. Implementing such an approach with user-friendly measurement devices would further enable longitudinal studies to investigate the relationship between GRF and RRI.

The shape of the vGRF curve during running depends on the runner’s footstrike pattern. Rearfoot (heel) strikers typically exhibit a double-peaked vGRF curve, whereas midfoot and forefoot strikers show a single-peaked pattern [11]. Most long-distance runners have a rearfoot strike pattern [12, 13]. In rearfoot strike patterns, the vGRF curve is composed of two distinct peaks: the first, known as the passive peak, corresponds to the initial contact with the ground, while the second, active peak, reflects the propulsive phase. Key features derived from this curve, such as peak vGRF, vertical loading rate and passive peak height, are frequently studied for their association with injury risk and running performance [14].

The gold standard for measuring GRF is using force plates embedded in the ground or within instrumented treadmills [15]. However, conducting longitudinal measurements in a laboratory is constrained by time and capacity limitations [15]. With the continuous development of wearable technologies, it is possible to move to an environment outside of the lab. Multiple studies have estimated GRF waveforms using inertial measurement units (IMUs) [16, 17]. Scheltinga et al. reported that GRF estimates had good repeatability on different days [18]. However, practical application of IMUs in real-life settings is challenging due to the need for precise placement [19], calibration and additional effort required by runners. Lacey et al. found that runners prefer user-friendly systems that fit their existing usage habits [20]. Fortunately, modern smartwatches and heart rate straps are equipped with integrated IMUs, creating a valuable opportunity: Instead of requiring runners to wear additional sensors, GRF can potentially be estimated using devices already used during training. Minimising the need for extra equipment is an essential step toward making biomechanical analysis more accessible and accepted in real-world settings. Using only the smartwatch for GRF estimation would be particularly valuable, since most runners already wear one during training [21]. To our knowledge, no study has explored the feasibility of wrist-only GRF prediction. In addition, it is shown that by using smartwatch-derived variables, such as speed and cadence, it is possible to estimate certain GRF metrics such as the cumulative peak vGRF and braking impulse [22].

Newton’s second law provides a physics-based method for estimating vGRF. However, its practical implementation depends on accurately capturing the acceleration of the body’s centre of mass (CoM) and the assumption that the human body can be modelled as a system of one or multiple rigid bodies. While the sternum is sometimes used as a proxy for the CoM, the study of Scheltinga et al. demonstrated that this location results in substantially higher prediction errors compared to using differently placed sensors [23]. This finding underscores the limitations of using the sternum for physical modelling. Moreover, IMUs placed on the wrist have no meaningful physical relationship

with CoM motion, limiting a physics-based approach. This highlights the need for using different approaches to estimate vGRF using the sternum and wrist IMU. Previous studies have successfully used recurrent Neural Networks and Long Short-Term Memory (LSTM) networks to predict GRF using IMU data [16, 17, 24, 25]. These studies reported RMSE values ranging from 0.16 to 0.27 body weight (BW), highlighting the ability of an LSTM model to effectively analyse and process sequential biomechanical data [26].

Ensemble learning is an increasingly popular approach for improving the predictive performance and robustness of machine learning models [27]. Ensemble learning combines multiple baseline models, referred to as ensemble members, together to create one unified ensemble model aiming to outperform individual ensemble members and prevent overfitting [27]. Ensemble learning is emerging within biomechanical modelling, with one recent study implementing it on GRF estimation [16]. In that study, seven distinct train-validation splits were used to train seven separate models for each test subject. By averaging the estimations of these models, an ensemble estimate was created. This approach reduced the bias towards the validation subjects, thus improving the generalisability of the model. The ensemble models were shown to consistently outperform the best individual ensemble members.

Although the model is initially aimed to be trained for indoor treadmill running, it is essential to address differences in kinetics and kinematics between treadmill running and overground outdoor running [28, 29]. Outdoor validation is therefore performed using pressure insoles, which estimate vGRF based on the applied pressure and can be used in both indoor and outdoor settings, providing an outdoor alternative to force plates.

This paper proposes a novel approach to GRF estimation that minimises the burden on the user by leveraging data from devices that most runners already wear. Specifically, it introduces a method that combines smartwatch-derived variables with IMUs positioned on the wrist and sternum, reflecting the configuration of commercially available wearables. Using LSTM networks, the model estimates the full vGRF waveform, rather than just peak values or summary metrics, enabling more comprehensive biomechanical analysis. Ensemble learning is implemented to enhance model performance and robustness. Outdoor validation is needed to show good generalisability to outdoor environments. This approach is a step toward enabling the longitudinal study of vGRF (characteristics) and their relationship to injury risk in a real-world environment.

2 Method

2.1 Participants

Twelve recreational distance runners (ten males, two females, body mass 79.4 ± 10.7 kg, height 190.9 ± 10.9 cm, age 25.6 ± 4.9) participated. Inclusion criteria were: rearfoot-strike running pattern, running ≥ 15 km/week over the past six months without injury, and the ability to complete a 400m lap at 14 km/h. All participants provided written informed consent. The protocol was approved by the University of Twente Ethics Committee (Application 250077).

2.2 Measurement setup

Pressure sensing insoles (Moticon, Munich, Germany) were placed in each participant's own running shoes, sampling at 100 Hz.

As part of a larger project, participants were equipped with seventeen IMUs following the full-body configuration of the Xsens MVN Link system (Movella, Enschede, The Netherlands). Only the sternum and wrist IMUs were used for this study. These IMUs, with a sampling frequency of 240 Hz, were attached to the body using double-sided tape and secured with additional tape to minimise soft-tissue artefacts. A sensor-to-segment calibration was performed per manufacturer's instructions.

Indoor trials were conducted on a dual-belt treadmill instrumented with force plates (Bertec Co., Columbus, OH, USA), which recorded 3D ground GRFs at 1200 Hz. Participants ran on one side of the treadmill. A schematic overview is shown in Figure 1. The outdoor protocol was performed on a tartan athletics track.

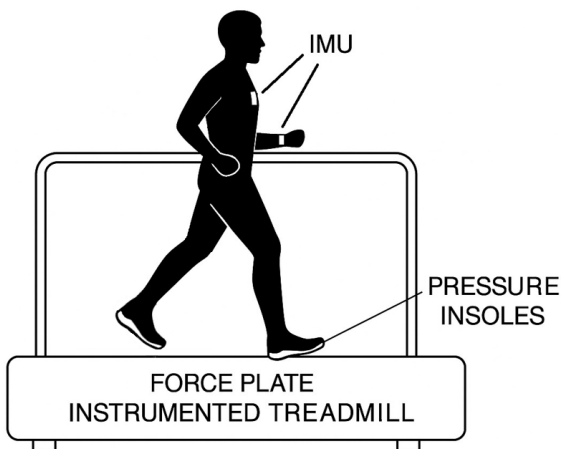


Figure 1: Schematic overview of the indoor measurement setup. In this overview, only the wrist and sternum IMU are shown as these are used for this research. vGRF is measured by the force plate instrumented treadmill and is estimated by the pressure insoles.

2.3 Data collection

The running protocol consisted of indoor and outdoor parts. During indoor trials, pressure insoles, IMUs, and an instrumented treadmill were used, whereas for outdoor trials, only insoles and IMUs were used.

2.3.1 Indoor

Body mass was determined via a static trial on the treadmill. Before and after each trial, participants jumped three times for temporal synchronisation of the pressure insoles, IMUs and force plates.

Participants then completed a 3-minute warm-up at a self-selected pace to familiarise themselves with the treadmill and measurement devices.

Twelve running trials of 90 seconds each were conducted in a randomised order at four different speeds (8, 10, 12, and 14 km/h). For each speed, trials were performed at the participant's preferred cadence, as well as at $\pm 10\%$ of it. Preferred cadence was determined during the first test at each speed by manually counting the steps. In the other trials, cadence was imposed using a metronome. A two-minute rest was provided after each trial to minimise fatigue.

2.3.2 Outdoor

After indoor trials, participants proceeded to the outdoor athletics track (400 m tartan running surface) at the University of Twente. As in the indoor protocol, participants performed three vertical jumps for temporal synchronisation before and after each trial.

Participants completed four laps at randomised speeds (8, 10, 12, and 14 km/h), with two minutes of rest after each trial. Speed was controlled by a cyclist pacing two meters ahead.

2.4 Data processing

Measured GRF (mGRF) and pressure insole data were low-pass filtered using a fourth-order zero-lag Butterworth filter (30 Hz cutoff). Both signals were resampled to 240 Hz to match the IMU data. IMU data included 3D accelerations (m/s^2) and angular velocities (deg/s). Temporal alignment was done using the data of the three jumps, by subtracting the lag corresponding to the maximum of the cross-correlation between the vertical acceleration of the sternum IMU and the vGRF. Data was segmented into individual steps. The stance phase is defined as the vGRF exceeding 30N. mGRF and pressure insole data were normalised by body weight. For each trial, 100 steps were analysed.

Step length and ground contact time (GCT) were determined for each step. GCT time was defined as the total stance phase duration per step. Step length was determined by multiplying the time between

consecutive steps by the speed of that trial. Cadence was determined across the whole measurement. A schematic overview of the pre-processing pipeline is shown in Figure 2.

2.5 Machine learning model

Machine learning models were created in MATLAB (R2023b, MathWorks Inc., Natick, MA, USA) using the Deep Learning Toolbox. The input matrix consisted of all sensor-free accelerations and angular velocities of both IMUs during the stance phase, combined with step-specific variables (speed, cadence, GCT and step length). The spatio-temporal features are constant for an individual step, so they were repeated across the time dimension to match the IMU sequence length.

The input per step had 16 features, forming a [sequence length \times 16] matrix. All features were normalised using z-score normalisation across the training set. A bidirectional LSTM layer with 256 hidden units processed the input, followed by a fully connected regression layer that produced an output of [sequence length \times 1] representing the estimated vGRF waveform. Training was done with the Adam optimiser (learning rate = 0.001, batch size = 32), using early stopping with a patience of 100 epochs and a maximum of 1000 epochs. Batches were shuffled each epoch. The loss function was mean squared error between predicted and target vGRF sequences.

To avoid overfitting, Leave-One-Subject-Out (LOSO) cross-validation was used. Additionally, an ensemble model was created [16]. The data of one subject was left out while the model was trained on the remaining ten subjects. Within this training set, seven random splits were performed, each time allocating seven subjects for training and three for validation.

After training, the left-out subject’s data was used to evaluate each model’s performance. To enhance robustness, model predictions were then averaged to obtain an ensemble model. A schematic overview can be found in Figure 3. This process was repeated, leaving every subject out once.

Indoor model performance was evaluated across all trials, including all speeds and cadences. Performance was assessed using the RMSE of the vGRF normalised to bodyweight [16, 24]. Pearson correlation coefficient and the active peak error were also calculated [16]. All evaluations were done during stance phase only.

2.6 Feature importance analysis

Feature importance analysis was performed using an ablation approach. In this method, predefined groups of features, either all signals from a single IMU or all spatio-temporal features, were systematically excluded from model training and resulting changes in performance metrics were evaluated. By comparing model results across all possible feature group combinations, the relative importance of each group was inferred. No distinctions were made between individual axes or signal types within a single IMU, since sensors equipped with IMUs typically record both angular velocity and acceleration across all axes. Statistical significance between models was assessed using paired t-tests.

2.7 Outdoor validation

To evaluate model performance in an outdoor setting, validation was conducted using pressure insoles, which estimate vGRF based on the applied pressure. Pressure insoles were first validated by comparing them to measured vGRF from the treadmill. The model used for outdoor validation included all features and had

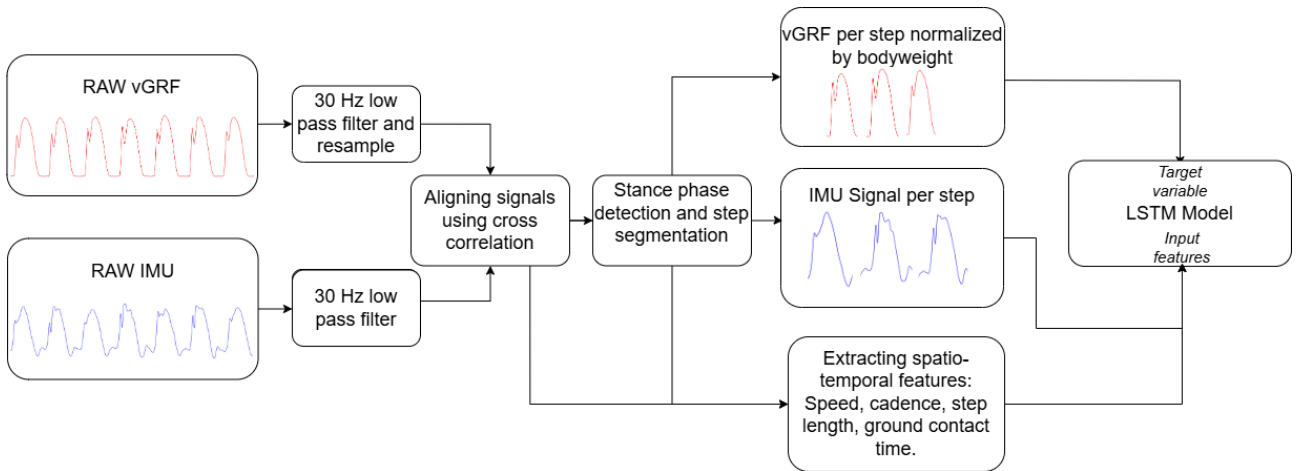


Figure 2: Overview of the pre-processing pipeline for the machine learning model. Raw vertical ground reaction force (vGRF) and inertial measurement unit (IMU) signals are first low-pass filtered at 30 Hz. Afterwards, vGRF data is resampled to match the IMU sampling frequency. The signals are temporally aligned using cross-correlation, after which stance phase detection is conducted and the data is segmented into different steps. vGRF is normalised by body weight. Spatio-temporal features are extracted and used as input features for the LSTM model. While the figure illustrates IMU acceleration signals only, angular velocity data (gyroscope) is also included in the model.

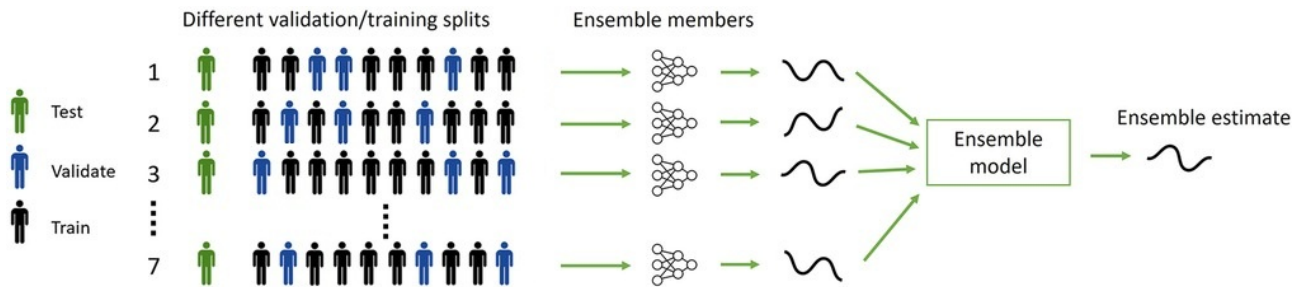


Figure 3: Ensemble model, seven distinct validation/training splits were used for training. The ensemble model was created by averaging over all ensemble members’ estimates. Figure adapted and modified with permission from Scheltinga et al. [16].

the same architecture and cross-validation procedure described in Section 2.5. The model was trained using vGRF measured by the treadmill as target variable and tested using vGRF estimated by the insoles. Performance was assessed using outdoor trials and indoor trials at preferred cadence, enabling direct comparison since outdoor trials were only performed at preferred cadence.

3 Results

Data from eleven subjects out of the twelve included are used for this study. One participant is excluded because of a forefoot strike pattern.

3.1 Part 1: Indoor model

Two measurement conditions from a single subject are excluded because one IMU disconnected, resulting in a total of 13,000 evaluated steps ($11 \cdot 12 \cdot 100 - 2 \cdot 100$). The results of the leave-one-subject-out (LOSO) cross-validation across all combinations of input features are presented in Table 1. The LSTM model using all features achieved an RMSE of 0.10 ± 0.025 BW.

No significant differences in RMSE, relative peak error or Pearson correlation coefficient were found between feature sets that included the sternum IMU ($p \geq 0.05$). In contrast, feature sets excluding the sternum IMU exhibit significantly higher RMSE, higher relative peak error and lower Pearson correlation coefficients ($p < 0.05$).

Among the feature sets that exclude the sternum IMU, using only the wrist IMU results in the lowest RMSE, although not significantly lower than using both the wrist IMU and spatio-temporal features ($p \geq 0.05$). However, using only spatio-temporal features leads to a significantly higher RMSE ($p < 0.05$).

Figure 4 shows model performance for participants with lowest and highest RMSE during the trial at preferred cadence at 12 km/h. While the ensemble model closely follows the actual vGRF for participant 1, it does not consistently capture the peaks accurately for participant 5. Similar inaccuracies in peak prediction were also observed in other participants with higher RMSE values.

The RMSE for each participant is shown in Figure 5. The RMSE of the ensemble model is compared to the mean and standard deviation of the seven individual ensemble members. The ensemble model consistently achieves a lower RMSE than the average of the individual ensemble members.

Table 2 reports the RMSE and relative RMSE (rRMSE, defined as RMSE normalised by the maximum vertical ground reaction force of each step) across running speeds for two feature sets: sternum and wrist only. For both sensor locations, a consistent trend is observed: RMSE and rRMSE increase with speed. Notably, the increase in RMSE across speeds is steeper than that of the rRMSE.

Table 1: Feature importance analysis results for vGRF estimation. All combinations of input features are shown along with their corresponding performance metrics: Root mean squared error (RMSE) expressed in body weight (BW) and Relative peak error expressed as a percentage. Values are reported as mean \pm standard deviation.

Feature Set	RMSE (BW)	Relative peak error	Pearson r
All Features	0.10 ± 0.025	$2.83 \pm 1.17\%$	0.99 ± 0.0042
Sternum IMU + Spatio-temporal Features	0.10 ± 0.026	$3.13 \pm 1.41\%$	0.99 ± 0.0040
Sternum IMU + Wrist IMU	0.10 ± 0.024	$2.85 \pm 1.18\%$	0.99 ± 0.0040
Sternum IMU Only	0.10 ± 0.026	$2.64 \pm 0.82\%$	0.99 ± 0.0046
Wrist IMU Only	0.17 ± 0.042	$6.41 \pm 2.66\%$	0.98 ± 0.0077
Wrist IMU + Spatio-temporal Features	0.18 ± 0.066	$7.57 \pm 4.64\%$	0.98 ± 0.011
Spatio-temporal Features Only	0.21 ± 0.045	$8.07 \pm 3.63\%$	0.97 ± 0.019

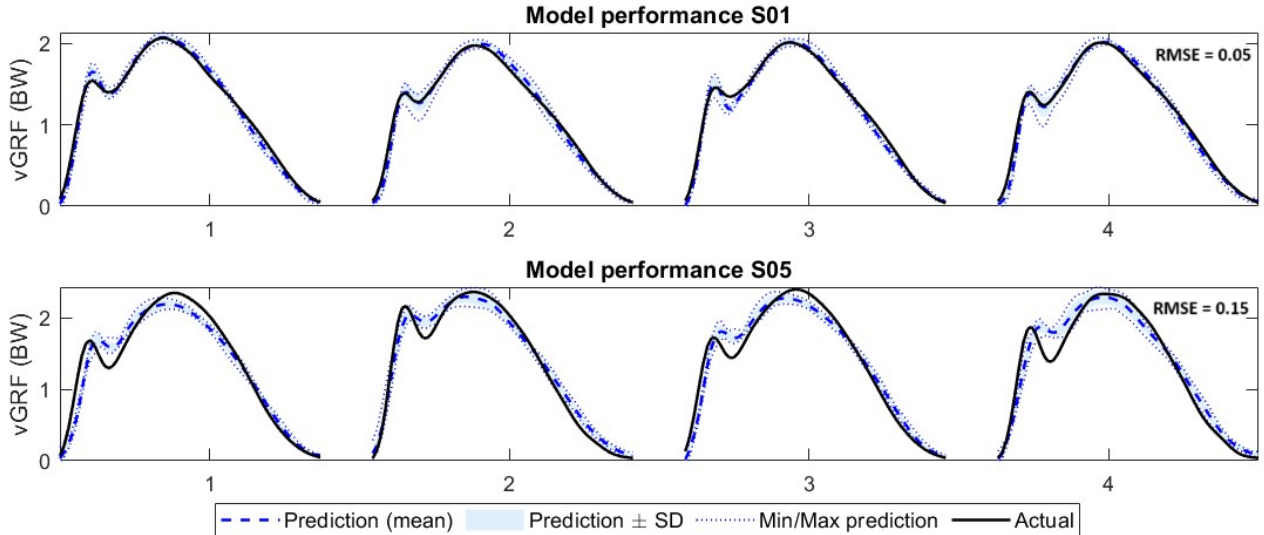


Figure 4: A typical example of model performance for the participants with the lowest (S01) and highest (S05) root mean squared error (RMSE) during the trial at 12 km/h with preferred cadence. The mean ensemble prediction, standard deviation (SD) and minimum and maximum predictions are shown alongside the measured vertical ground reaction force (vGRF).

Figure 6 shows the measured versus predicted peak vGRF for participant 7 across all speeds. Both measured and predicted peak vGRF values increase with speed. There is a strong linear relationship, with a Pearson correlation of 0.97. Most data points lie above the identity line ($y = x$), indicating that the model tends to overestimate the peak vGRF for this participant. Similar linear relationships are observed in other participants, although for some, the relationship between speed and peak vGRF is less consistent compared to this participant (See Appendix B for further details).

3.2 Part 2: Outdoor validation

One outdoor measurement condition from a single subject is excluded due to corrupted pressure insole data, resulting in a total of 4,300 evaluated outdoor steps ($11 \cdot 4 \cdot 100 - 100$). For the indoor data, pressure insole estimations are compared to vGRF measured by the instrumented treadmill, yielding an RMSE of 0.24 BW and a Pearson correlation coefficient of 0.95.

Outdoor validation is performed using the model with all features. Figure 7 presents the RMSE (BW), Pearson correlation coefficient and relative peak error

for indoor and outdoor trials, shown per participant. On average, both conditions yield an RMSE of 0.19 BW. Although variability in performance is observed between subjects, no significant differences are found at the group level between indoor and outdoor results ($p \geq 0.05$).

4 Discussion

An LSTM model, capable of estimating vertical ground reaction forces, using smartwatch-derived data in combination with IMU sensors on the wrist and sternum, was developed. Model validation based on RMSE, Pearson correlation coefficient and relative peak error showed improved performance compared to previous work. This study reported an RMSE of 0.10 ± 0.02 BW. Other studies reported RMSEs of 0.16 to 0.27 BW [16, 17, 24, 25]. A relative peak error (in %) of 2.64 ± 0.82 was found while Scheltinga et al. found a relative peak error of 3.61 ± 1.80 [16]. Unlike previous works, this model only uses sensors that runners will already carry with them, as IMUs are nowadays also integrated in smartwatches and heart rate straps. This makes the step to wider application smaller.

Table 2: RMSE (in BW) and relative RMSE (rRMSE) per speed for sternum and wrist sensors. Values are reported as mean \pm standard deviation.

Speed (km/h)	Sternum only		Wrist only	
	RMSE	rRMSE	RMSE	rRMSE
8	0.081 ± 0.022	0.034 ± 0.008	0.143 ± 0.051	0.059 ± 0.016
10	0.095 ± 0.026	0.038 ± 0.008	0.157 ± 0.051	0.064 ± 0.019
12	0.109 ± 0.030	0.043 ± 0.010	0.163 ± 0.041	0.064 ± 0.013
14	0.125 ± 0.038	0.048 ± 0.013	0.186 ± 0.060	0.072 ± 0.020

The LOSO validation revealed noticeable variation in RMSE between subjects, suggesting that generalisation across subjects can be improved. No clear explanation was found for why specific subjects had lower or higher RMSE. Including more subjects with diverse characteristics could further improve generalisation and thus potentially reduce variabilities in RMSE, although some outliers may always remain.

Feature importance analysis showed that the sternum IMU data provided the most valuable information for estimating vGRF. Adding data from additional features did not significantly improve model performance compared to using only the sternum IMU. However, for practical applications, relying solely on the wrist IMU and static features may be preferable, as not all runners wear a heart rate monitor belt, whereas smartwatches are commonly used. Using only wrist-derived features, the model achieved an RMSE of 0.17 ± 0.042 BW, which still compares well to previously reported values in the literature (0.16-0.27 BW). A trade-off must be made between model accuracy and user convenience when implementing this approach in real-world settings. Smartwatch-derived variables (speed, cadence, ground contact time and step length) did not significantly improve model performance when combined with IMU signals. Similarly, smartwatch-derived variables alone (i.e., without

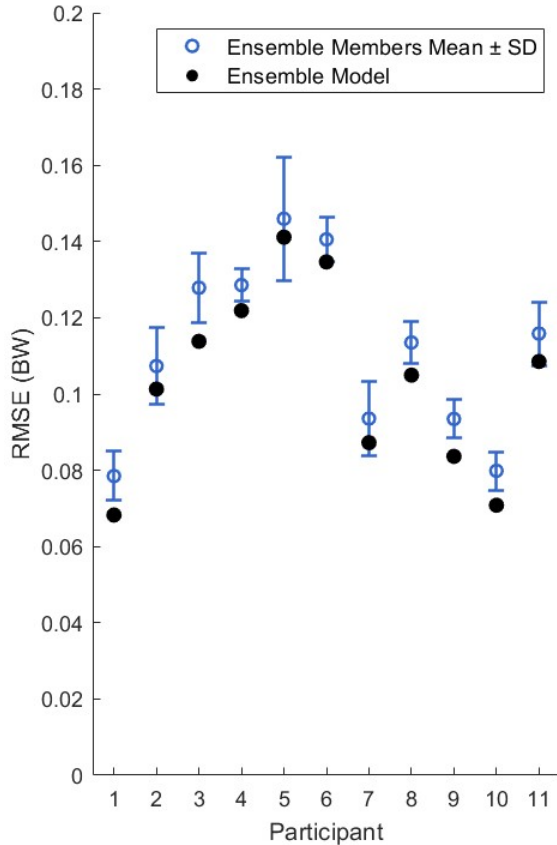


Figure 5: RMSE (BW) for all participants. In blue, the mean and standard deviation of the seven ensemble members are shown. In black, the eventual ensemble model is shown.

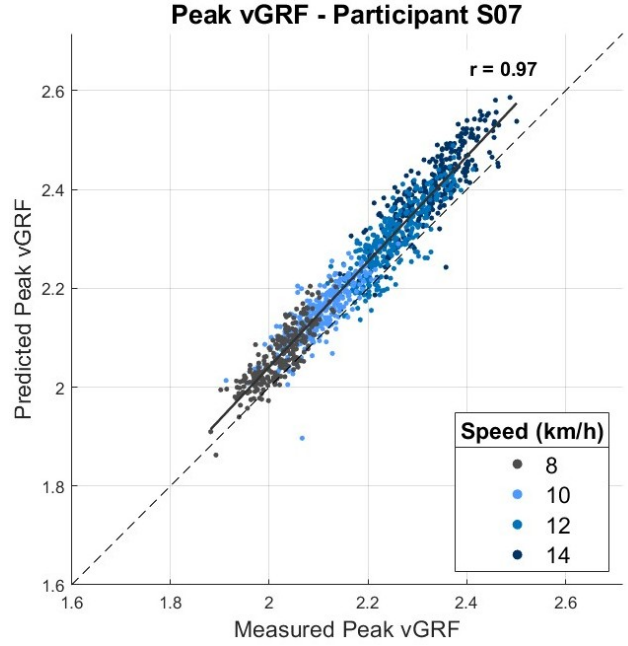


Figure 6: Measured versus predicted peak vGRF for participant 7. The dotted line shows the identity line and the solid line represents the best-fit linear regression of the measured versus predicted values. Point colour reflects running speed.

IMU input) were insufficient to achieve comparable accuracy, highlighting the importance of using other wearable sensors such as IMUs for reliable vGRF estimation.

Both RMSE and rRMSE increased with running speed. The rise in RMSE is expected, as vGRF waveforms generally have higher amplitudes at higher speeds [30]. Although rRMSE accounts for this and was expected to be similar across speeds, a slight increase was observed, suggesting that the error is not solely due to higher vGRF values. Other potentially contributing factors are shorter ground contact times at higher speeds, resulting in fewer data points per step and increased variability in running patterns due to greater effort at higher running speeds. However, the model remained sensitive to changes in signal amplitude, as increases in peak vGRF resulted in an increase in the predicted peak vGRF, indicating that the model is sensitive to speed changes.

Outdoor validation is essential to bridge the gap to real-world monitoring, as current research focuses mainly on indoor running [31] and there are substantial differences in kinetics and kinematics between treadmill and outdoor running [28, 29]. In this study, outdoor validation was performed using pressure insoles, which have been shown to correlate highly with measured GRF (Pearson coefficient = 0.95), but absolute agreement is limited, with an RMSE of 0.24 BW compared to treadmill measurements. No significant differences were found between indoor and outdoor test sets, suggesting consistent indoor and outdoor performance. However, due to the relatively high error of the pressure insoles, these results should

Ensemble Model Comparison: Indoor vs Outdoor

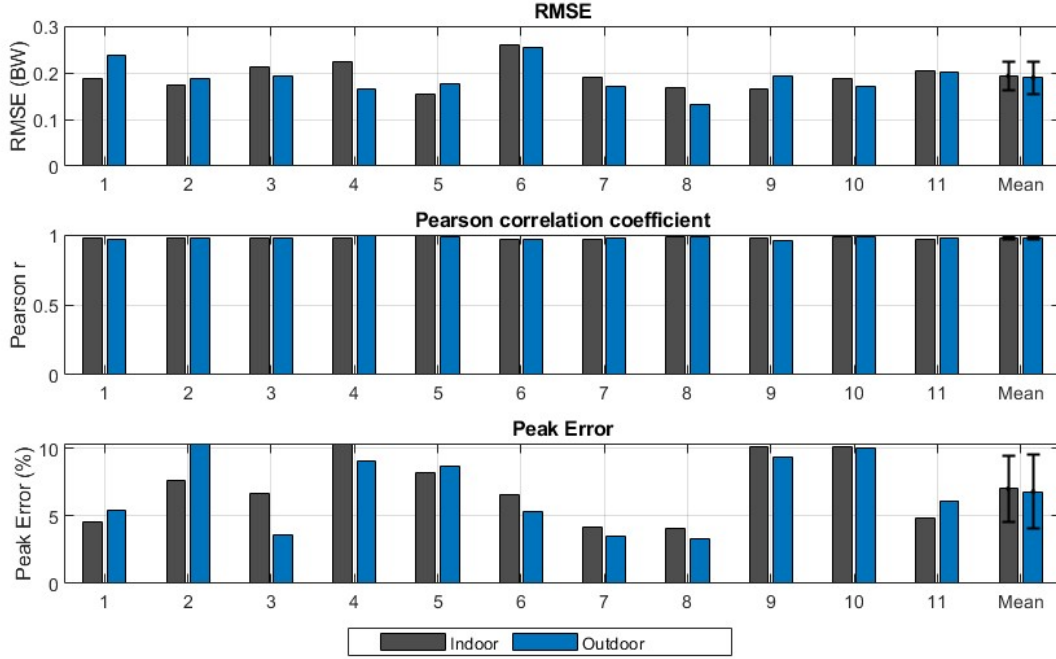


Figure 7: Comparison of ensemble model performance between indoor and outdoor trials. RMSE, Pearson correlation coefficient and relative peak error are presented for each participant individually. The final bar in each subplot represents the mean performance across all participants, with error bars indicating the standard deviation.

be interpreted with caution. Further validation remains essential, especially as the literature faces challenges in establishing a reliable outdoor ground truth. Additional challenges lie in the variability of outdoor running. Differences in surface type (e.g., tartan, concrete, sand or grass) can affect GRF characteristics and running kinematics [32], while courses with frequent corners or hilly terrain introduce additional variability in GRF waveforms [33, 34]. These factors must be considered to develop and validate a robust and generalizable model. This can be done by including multiple conditions during data collection, such that the model can also generalise to these environments. For instance, Alcantara et al. incorporated different running slopes for training their model [24].

Future work could focus on implementing this model within commercially available smartwatches and heart rate straps. This would enable longitudinal studies to investigate the relation between GRF and RRI or running performance. To date, only one study has considered this relation [10], but only peak vGRF has been considered. Estimating the full GRF waveform, as done within the present study, enables the extraction of additional biomechanical GRF features, such as vertical loading rate, vertical impulse or passive peak height, which could play a role in injury development and performance enhancement. As research has already shown that not only the peak vGRF, but also other vGRF metrics are associated with certain injuries [14, 35].

A key distinction between the IMUs used in this study and those in commercially available smartwatches lies in the sampling frequency. The IMUs used in this study are part of the Xsens MVN Link system (Movella, Enschede, The Netherlands), which consists of sensors wired to a body pack and measures at a sampling frequency of 240 Hz. In contrast, IMUs embedded in commercially available smartwatches and heart rate straps typically operate at lower sampling frequencies. Previous research by Haitjema et al. has shown that different sampling frequencies can significantly influence the outcome of acceleration data [36]. Additionally, sensor-to-segment calibration was done according to manufacturer’s instructions and algorithm. In real-world applications, smartwatches and heart rate straps will be placed inconsistently and will move during activities, which limits reliability of the data. In addition, variations in IMU orientation between devices introduce further challenges. Moreover, requiring a sensor-to-segment calibration before each run is impractical for recreational runners. There is a clear need to develop models that are robust to these conditions and trained on data reflecting such real-world scenarios.

This study only utilises a single bidirectional LSTM layer with a standard mean squared error loss function. However, the choice of loss function can significantly influence which parts of the vGRF curve the model prioritises during learning (i.e., passive or active peak). Alternative loss functions could shift the focus of the machine learning model on correctly estimating the timing and amplitude of the passive and active peaks,

as these have been shown to be common sources of error. Moreover, custom loss functions could be designed to prioritise other biomechanically relevant features of the vGRF curve, such as impulse or loading rate, depending on the specific application.

Ensemble models performed better than most individual ensemble members, consistent with previous findings [16, 37]. Ensemble learning also improved Pearson correlation coefficient, although the relative peak error was not always better compared to individual ensemble members. Since each ensemble member requires training a separate model, computational cost increases linearly with ensemble size. Therefore, ensemble learning is recommended for studies where low prediction error is essential and resources allow.

This study focused on predicting vGRF in runners with a rearfoot strike. This is because most people have a rearfoot strike pattern in long-distance running. Hasegawa et al. reported that 74.9% of the participants in an elite-level half marathon were rearfoot strikers and Larson et al. even reported 88.9% of the participants in a (half) marathon were rearfoot strikers [12, 13]. By also focusing on midfoot and forefoot strikers, the generalisability of the models would improve. Alcantara et al. used footstrike pattern as input to the LSTM model, which improved model performance [24].

5 Conclusion

The study demonstrated that vGRF can be estimated using IMUs on the wrist and sternum combined with smartwatch-derived variables, using an LSTM network. An RMSE of 0.10 BW was found using the sternum IMU and 0.17 BW with only the wrist IMU. These results show promising potential in the direction of longitudinally estimating vGRF in an outdoor setting without requiring additional sensors.

References

- [1] R. Van Gent, D. Siem, M. van Middelkoop, A. Van Os, S. Bierma-Zeinstra, and B. Koes, "Incidence and determinants of lower extremity running injuries in long distance runners: a systematic review," *British journal of Sports Medicine*, vol. 41, no. 8, pp. 469–480, 2007.
- [2] S. Videbæk, A. M. Bueno, R. O. Nielsen, and S. Rasmussen, "Incidence of running-related injuries per 1000 h of running in different types of runners: a systematic review and meta-analysis," *Sports Medicine*, vol. 45, pp. 1017–1026, 2015.
- [3] A. H. Gruber, "The "impacts cause injury" hypothesis: Running in circles or making new strides?" *Journal of Biomechanics*, vol. 156, p. 111694, 2023.
- [4] M. A. Zandbergen, X. J. Ter Wengel, R. P. van Middelaar, J. H. Buurke, P. H. Veltink, and J. Reenalda, "Peak tibial acceleration should not be used as indicator of tibial bone loading during running," *Sports Biomechanics*, pp. 1–18, 2023.
- [5] E. S. Matijevich, L. M. Branscombe, L. R. Scott, and K. E. Zelik, "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*, vol. 14, no. 1, p. e0210000, 2019.
- [6] S. Willwacher, M. Kurz, J. Robbin, M. Thelen, J. Hamill, L. Kelly, and P. Mai, "Running-related biomechanical risk factors for overuse injuries in distance runners: a systematic review considering injury specificity and the potentials for future research," *Sports Medicine*, vol. 52, no. 8, pp. 1863–1877, 2022.
- [7] L. Ceyskens, R. Vanelderden, C. Barton, P. Malliaras, and B. Dingenen, "Biomechanical risk factors associated with running-related injuries: a systematic review," *Sports Medicine*, vol. 49, pp. 1095–1115, 2019.
- [8] M. Bertelsen, A. Hulme, J. Petersen, R. K. Brund, H. Sørensen, C. Finch, E. T. Parner, and R. Nielsen, "A framework for the etiology of running-related injuries," *Scandinavian journal of medicine & science in sports*, vol. 27, no. 11, pp. 1170–1180, 2017.
- [9] W. B. Edwards, "Modeling overuse injuries in sport as a mechanical fatigue phenomenon," *Exercise and Sport Sciences Reviews*, vol. 46, no. 4, pp. 224–231, 2018.
- [10] D. Kiernan, D. A. Hawkins, M. A. Manoukian, M. McKallip, L. Oelsner, C. F. Caskey, and C. L. Coolbaugh, "Accelerometer-based prediction of running injury in national collegiate athletic association track athletes," *Journal of Biomechanics*, vol. 73, pp. 201–209, 2018.
- [11] D. Janoušek and P. Stejskal, "Forefoot strike, rear foot strike or running shoes. does it matter?" *Studia sportiva*, vol. 13, no. 1, pp. 49–54, 2019.
- [12] H. Hasegawa, T. Yamauchi, and W. J. Kraemer, "Foot strike patterns of runners at the 15-km point during an elite-level half marathon," *The Journal of Strength & Conditioning Research*, vol. 21, no. 3, pp. 888–893, 2007.
- [13] P. Larson, E. Higgins, J. Kaminski, T. Decker, J. Preble, D. Lyons, K. McIntyre, and A. Normile, "Foot strike patterns of recreational and sub-elite runners in a long-distance road race," *Journal of Sports Sciences*, vol. 29, no. 15, pp. 1665–1673, 2011.

- [14] H. Van der Worp, J. W. Vrieling, and S. W. Bredeweg, "Do runners who suffer injuries have higher vertical ground reaction forces than those who remain injury-free? a systematic review and meta-analysis," *British journal of Sports Medicine*, vol. 50, no. 8, pp. 450–457, 2016.
- [15] A. Ancillao, S. Tedesco, J. Barton, and B. O’Flynn, "Indirect measurement of ground reaction forces and moments by means of wearable inertial sensors: A systematic review," *Sensors*, vol. 18, no. 8, p. 2564, 2018.
- [16] B. L. Scheltinga, J. N. Kok, J. H. Buurke, and J. Reenalda, "Estimating 3d ground reaction forces in running using three inertial measurement units," *Frontiers in Sports and Active Living*, vol. 5, p. 1176466, 2023.
- [17] S. R. Donahue and M. E. Hahn, "Estimation of ground reaction force waveforms during fixed pace running outside the laboratory," *Frontiers in Sports and Active Living*, vol. 5, p. 974186, 2023.
- [18] B. L. Scheltinga, J. H. Buurke, J. N. Kok, and J. Reenalda, "Repeatability of vertical ground reaction force estimation during running on the athletics track on 3 different days," *Journal of Applied Biomechanics*, vol. 41, no. 2, pp. 167–178, 2025.
- [19] W. Johnston and B. Heiderscheit, "Mobile technology in running science and medicine: are we ready?" *Journal of Orthopaedic & Sports Physical Therapy*, vol. 49, no. 3, pp. 122–125, 2019.
- [20] A. Lacey, E. Whyte, S. O’Keeffe, S. O’Connor, and K. Moran, "A qualitative examination of the factors affecting the adoption of injury focused wearable technologies in recreational runners," *PloS one*, vol. 17, no. 7, p. e0265475, 2022.
- [21] K. Helsen, M. Janssen, S. Vos, and J. Scheerder, "Two of a kind? similarities and differences between runners and walkers in sociodemographic characteristics, sports related characteristics and wearable usage," *International Journal of Environmental Research and Public Health*, vol. 19, no. 15, p. 9284, 2022.
- [22] A. Backes, S. D. Skej , P. Gette, R.  . Nielsen, H. S rensen, C. Morio, and L. Malisoux, "Predicting cumulative load during running using field-based measures," *Scandinavian Journal of Medicine & Science in Sports*, vol. 30, no. 12, pp. 2399–2407, 2020.
- [23] B. L. Scheltinga, H. Usta, J. Reenalda, and J. H. Buurke, "Estimating vertical ground reaction force during running with 3 inertial measurement units," *Journal of Biomedical Engineering and Biosciences*, vol. 9, pp. 31–38, 2022.
- [24] R. S. Alcantara, W. B. Edwards, G. Y. Millet, and A. M. Grabowski, "Predicting continuous ground reaction forces from accelerometers during uphill and downhill running: a recurrent neural network solution," *PeerJ*, vol. 10, p. e12752, 2022.
- [25] F. J. Wouda, M. Giuberti, G. Bellusci, E. Maartens, J. Reenalda, B.-J. F. Van Beijnum, and P. H. Veltink, "Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors," *Frontiers in physiology*, vol. 9, p. 218, 2018.
- [26] R. D. Gurchiek, N. Cheney, and R. S. McGinnis, "Estimating biomechanical time-series with wearable sensors: A systematic review of machine learning techniques," *Sensors*, vol. 19, no. 23, p. 5227, 2019.
- [27] A. Mohammed and R. Kora, "A comprehensive review on ensemble deep learning: Opportunities and challenges," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, 2023.
- [28] B. M. Nigg, R. W. De Boer, and V. Fisher, "A kinematic comparison of overground and treadmill running," *Medicine and science in sports and exercise*, vol. 27, pp. 98–98, 1995.
- [29] A. F. D. Lempke, A. P. Audet, M. G. Wasserman, A. C. Melvin, K. Soldes, E. Heithoff, S. Shah, K. M. Kozloff, and A. S. Lepley, "Biomechanical differences and variability during sustained motorized treadmill running versus outdoor overground running using wearable sensors," *Journal of Biomechanics*, vol. 178, p. 112443, 2025.
- [30] J. Hamill, B. Bates, K. Knutzen, and J. Sawhill, "Variations in ground reaction force parameters at different running speeds," *Human Movement Science*, vol. 2, no. 1-2, pp. 47–56, 1983.
- [31] L. C. Benson, A. M. R ais nen, C. A. Clermont, and R. Ferber, "Is this the real life, or is this just laboratory? a scoping review of imu-based running gait analysis," *Sensors*, vol. 22, no. 5, p. 1722, 2022.
- [32] S. J. Dixon, A. C. Collop, and M. E. Batt, "Surface effects on ground reaction forces and lower extremity kinematics in running," *Medicine & Science in Sports & Exercise*, vol. 32, no. 11, pp. 1919–1926, 2000.
- [33] J. S. Gottschall and R. Kram, "Ground reaction forces during downhill and uphill running," *Journal of Biomechanics*, vol. 38, no. 3, pp. 445–452, 2005.
- [34] L. R. Williams, T. W. Standifird, A. Creer, H. B. Fong, and D. W. Powell, "Ground reaction force profiles during inclined running at iso-efficiency speeds," *Journal of Biomechanics*, vol. 113, p. 110107, 2020.

- [35] C. D. Johnson, A. S. Tenforde, J. Outerleys, J. Reilly, and I. S. Davis, “Impact-related ground reaction forces are more strongly associated with some running injuries than others,” *The American journal of sports medicine*, vol. 48, no. 12, pp. 3072–3080, 2020.
- [36] A. Haitjema, F. J. Wouda, P. H. Veltink, G. Miller, and J. Reenalda, “Ambulatory monitoring of injury risk in runners: The effect of sampling frequency on peak tibial acceleration and impulse,” in *2024 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2024, pp. 1–6.
- [37] E. Grzesiak, J. Sloboda, and H. C. Siu, “Predicting ankle moment trajectory with adaptive weighted ensemble of lstm networks,” in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE, 2022, pp. 1–7.

Appendix

A AI statement

During the preparation of this work, I used ChatGPT to improve sentence flow and assist with text editing. After using this tool/service, I thoroughly reviewed and revised the content as needed, and take full responsibility for the final version of the manuscript.

B Extra analysis speed sensitivity

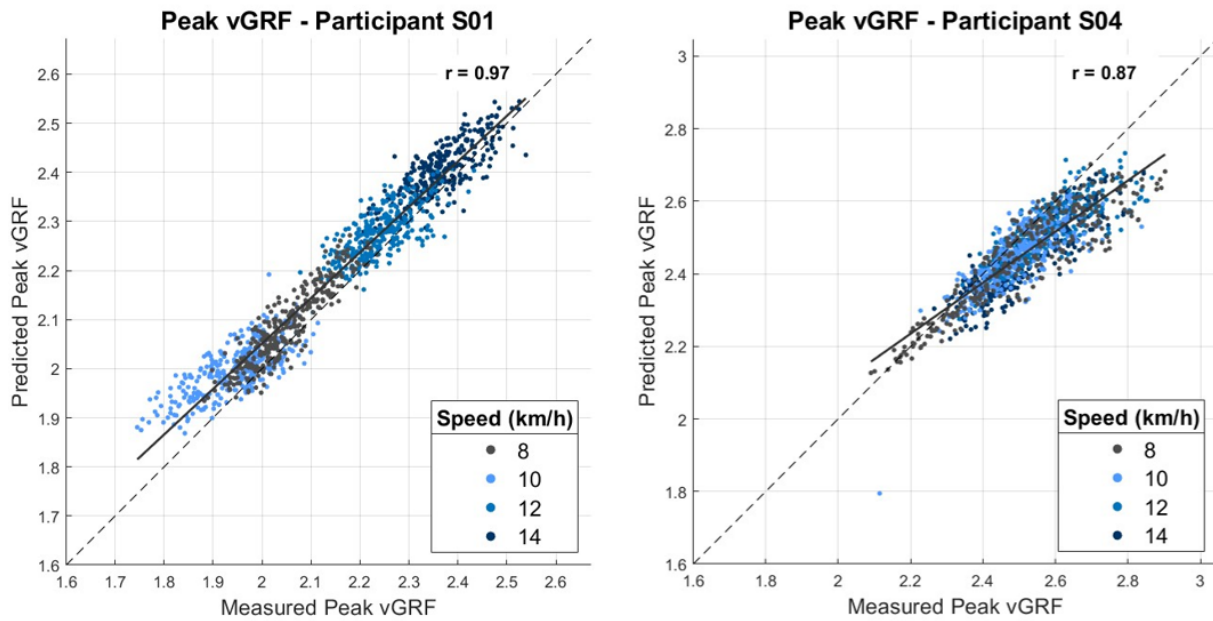


Figure 8

Figure 8 displays the measured versus predicted peak vGRF for participants 1 and 4. They show a similar linear increase compared to participant 7 in section 3.1, although participant 4 does have a lower Pearson correlation coefficient.

Participant 1 shows distinct clustering by speed. Interestingly, the predicted and measured peak vGRF values for 8 km/h are higher than those for 10 km/h, which is counterintuitive. A possible explanation is that this participant adopts a running style with increased vertical displacement or greater vertical oscillation at lower speeds, resulting in higher vertical ground reaction forces during push-off.

In contrast, participant 4 shows no clear differences between running speeds, either in the predicted or measured peak vGRF values. This suggests that the model is still capable of distinguishing between high and low peak vGRF values even in the absence of clear speed-dependent variation.