

Subject-Independent Estimation of Vertical Ground Reaction Force from a Wearable Accelerometer: A Reproducible Benchmark of Accuracy, Robustness, and Inference Latency

Sergio de N. R. L. Jr, Ronaldo de F. Zampolo, Antonio Pereira Jr

Abstract—Estimating vertical ground reaction force (vGRF) from wearable inertial sensors may enable gait assessment outside laboratory environments; however, subject-independent generalization, robustness to sensing disturbances, and computational cost remain important challenges. This study presents a reproducible benchmark for continuous vGRF estimation from a single triaxial accelerometer positioned on the right tibia. Nine predictive approaches, including classical, ensemble-based, and temporal neural network models, were evaluated under a common preprocessing and causal windowing pipeline using leave-one-subject-out (LOSO) validation across nine participants. Robustness was assessed by applying controlled perturbations to the test signals, including additive noise, sensor rotation, axis ablation, and temporal occlusion, while computational efficiency was quantified through inference latency per window. Random Forest achieved the lowest mean RMSE (130.93 ± 44.80 N), followed by GRU (138.93 ± 42.05 N). TCN and CNN1D provided lower inference latency with intermediate predictive accuracy, highlighting a favorable trade-off for computationally constrained applications. Robustness analyses revealed model-specific sensitivity, particularly to additive noise, final-window occlusion, and z -axis ablation. The Friedman test indicated significant performance differences among models ($p = 1.38 \times 10^{-10}$), with strong agreement in model ranking across participants ($W = 0.8708$). These findings demonstrate that model selection for wearable vGRF estimation should jointly consider predictive accuracy, robustness, and computational efficiency, providing a methodological basis for future validation in independent cohorts and embedded sensing platforms.

Index Terms—vertical ground reaction force, wearable accelerometers, inertial sensors, subject-independent validation, machine learning, temporal neural networks, robustness analysis, gait analysis.

I. INTRODUCTION

Ground reaction force (GRF) is widely used to characterize locomotion, detect gait events, assess asymmetries, and monitor neuromuscular function [1], [2]. Its vertical component (vGRF) is typically measured using force platforms, which are considered the biomechanical gold standard [1]. Instrumented insoles provide a portable alternative; however, cost, calibration requirements, and operational constraints limit their prolonged use outside laboratory environments [3], [4].

Inertial measurement units (IMUs) offer a low-cost and highly portable alternative. However, estimating vGRF from

segmental accelerations constitutes a nonlinear and temporally dependent inverse problem [5], [6].

Different machine learning models have been applied to GRF estimation from wearable sensors [5]–[11]. However, heterogeneity in sensor configurations, activities, target signals, and evaluation protocols makes direct comparison of results difficult. Furthermore, few studies jointly evaluate cross-subject generalization, robustness, and inference cost under a common experimental protocol.

A. Positioning Relative to Previous Work

Table I summarizes representative studies on GRF estimation using wearable sensors.

Although LOSO validation has already been adopted in previous studies, the works considered here do not jointly evaluate accuracy, robustness, and computational cost under identical experimental conditions. The present study introduces a reproducible benchmark of nine models using a single triaxial tibial accelerometer, LOSO partitioning, a common preprocessing pipeline, controlled perturbations, per-window inference latency, and statistical comparison. Therefore, the main contribution lies in the integrated assessment of the trade-offs among predictive performance, robustness, and deployment feasibility, rather than in the proposal of a new model architecture.

B. Main Contributions

The main contributions of this study are:

- a reproducible benchmark of nine models for continuous vGRF estimation from a single triaxial tibial accelerometer;
- a LOSO validation protocol with target normalization restricted to the training data and standardized robustness tests;
- a joint analysis of predictive accuracy, robustness, inference latency, and statistical differences among the evaluated models.

II. MATERIALS AND METHODS

A. Dataset Description

The publicly available multimodal dataset introduced by Grouvel et al. [12] was used in this study. The dataset comprises synchronized recordings from optoelectronic motion

All authors are with the Signal Processing Laboratory, Federal University of Pará (UFPA), 66075-110 Belém, Brazil.

Corresponding author: Antonio Pereira Jr (e-mail: apereira@ufpa.br).

TABLE I
COMPARISON WITH STUDIES ON WEARABLE SENSOR-BASED GRF ESTIMATION. NR: NOT REPORTED; GRF/M: GROUND REACTION FORCES AND MOMENTS.

Study	Model and target	Sensors and condition	Validation	Robustness	Latency/cost
Johnson et al. [5]	CNN; 3D GRF/M	Five accelerometers; running and change-of-direction tasks	External cross-laboratory testing	No	Near real-time; latency NR
Alcantara et al. [6]	RNN; continuous GRF	Sacral and shoe-mounted accelerometers; running at different speeds and inclines	LOSO	No	NR
Hossain et al. [7]	Deep <i>ensemble</i> ; 3D GRF and joint moments	Three IMUs; multiple walking conditions	LOSO	No	Complexity discussed; latency NR
Scheltinga et al. [8]	ANN <i>ensemble</i> ; 3D GRF	Three IMUs; running	LOSO	No	NR
Bogaert et al. [9]	Lasso; vGRF characteristics	One sacral IMU; running at multiple speeds	Subject-wise testing and internal LOSO	No	NR
Song et al. [10]	SER, KNN, and LSTM; continuous GRF	IMUs on the sacrum and tibias; running	Subject-dependent subject-independent scenarios	and No	Comparison on an edge device
Yilmazgün et al. [11]	CNN; 3D GRF	One to seven IMUs; six locomotor tasks	LOSO	No	NR
Present study	Nine models; continuous vGRF	One tibial accelerometer; over-ground walking	LOSO	Noise, rotation, ablation, and occlusion	Latency per window

capture systems, inertial sensors, instrumented insoles, and force platforms.

Participants: The original dataset includes ten asymptomatic adults. One participant was reserved for future external evaluation and excluded from all stages of the present study, resulting in nine participants and nine LOSO *folds*.

The participants had no neurological or musculoskeletal conditions affecting gait and had undergone no lower-limb or spinal surgery within the previous two years. The original study was approved under protocol CCER-2020-00358 and conducted in accordance with the Declaration of Helsinki, with informed consent obtained from all participants [12].

Experimental Protocol: Each participant completed a single laboratory session. Only overground walking trials performed at a self-selected comfortable speed along a 10 m walkway were considered.

The right-limb vGRF was used as the target variable, while the input signals consisted of triaxial accelerations from the IMU positioned on the right tibia, expressed in the sensor's local coordinate system.

Synchronization and Data Format: The dataset provides synchronized CSV files resampled to 100 Hz, with temporal alignment performed *offline* using an impact-based procedure. These files were used directly, without any additional synchronization.

B. Preprocessing

Each accelerometer axis was filtered using a causal fourth-order Butterworth low-pass filter with a cutoff frequency of 10 Hz, a configuration consistent with the spectral range of human gait and with real-time applications [1], [13], [14].

Subsequently, *z-score* normalization was applied separately to each axis and file.

C. Temporal Windowing and Dataset Construction

The signals were segmented into causal windows of $L = 30$ samples, with $STEP \in \{1, 3, 5\}$. Each window was represented as $\mathbf{X}_t \in \mathbb{R}^{3 \times L}$. The neural networks received this

sequential representation, whereas Ridge, Random Forest, and Boosting used its vectorized form, $\mathbf{x}_t = \text{vec}(\mathbf{X}_t) \in \mathbb{R}^{90}$.

The target was defined as the vGRF value at the final time instant of the window:

$$y_t = \text{vGRF}[t + L - 1]. \quad (1)$$

Metadata corresponding to participant, file, temporal indices, sampling frequency, and *STEP* were preserved.

D. Data Partitioning and Validation Protocol

A *leave-one-subject-out* (LOSO) validation protocol was adopted, comprising $N = 9$ *folds*. In each *fold*, one participant was held out exclusively for testing, while the remaining eight participants were used for training. Subject-level partitioning prevented correlated windows from being shared between the training and test sets and reduced the risk of information leakage [15].

The vGRF target was normalized using statistics computed exclusively from the training set of each *fold*. Predictions were transformed back to the original scale before the performance metrics were computed.

E. Evaluation and Reproducibility

All models were evaluated using the same preprocessing, windowing, and LOSO partitioning procedures. The following metrics were computed: RMSE, MAE, $p95(|e|)$, R^2 , MAPE, peak error, and mean inference latency per window.

Robustness was evaluated exclusively on the test set by applying additive noise (SNR of 20 and 10 dB), a 15° rotation, *z*-axis ablation, and final temporal occlusion. The impact of each perturbation was quantified as

$$RI = \frac{\text{RMSE}_{\text{stress}} - \text{RMSE}_{\text{base}}}{\text{RMSE}_{\text{base}}}. \quad (2)$$

F. Computational Cost

Computational cost was assessed using the mean inference latency per window, measured with a batch size of one.

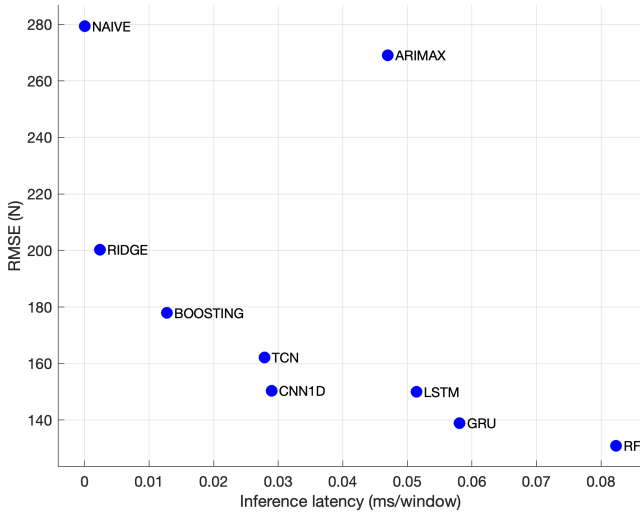


Fig. 1. Mean RMSE as a function of mean inference latency per window across the nine LOSO folds. Lower values on both axes indicate a more favorable trade-off.

G. Statistical Analysis

Differences in RMSE among the models were evaluated using the Friedman test, with the LOSO folds considered as blocks. The consistency of the model ranking was quantified using Kendall's coefficient of concordance (W). The p -value and W statistic were reported [16].

III. RESULTS

A. Predictive Performance under LOSO Validation

Table II presents the results across the nine LOSO folds. RF achieved the lowest mean RMSE, followed by GRU. RIDGE, ARIMAX, and NAIVE exhibited the largest errors, with NAIVE yielding a negative mean R^2 .

B. RMSE–Latency Trade-off

Figure 1 illustrates the trade-off between predictive performance and inference latency. RF achieved the lowest RMSE but exhibited the highest latency, whereas RIDGE was the fastest model but provided lower predictive accuracy. TCN and CNN1D combined low inference latency with intermediate predictive performance.

C. Robustness under Signal Perturbations

Table III presents, for each model, the perturbation that produced the largest mean increase in RMSE. RIDGE was most sensitive to 10 dB additive noise. For RF, GRU, and LSTM, the largest degradation occurred under final temporal occlusion, whereas for CNN1D and TCN it occurred under z -axis ablation.

D. Statistical Results

Figure 2 presents the distribution of RMSE across the nine LOSO folds. RF achieved the lowest median, followed by GRU.

The Friedman test indicated significant differences among the models ($p = 1.38 \times 10^{-10}$), with high concordance in the ranking across participants ($W = 0.8708$).

E. Representative Inference in an Unseen Subject

Figure 3 illustrates, for $\text{STEP} = 3$, the reconstruction of vGRF in a participant not used during training.

In this example, RF and LSTM showed closer agreement with the reference signal, followed by GRU, whereas RIDGE exhibited larger amplitude and phase deviations. This behavior is consistent with the global results presented above.

IV. DISCUSSION

A. Comparative Performance and Modeling Implications

Under identical experimental conditions, RF achieved the lowest mean RMSE, followed by GRU. The Friedman test indicated global differences among the models, while the high Kendall coefficient showed consistency in their ranking across participants. However, in the absence of *post hoc* tests, statistical superiority between specific pairs of models cannot be established.

The performance of RF, GRU, and LSTM suggests that nonlinear relationships and short-term temporal dependencies are relevant for vGRF estimation. In contrast, the higher errors obtained with RIDGE and ARIMAX indicate limitations of linear mappings and low-order autoregressive structures. TCN and CNN1D exhibited intermediate predictive performance with lower inference latency, making them potential alternatives for computationally constrained applications.

The errors obtained are of the same order of magnitude as those reported in previous studies [5], [6]. However, direct comparisons should account for differences in activity, sensor configuration, validation protocol, target variable, and metric units. Therefore, the main contribution of this study lies in the integrated comparison of accuracy, robustness, and inference latency, rather than in the proposal of a new model architecture.

B. Biomechanical Interpretation

The performance degradation under final temporal occlusion indicates that samples close to the target at $t + L - 1$ contain relevant predictive information. Sensitivity to the ablation of a_z suggests a dependence on this component of tibial acceleration.

RIDGE exhibited high sensitivity to 10 dB additive noise, whereas the nonlinear models experienced smaller performance degradation. These findings indicate greater relative tolerance of *ensemble*-based and neural architectures to the evaluated perturbations.

C. Accuracy–Latency Trade-off

RF achieved the lowest RMSE but exhibited the highest inference latency, whereas TCN and CNN1D combined lower latency with intermediate predictive performance. Measurements were performed on a per-window basis using a batch size of one, thereby approximating continuous streaming operation.

These results enable a relative comparison among the models in the evaluated computational environment but do not demonstrate their feasibility on microcontrollers or edge devices, which requires validation on the target hardware.

TABLE II
MEAN PERFORMANCE $\pm$ STANDARD DEVIATION ACROSS THE NINE LOSO *folds*.

Model	RMSE (N)	MAE (N)	$p95(e)$ (N)	R^2	Latency (ms/window)
RF	130.93 $\pm$ 44.80	93.82 $\pm$ 35.67	269.04 $\pm$ 79.03	0.757 $\pm$ 0.108	0.082 $\pm$ 0.015
GRU	138.93 $\pm$ 42.05	99.30 $\pm$ 35.04	288.75 $\pm$ 70.58	0.720 $\pm$ 0.125	0.058 $\pm$ 0.016
LSTM	150.00 $\pm$ 49.35	107.29 $\pm$ 40.18	312.88 $\pm$ 78.44	0.672 $\pm$ 0.165	0.051 $\pm$ 0.015
CNN1D	150.38 $\pm$ 38.59	116.41 $\pm$ 32.42	298.68 $\pm$ 67.57	0.680 $\pm$ 0.113	0.029 $\pm$ 0.006
TCN	162.20 $\pm$ 34.48	124.62 $\pm$ 27.49	317.99 $\pm$ 72.29	0.633 $\pm$ 0.074	0.028 $\pm$ 0.006
BOOSTING	177.94 $\pm$ 36.81	131.35 $\pm$ 29.66	367.99 $\pm$ 66.70	0.530 $\pm$ 0.222	0.013 $\pm$ 0.003
RIDGE	200.33 $\pm$ 44.23	155.63 $\pm$ 33.59	401.59 $\pm$ 90.90	0.424 $\pm$ 0.228	0.002 $\pm$ 0.001
ARIMAX	269.10 $\pm$ 61.94	230.95 $\pm$ 58.81	466.88 $\pm$ 102.63	0.018 $\pm$ 0.078	0.047 $\pm$ 0.016
NAIVE	279.34 $\pm$ 64.42	249.39 $\pm$ 57.54	439.45 $\pm$ 150.51	-0.061 $\pm$ 0.064	0.000 $\pm$ 0.000

TABLE III
WORST-CASE PERTURBATION FOR EACH MODEL. RMSE AND Δ RMSE VALUES ARE REPORTED IN N.

Model	Perturbation	RMSE _{base}	RMSE _{pert}	Δ RMSE	RI
RIDGE	SNR10	200.33	728.87	528.54	2.638
RF	OCC_FINAL	130.93	238.12	107.19	0.819
GRU	OCC_FINAL	138.93	253.38	114.45	0.824
CNN1D	ABL_Z	150.38	272.91	122.53	0.815
TCN	ABL_Z	162.20	277.58	115.38	0.711
LSTM	OCC_FINAL	150.00	249.48	99.48	0.663
BOOSTING	OCC_FINAL	177.94	271.44	93.50	0.525
ARIMAX	OCC_FINAL	269.10	272.30	3.20	0.012
NAIVE	ABL_Z	279.34	279.34	0.00	0.000

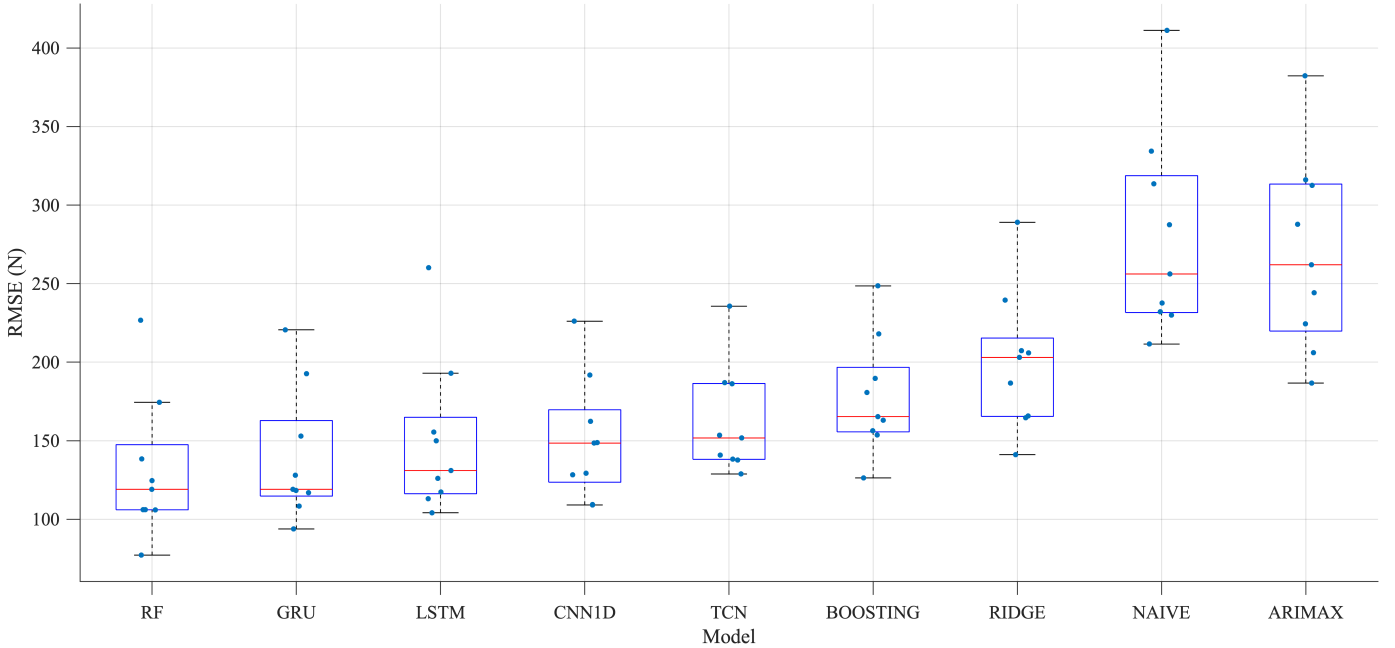


Fig. 2. Distribution of RMSE across the nine LOSO *folds*. The boxes represent the interquartile range, the central line indicates the median, the *whiskers* correspond to $1.5 \times \text{IQR}$, and the points represent the individual *folds*.

D. Cross-Subject Generalization

The LOSO protocol ensured that all windows from each participant were assigned exclusively to either the training or test set, thereby reducing the risk of information leakage. The high Kendall concordance indicates consistency in model ranking, while the variability in RMSE across the *folds* highlights inter-participant differences.

These results provide an estimate of generalization to unseen individuals drawn from the same population and acqui-

sition protocol, but they do not replace external validation in an independent cohort.

E. Limitations

The small number of participants limits the precision of the estimates and the external validity of the findings.

In addition, only asymptomatic adults, self-selected walking speed, a single sensor location, and synthetic perturbations were evaluated. Therefore, the results should be interpreted

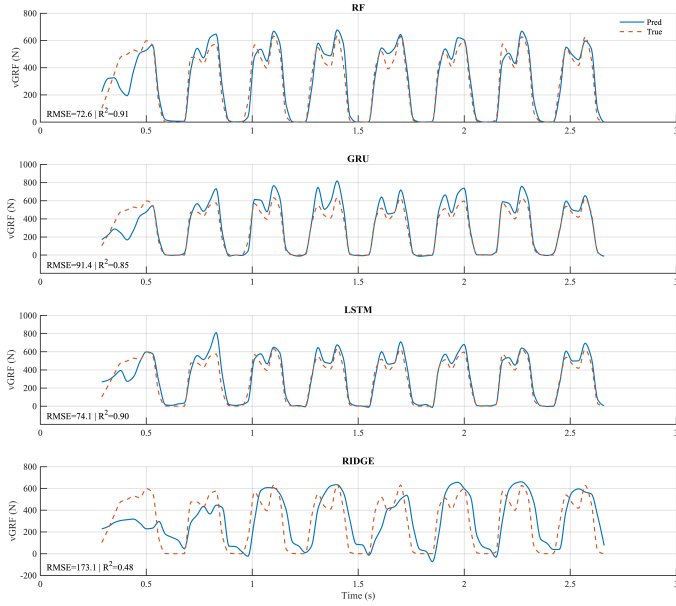


Fig. 3. Example of vGRF prediction for a participant not seen during training, with $STEP = 3$. Dashed lines indicate the reference signal, whereas solid lines represent the predictions of RF, GRU, LSTM, and RIDGE. RMSE and R^2 were computed at the final points of the windows; interpolation was used for visualization purposes only.

as a methodological benchmark restricted to the investigated cohort and experimental protocol.

Future work should include larger and independent cohorts, clinical populations, different walking speeds and sensor placements, as well as evaluations under real-world acquisition conditions and on the target hardware.

F. Clinical and Rehabilitation Relevance

The LOSO validation protocol and robustness tests approximate requirements of wearable systems intended for deployment in previously unseen users. However, the clinical relevance remains limited because clinical populations, external cohorts, and embedded platforms were not evaluated.

RF achieved the lowest RMSE, whereas TCN and CNN1D provided lower inference latency. Model selection should therefore consider this trade-off in relation to the target hardware.

V. CONCLUSION

This study presented a reproducible benchmark for estimating vGRF from a single triaxial tibial accelerometer, integrating LOSO validation, robustness testing, statistical analysis, and inference latency assessment.

RF achieved the lowest mean RMSE, followed by GRU, whereas TCN and CNN1D provided lower inference latency with intermediate predictive performance. The perturbation analysis revealed distinct model sensitivities, particularly to additive noise, temporal occlusion, and z -axis ablation. The Friedman test indicated global performance differences among the models, with high concordance in their ranking.

The proposed framework enables comparison of the trade-offs among predictive performance, robustness, and inference

latency, providing a methodological basis for future studies involving external cohorts, clinical populations, and embedded hardware.

DATA AND CODE AVAILABILITY

The source code used to preprocess the signals, train the models, and reproduce all experiments and figures, together with the resulting tables and performance metrics, is publicly available at:

https://github.com/Sergio-Jr86/analysis_comparative_grf

REFERENCES

- [1] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, 4th ed. Hoboken, NJ, USA: John Wiley & Sons, 2009.
- [2] D. Levine, J. Richards, and M. W. Whittle, *Whittle's Gait Analysis*, 5th ed. Edinburgh, UK: Churchill Livingstone Elsevier, 2012.
- [3] J. M. Burnfield, C. D. Few, O. S. Mohamed, and J. Perry, "The influence of walking speed and footwear on plantar pressures in older adults," *Clinical Biomechanics*, vol. 19, no. 1, pp. 78–84, 2004.
- [4] J. Taborri, E. Palermo, S. Rossi, and P. Cappa, "Gait partitioning methods: A systematic review," *Sensors*, vol. 16, no. 1, p. 66, 2016.
- [5] W. R. Johnson, A. Mian, M. A. Robinson, J. Verheul, D. G. Lloyd, and J. A. Alderson, "Multidimensional ground reaction forces and moments from wearable sensor accelerations via deep learning," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 1, pp. 289–297, 2021.
- [6] 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.
- [7] M. S. B. Hossain, Z. Guo, and H. Choi, "Estimation of lower extremity joint moments and 3D ground reaction forces using IMU sensors in multiple walking conditions: A deep learning approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 6, pp. 2829–2840, 2023.
- [8] 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.
- [9] S. Bogaert, J. Davis, and B. Vanwanseele, "Predicting vertical ground reaction force characteristics during running with machine learning," *Frontiers in Bioengineering and Biotechnology*, vol. 12, p. 1440033, 2024.
- [10] B. Song, M. Paolieri, H. E. Stewart, L. Golubchik, J. L. McNitt-Gray, V. Misra, and D. Shah, "Estimating ground reaction forces from inertial sensors," *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 2, pp. 595–608, 2025.
- [11] B. Yilmazgün, J. Weber, T. Stein, S. Sell, and B. J. Stetter, "Predicting 3D ground reaction forces across various movement tasks: A convolutional neural network study comparing different inertial measurement unit configurations," *Journal of Biomechanics*, vol. 192, p. 112888, 2025.
- [12] G. Grouvel, L. Carcreff, F. Moissenet, and S. Armand, "A dataset of asymptomatic human gait and movements obtained from markers, IMUs, insoles and force plates," *Scientific Data*, vol. 10, no. 1, p. 180, 2023.
- [13] D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, and S. N. Whittlesey, *Research Methods in Biomechanics*, 2nd ed. Champaign, IL, USA: Human Kinetics, 2014.
- [14] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, 1999.
- [15] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, p. 91, 2006.
- [16] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. [Online]. Available: <https://jmlr.org/papers/v7/demsar06a.html>