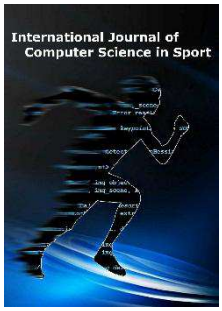
International Journal of Computer Science in Sport

Volume 24, Issue 2, 2025

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2025-0014



Evaluating the Influence of Sensor Configuration and Hyperparameter Optimization on Wearable-Based Knee Moment Estimation During Running

Höschler, L.[1], Halmich, C.[2,3], Schranz, C.[3], Koelewijn, A.D.[4], Schwameder, H.[1]

[1] *Department of Sport and Exercise Science, University of Salzburg*

[2] *Department of Artificial Intelligence and Human Interfaces, University of Salzburg*

[3] *Salzburg Research Forschungsgesellschaft mbH*

[4] *Department of Electrical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg*

Abstract

Wearable sensors combined with machine learning (ML) offer a promising approach for estimating joint kinetics in real-world settings, with potential applications in athlete monitoring and injury prevention. However, the variety of sensor configurations in previous studies complicates comparisons and optimal configuration selection. This study compared different wearable sensor configurations, comprising inertial measurement units (IMUs) and pressure insoles (PIs), to determine their influence on the accuracy of ML–based predictions of 3D knee moments during running. Sensor configurations ranged from one to four IMUs, with and without PIs. The dataset consisted of wearable and ground truth knee moment data from 19 recreational runners during treadmill running. Model performance of the convolutional neural networks was evaluated on an independent test set. Hyperparameter optimization (HPO) was applied to refine model architectures and training parameters. Performance gains by PIs and a greater number of IMUs were small but significant. The results after HPO confirmed similar performances between single- and multi-sensor configurations, suggesting only small benefits from additional sensors. Our findings highlight that both sensor configuration and model optimization play critical roles in achieving optimal performance. We provide practical recommendations for sensor selection, balancing accuracy and feasibility, to enable biomechanical assessments in real-world environments.

KEYWORDS: IMU, PRESSURE INSOLE, MACHINE LEARNING, JOINT KINETICS.

Introduction

Wearable sensors such as inertial measuring units (IMUs) and pressure insoles (PIs) have shown great potential for estimating joint biomechanics outside of laboratory settings, which could improve accessibility and reduce costs (Benson, Räisänen, Clermont, & Ferber, 2022; Caldas, Mundt, Potthast, Buarque de Lima Neto, & Markert, 2017; Gurchiek, Cheney, & McGinnis, 2019; C. J. Lee & Lee, 2022; Mundt, 2023; Weygers et al., 2020). Originally developed for vehicle navigation, IMUs are now small lightweight sensors that can be unobtrusively attached to the body with tape, straps, or tight-fitting fabrics (Cereatti et al., 2024). They measure three-dimensional linear accelerations, angular velocities, and magnetic field changes, enabling biomechanical analysis in various environments. IMU configurations in biomechanical studies range from complex 17-sensor full-body networks (Khurelbaatar, Kim, Lee, & Kim, 2015) to single sensors attached to the foot (Hossain, Zhishan, & Hwan, 2022), shank (Long et al., 2023), or pelvis (Lim, Kim, & Park, 2019). PIs are insole-like sensors that are inserted into footwear to measure plantar pressure during activity. While PIs allow for lab-independent measurements of ground contact kinetics, they are used less frequently than IMUs due to higher costs and practical challenges in integration and data collection (Adesida, Papi, & McGregor, 2019; Prasanth et al., 2021; Zhang, Novak, Brouwer, & Li, 2013).

Today, the lab-independent measurement of joint and segmental kinematics (angle, position and their time derivatives: velocity and acceleration) using IMUs is a largely solved problem. Research-grade and commercial solutions leverage sensor fusion techniques, such as extended Kalman filters and strapdown integration, to estimate kinematics reliably (Weygers et al., 2020; Zhang et al., 2013). In contrast, assessing joint kinetics, such as joint forces and joint moments, is more challenging. Unlike kinematics, which can rely on direct sensor data from IMUs, kinetics estimation requires complex physical modeling of forces and moments acting on the body, involving assumptions about joint mechanics, segmental properties, and ground reaction forces (Whittlesey & Robertson, 2004). While some studies have demonstrated simplified methods to estimate kinetics (Khurelbaatar et al., 2015), there is no straightforward, off-the-shelf solution comparable to those for kinematics. However, kinetic information is crucial for athlete monitoring and injury prevention. For example, an elevated knee abduction impulse is an injury risk factor in distance running (Stefanyshyn, Stergiou, Lun, Meeuwisse, & Worobets, 2006).

Most previous studies that propose wearable sensor-based methods for assessing kinetics use physics-based approaches that require extensive sensor configurations and careful sensor-to-segment calibration to achieve accurate results (C. J. Lee & Lee, 2022). Alternatively, machine learning (ML) methods have shown to provide more accurate results, even with relatively sparse sensor configurations (Carter, Chen, Cazzola, Trewartha, & Preatoni, 2024; Hossain et al., 2022; Matijevich, Scott, Volgyesi, Derry, & Zelik, 2020; Moghadam, Yeung, & Choisne, 2023; Mundt et al., 2020; Stetter, Krafft, Ringhof, Stein, & Sell, 2020; van Hooren, van Rengs, & Meijer, 2024). Rather than modeling physical principles, ML establishes statistical relationships between wearable sensor data and kinetics. By learning an optimized mapping function between inputs (sensor readings) and targets (joint kinetics) (Goodfellow, Bengio, & Courville, 2016), ML eliminates the need for complex biomechanical assumptions, enabling faster inference and reduced processing work.

Despite the growing body of research on wearable based biomechanics, sensor configurations vary widely across studies. Some studies rely on multi-IMU configurations with two (Stetter et al., 2020), three (Mundt et al., 2020), or four (Dorschky et al., 2020) different sensor locations, while others have used single IMUs attached to the shank (Long et al., 2023) or pelvis (Lim et al., 2019). Others combine IMUs with PI sensors (Carter et al., 2024; Höschler et al., 2025;

Matijevich et al., 2020; McCabe, van Citters, & Chapman, 2022; van Hooren et al., 2024). With each study typically using a unique configuration for its specific research goals, comparisons between methods are challenging. So far, no best practices regarding sensor type and placement have emerged. Establishing evidence-based guidelines is essential, as sensor configuration choice may affect accuracy, generalizability, and practicality for real-world applications.

Only few previous studies have systematically compared the influences of different sensor configurations on the accuracy of joint kinetics estimation during walking and running. Dorschky, Nitschke et al. (2025) report that adding more sensors improves their (physics-based) model's accuracy and recommend a lower-body configuration with at least foot, thigh, and pelvis IMUs for optimal performance. In contrast, Moghadam et al. (2023), have observed only small improvements when adding more sensors beyond a single shank or foot IMU. Using a combination of two IMUs and PIs, Matijevic et al. (2020), have found the highest ML prediction accuracy when all sensors are included. They have identified PIs as the primary contributor to model accuracy. Similarly, Carter et al. (2024) have combined each of three upper-body IMU locations (chest, wrist, pelvis) with a foot IMU and PIs, finding comparable performance across locations, with only the chest IMU yielding slightly better results than the full sensor configuration. Overall, findings are inconclusive: some studies emphasize the value of more IMUs, while others highlight the importance of PIs or suggest minimal to no accuracy gains from additional sensors.

Balancing accuracy and practicality of minimal sensor configurations presents a key dilemma when using wearable sensors in biomechanics research. Fewer sensors enhance usability by reducing costs, setup complexity, participant obtrusion, and risk of data loss due to sensor malfunction. According to ML principles, increasing the amount of input information (i.e., by using larger sensor configurations) would typically improve the model accuracy (Goodfellow et al., 2016). However, sensors that contribute redundant or noisy data could lead to overfitting and increased computational complexity. A fair comparison of different sensor configurations requires optimizing model architecture and training parameters for each configuration separately to prevent biases (Bischl et al., 2023). Creating a successful solution therefore depends on finding a minimal yet effective sensor configuration, combined with appropriate model optimization.

Building on these results, this study aimed to identify the most effective sensor configuration and further optimize performance. Specifically, we investigated how different combinations of wearable sensors, including four IMUs and a pair of PIs, influence model accuracy. First, we compared the sensor configurations using a baseline ML model with consistent training settings to quantify the effects of PIs and the number of IMUs. We hypothesized that the inclusion of PIs would significantly improve prediction accuracy and expected no significant performance improvements with greater numbers of IMUs. Second, we optimized the model hyperparameters for selected, promising sensor configurations identified in the initial evaluation. Thereby, we aimed to refine the accuracy further and provide practical recommendations for wearable based ML applications.

In a previous study, we developed an ML-based method to estimate joint kinetics during running based on combined input from four lower-body IMUs and a pair of PIs (Höschler et al., 2025). Building on this, the present study aimed to identify the most effective sensor configuration and further optimize performance. Specifically, we investigated how PIs and different numbers of IMUs influence prediction accuracy. First, we compared sensor configurations using a baseline ML model with consistent training settings. We hypothesized that the inclusion of PIs would significantly improve prediction accuracy and expected no significant performance improvements with greater numbers of IMUs. Second, we optimized the model hyperparameters

for selected, promising sensor configurations to refine accuracy further and provide practical recommendations for wearable-based ML applications.

Methods

Part 1: Baseline Comparison

Dataset

The dataset for this study is described in detail in our previous publication (Höschler et al. 2025). It consisted of synchronized continuous 3D knee moment, IMU, and PI data of 19 recreational runners (7 females, 12 males) during various treadmill running conditions using three different types of running footwear. With each shoe condition, 60 seconds of running were recorded at three slopes (level 0 % and ± 5 %) and three speeds (self-selected speed ± 1 km/h).

From the force and marker data of the instrumented treadmill (Gait 3D, h/p/- Cosmos, Traunstein, Germany) and the motion capture system (Vicon Nexus, Oxford, United Kingdom) ground truth reference 3D knee moments were calculated via inverse dynamics and normalized to the participant's body mass. Moment data were down sampled from 200 to 100 Hz to match the sampling rate of the PIs. The participants were equipped with a pair of high resolution PIs (Intelligent Insoles Pro, XSens, Calgary, AB, Canada) and seven IMUs (Wavetrack, Cometa, Milano, Italy; sampling rate 2000 Hz, range ± 16 g and ± 2000 deg/s). The IMUs were placed on the participant's bilateral dorsal feet, medial shanks, lateral thighs, and sacrum (Figure 1). From the PI data, sampled at 100 Hz, time-series of mean pressure (normalized to body mass) were calculated for seven functional foot segments (medial anterior/posterior forefoot, lateral forefoot, medial/lateral midfoot, and medial/lateral rearfoot) along with center of pressure coordinates (antero-posterior, medio-lateral). The 3D angular velocity and linear acceleration signals from each IMU were down sampled from 2000 to 400 Hz, low-pass filtered with a 2nd order bi-directional Butterworth filter with custom cut-off frequencies per channel (Yu, Gabriel, Noble, & An, 1999), and z-score normalized. For the prediction of knee moments of a respective body side, only the data of sensors at the corresponding stance leg plus pelvis IMU were used (i.e., left PI, left foot, left shank, left thigh and pelvis IMU for the prediction of left knee moments). All preprocessing steps were implemented in Python 3.9 using custom scripts.



Figure 1. Placement of the wearable sensors. Seven inertial measuring units (IMUs, blue) were placed on the following locations: bilateral dorsal feet, medial shanks, lateral thighs, and sacrum. A pair of pressure insoles (PIs, green) were inserted bilaterally into the footwear at both sides.

Sensor Configurations & Baseline Model Architecture

The combination of the four IMU positions together with PIs resulted in a total of 31 possible configurations. The sensor configurations were abbreviated based on the combination of sensors included, using the following notation: foot (F), shank (S), thigh (T), pelvis (P), and pressure insoles (indicated with "+").

Given the success of convolutional neural networks (CNNs) in predicting biomechanical time-series data (Dorschky et al., 2020; Liew et al., 2021; Mundt et al., 2021), a CNN based architecture was chosen and implemented using `PyTorch`. The baseline model consisted of seven one-dimensional convolutional layers (`torch.nn.Conv1d`), designed to map IMU and PI data to 3D knee moments (Figure 2). For configurations that included both sources, IMU (Figure 2, blue) and PI (Figure 2, green) data were processed separately in a first layer. To compensate for the four times higher sampling rate of the IMU data, the IMU layer used a kernel four times wider than the PI layer, with a stride of 4. After this initial feature extraction, the outputs from both layers were concatenated and merged into a dedicated layer for simultaneous feature learning (Figure 2, green & blue). This was followed by two auxiliary layers (Figure 2, purple) that incorporated information on running slope (level, uphill, downhill, encoded as [0,0], [1,0], [0,1], respectively). Finally, the model passed through two additional Conv1D layers (Figure 2, yellow) before outputting predictions.

To handle variations in sensor configuration, the IMU layer dynamically adjusted its input dimensions based on the available number of IMUs (n_{IMUs}), specifically $n_{IMUs} \times 6$ (accounting for 3D accelerometer and 3D gyroscope data per IMU). If a configuration used only one data source (IMU or PI), the unused layer was simply omitted. The ReLU activation function was used for all, but the final layer.

To prevent overfitting, dropout regularization with a probability of 0.2 was applied to all layers except the initial and auxiliary layers. Additionally, random noise, scaled to 0.1 of the value range for each IMU and PI sample, was introduced to enhance generalizability. The hyperparameters of the baseline architecture are shown in Table 1. The model was trained using 10-second data slices to facilitate rapid application on long data recordings.

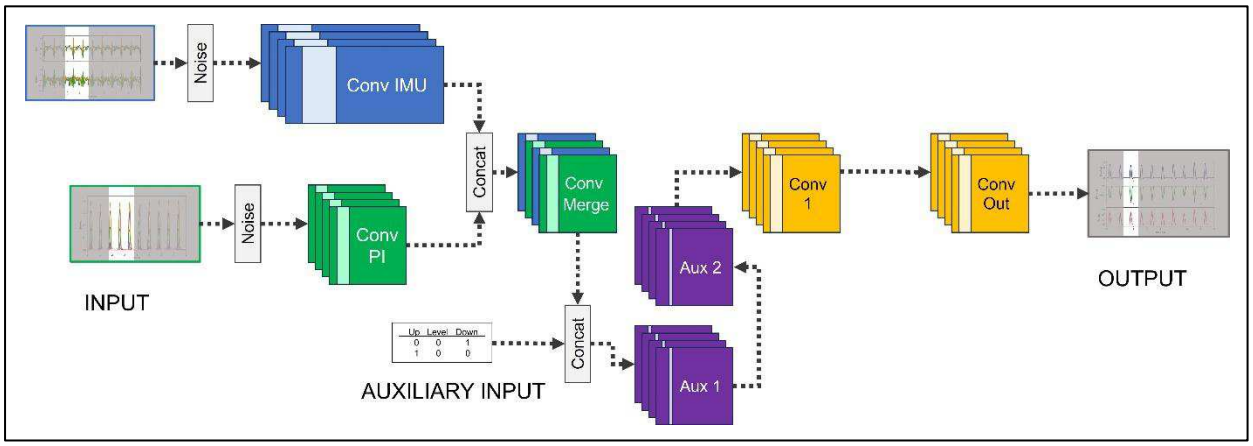


Figure 2. Illustration of the baseline model architecture consisting of multiple 1D convolutional layers. First noise is added to the inputs of IMU (blue) and PI (green) before they are processed in individual layers. Outputs are concatenated, then passed through a merging layer (green-blue). Slope information is incorporated as auxiliary input before passing through two auxiliary layers (purple) and two additional layers (yellow) before generating 3D knee moment outputs.

Training & Evaluation

Model training was conducted on an RTX A6000 GPU with 48 GB RAM (Nvidia, Santa Clara, CA, United States). The baseline models for each sensor configuration were trained for 20 epochs with a batch size of 64 using an Adam optimizer. A cyclic learning rate scheduler (`OneCycleLR`) with a base learning rate of 0.001 was employed, increasing to 0.004 every 5 epochs. A custom loss function based on the rooted `torch.nn.MSELoss` was used to calculate the loss at each iteration. This function calculated a weighted mean across the losses in three dimensions (sagittal, frontal, transverse), with losses in the non-sagittal planes scaled by a factor of 7 to accelerate learning in the minor motion planes (Höschler et al., 2025). The models were trained using a leave-one-subject-out cross-validation procedure, where the data of each subject (n = 19) was withheld from the training set to serve as validation set in turn. All models were trained on the same data slices to ensure consistency across sensor configurations. A controlled random seed was used to ensure reproducibility and prevent variations in data allocation. Model performance for each sensor configuration was evaluated by the mean (non-weighted) RMSE between the estimated and ground truth knee moment time-series across the three dimensions averaged over all validation runs.

Table 1. Baseline hyperparameters of the 1D convolutional layers. n_{IMUs} – number of IMUs in the respective configurations. * IMU kernel size is multiplied by 4 due to higher sampling rate.

Parameter	IMU	PI	Merge	Aux 1	Aux 2	Conv 1	Out
In channels	$n_{IMUs} \times 6$	9	64	66	64	64	64
Out channels	32	32	64	64	64	64	3
Kernel size	51*	51	27	1	1	15	7
Stride	4	1	1	1	1	1	1
Groups	1	1	4	1	1	4	1

Data Analysis

After reviewing the results of the cross-validation, three configurations (P+, P, and +) were excluded from further analysis due to their poor performance. On the remaining 28 configurations, a linear mixed effects models (LMM) approach was used to examine the effect of PIs (binary: 0 or 1) and the number of IMUs (n_{IMUs}, categorical: 1, 2, 3, or 4) on the prediction accuracy (RMSE). Therefore, the data was organized into a long format, with each sensor configuration (with and without PI, and across varying n_{IMUs}) treated as an observation for each subject. Each statistical model included a random intercept for subjects to account for individual variability.

LMM was chosen due to the multilevel structure of the data and the uneven distribution of n_{IMUs} across the configurations. The n_{IMUs} per configuration ranged from 1 to 4, with unequal representation in each category (2 configurations with 4 IMUs, 8 configurations with 3 IMUs, 12 configurations with 2 IMUs, and 6 configurations with 1 IMU). This imbalance, combined with the hierarchical nature of the data, made LMM the most appropriate method to account for both fixed effects (i.e., PI and n_{IMUs}) and random effects (i.e., subject-specific variability). Five models with increasing complexity were fit to determine the best explanation of RMSE:

1. A null model (*Model_0*) containing only an intercept and a random effect for subjects.
2. A model (*Model_1.1*) with a fixed effect for PI and a random effect for subjects.
3. A model (*Model_1.2*) with a fixed effect for n_{IMUs} and a random effect for subjects.
4. A model (*Model_2*) with fixed effects for both PI and n_{IMUs}, and a random effect for subjects.
5. A full interaction model (*Model_3*) with both fixed effects and their interaction term, as well as a random effect for subjects.

Using ANOVA as Likelihood Ratio Test, a stepwise model comparison was performed to identify the best-fitting model. Improvements in model fit were evaluated pairwise by changes in log-likelihood values and their statistical significance as well as by lower Akaike Information Criterion (AIC) and Bayesian Information Criterion values (BIC) (Meteyard & Davies, 2020). The assumptions of constant variance (homoscedasticity) and normal distribution of the residuals were validated by visual inspection of the residual fit, histogram, and quantile-quantile plots. Using marginal R^2 values (Nakagawa & Schielzeth, 2012), we calculated effect size as Cohen's f^2 for the best fitting LMM model. Overall effect size was obtained with fixed effects included (R^2_{full}) (1), while local effects for PIs and n_{IMUs} were calculated using (2), with the respective fixed effect removed from the LMM model ($R^2_{reduced}$). Effect sizes above 0.02, 0.15, and 0.35 were interpreted as small, medium, and large, respectively (Cohen, 1988). Statistical analyses were conducted using **R** (version 4.3), utilizing the `nlme` package.

$$f^2_{overall} = \frac{R^2_{full}}{1 - R^2_{full}} \quad (1)$$

$$f^2_{local} = \frac{R^2_{full} - R^2_{reduced}}{1 - R^2_{full}} \quad (2)$$

Part 2: Optimization of Selected Models

Model Selection

To further improve model performance and thus enable future practical applications, the model

architecture and training parameters of selected sensor configurations were optimized in the second part of the study. Due to computational limitations, an optimization of all 31 configurations was not feasible. Therefore, the most promising configurations were selected based on the results of part 1. The overall goal was to achieve the best accuracy, while minimizing the number of sensors, aiming to keep the final configuration as simple and unobtrusive as possible. Only configurations without PIs were chosen for the optimization, as using two systems (IMUs and PIs) would increase the measurement and processing complexity, including challenges with synchronization and data fusion. Moreover, relying on multiple systems would raise the risk of data loss, as was observed during data collection with frequent malfunctions of the PI system (Höschler et al., 2025). Based on the results of part 1, the two best-performing configurations (with the lowest RMSE) were selected for each n_{IMUs} (1, 2, and 3 IMUs), along with the configuration with all 4 IMUs.

Optimization

For the optimization, the model architecture and training settings were refined simultaneously, as recommended by Bischl et al. (2023). After removing PI data input, we assessed the necessity of the merge layer, which originally combined time-series data from both sensor sources, before the auxiliary layer. Additionally, up to two extra layers could be added before the output layer, with opportunities to optimize the number of channels in specific layers. Specifically, the number of output channels of the IMU layer, the first auxiliary layer, the first Conv1D layer after the auxiliary layers, and the potential extra layers could be adjusted. Finally, kernel sizes could be modified for all layers except the auxiliary layers (Table 2).

Table 2. Optimized hyperparameters, search spaces, and, if applicable, layers, to which the parameter applies, lr_{min}: minimum learning rate, lr_{max}: maximum learning rate. *only if layer exists, **IMU kernel size is multiplied by 4 due to higher sampling rate, n.a.: not applicable (for the general training parameters).

Parameter	Search Space	Applicable Layers
Exists	{True, False}	Merge, Extra 1, Extra 2
Out Channels	{32, 64, 128, 256}	IMU, Aux 1, Conv 1, Extra 1*, Extra 2*
Kernel Size	{7, 15, 25, 51}	IMU**, Merge*, Conv 1, Extra 1*, Extra 2*, Out
Epochs	{25, 50}, increment of 5	n.a.
Batch Size	{8, 16, 32, 64, 128}	n.a.
lr_{min}	$\{10^{-4}, 10^{-3}\}$	n.a.
lr_{max} ratio	{1, 10}	n.a.

In addition to the architectural adjustments, the following training parameters were optimized: the number of epochs, batch size, minimum learning rate, and maximum learning rate, set to a floating point multiple of the minimum learning between one and ten.

In hyperparameter optimization (HPO), the vast number of possible configurations with more than three parameters makes exhaustive search methods, like grid search, computationally impractical. Sequential algorithms, such as gradient-based or probabilistic methods, address this by focusing on promising regions of the search space, which helps them find suitable

hyperparameters faster by learning from previous iterations. Therefore, we used a Tree-structured Parzen Estimator (TPE) sampler implemented in the `Optuna` package, which uses Bayesian optimization to efficiently explore the search space of all possible parameter combinations to find the combination with the highest potential for performance improvement (Watanabe, 2023).

For each of the seven selected sensor configurations, a typical HPO study with 500 iterations and repeated resampling was conducted to refine the model architecture and training parameters. At each iteration, a combination of hyperparameters from the search space was suggested. Model performance for each of these combinations was evaluated using a three-fold cross-validation, with the same participants consistently assigned to each fold. The optimization criterion was the minimal mean validation loss (weighted mean RMSE) across the three folds. A Hyperband pruner (`optuna.pruners.HyperbandPruner`) was used to terminate underperforming trials early, reallocating resources to more promising combinations to improve efficiency. Log files were used to track all executed parameter combinations and their respective performance.

Evaluation

For each of the seven selected sensor configurations, two models were trained on the training dataset (with data from 19 participants): (1) using the baseline model configuration, and (2) using the best performing configuration from the HPO study. Model performance was evaluated on an independent test set that was created using data of five new participants from the original experiment. These participants were excluded from the initial dataset due to incomplete data caused by PI malfunctions. However, since the sensor configurations selected for the hyperparameter optimization study did not rely on PI data, their data were suitable for use as a test set. To comprehensively evaluate model performance, RMSE, normalized RMSE (nRMSE), and intra-class correlation coefficient (ICC, two-way mixed effects, absolute agreement, single measurement, single rater) were calculated for each dimension (sagittal, frontal, transverse) for continuous predictions (CONT) and during stance-phases only (PHSS) (Höschler et al., 2024). PHSS results were calculated because stance-phase joint moments are of primary relevance from a biomechanical perspective. To calculate nRMSE, the RMSE was normalized to the value range (max-min) of the reference data in each sample for the respective dimension.

Results

Part 1: Baseline Comparison

Evaluation Results

The results of the baseline comparison (Figure 3) showed that, with three exceptions (P, P+, and +), all sensor configurations resulted in similar model performance, with a mean RMSE across all three dimensions below 0.14 Nm/kg. Overall, the FST+ configurations had the lowest RMSE with 0.122 Nm/kg.

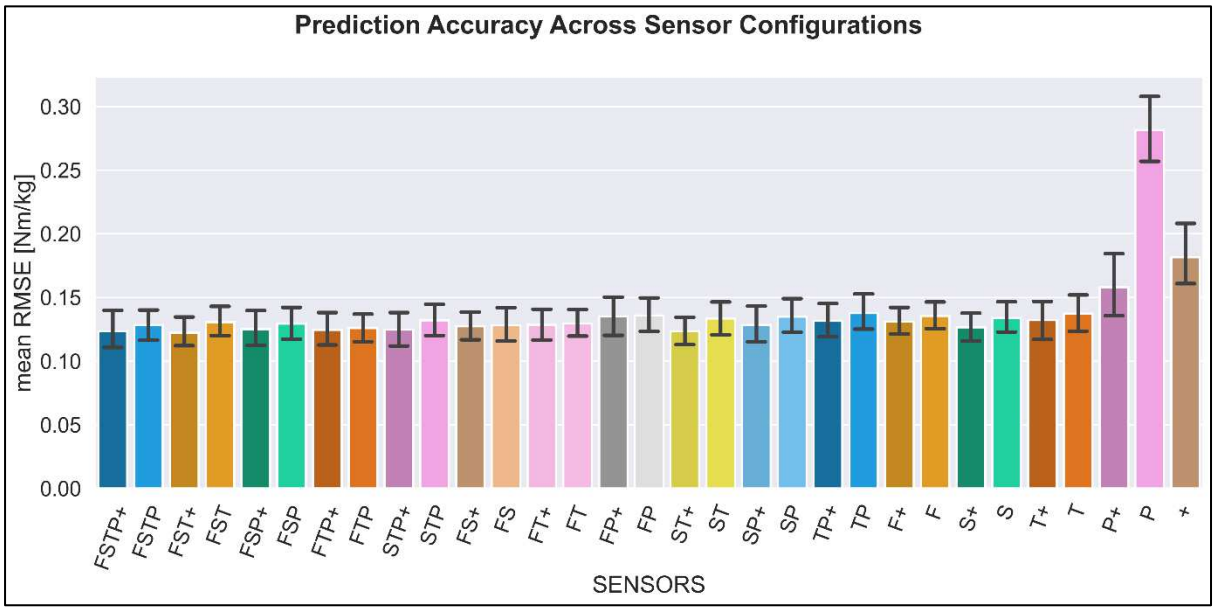


Figure 3. Mean RMSE (across the three dimensions) after cross-validation (n=19) of all 31 configurations using the baseline model settings. Caps indicate the standard deviation. The abbreviation of the configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P, PIs: +). Configurations using the same IMUs with and without PIs are colored similarly and placed side by side.

Influence of Pressure Insoles and Number of IMUs

The results of the stepwise model comparison are reported in the appendix (Table A1). *Model_2*, which includes fixed effects for both PI and n_{IMUs}, was the best-fitting LMM model, providing the lowest AIC, BIC, and a significant likelihood ratio test ($p < 0.001$) compared to simpler models. Adding an interaction term (*Model_3*) did not significantly improve the fit ($p = 0.97$). Therefore, the results of *Model_2* are reported (Figure 4). Compared to the intercept (0.135 Nm/kg, PI: False, 1 IMU), the addition of PIs was associated with a slight reduction in RMSE by 0.005 Nm/kg (3.6 %), indicating a significantly improved model accuracy with minimal effect size ($f^2 = 0.008$). While adding a second IMU did not significantly reduce RMSE ($p = 0.29$), increasing to three and four IMUs led to a statistically significant RMSE reduction by 0.006 (4.4 %, $f^2 = 0.011$) and 0.007 Nm/kg (5.1 %, $f^2 = 0.008$), respectively.

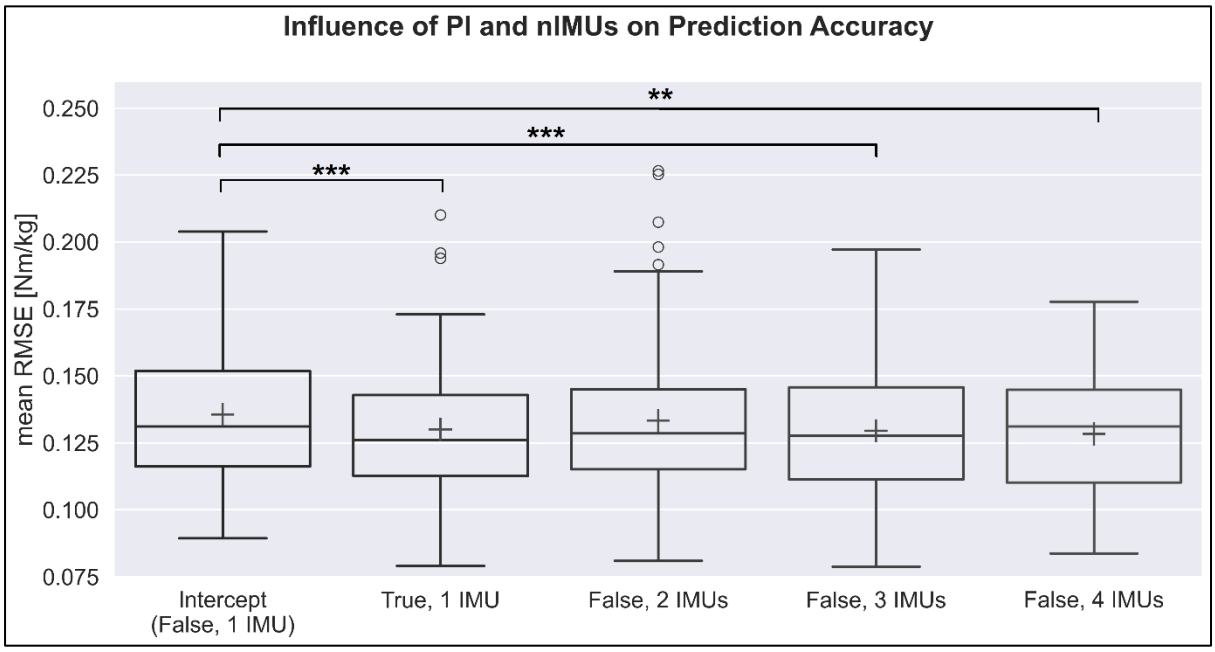


Figure 4. Mean RMSE (across the three dimension) of all validation runs (n=19) per configuration. The first and second boxes include three configurations each (excluding those with pelvis IMU). The third, fourth, and fifth boxes include six, three and one configurations, respectively. The boxes contain the lower and upper quartiles with median indicated as solid line. The whiskers display the 1.5 times interquartile range. Outliers are indicated by circles, the mean is indicated by a +. ** and *** indicate statistical significance of the fixed effects *PI* and *number of IMUs* compared to the intercept (PI: false, 1 IMU) below a p-value of 0.01 and 0.001, respectively.

Part 2: Optimization of Selected Models

Sensor Configuration Selection

For the HPO study, the following best-performing configurations from Part 1 without PIs for each n_{IMUs} were selected:

- 4 IMUs: FSTP (mean RMSE: 0.128 Nm/kg)
- 3 IMUs: FSP (mean RMSE: 0.129 Nm/kg), FTP (mean RMSE: 0.126 Nm/kg)
- 2 IMUs: FS (mean RMSE: 0.128 Nm/kg), FT (mean RMSE: 0.13 Nm/kg),
- 1 IMU: F (mean RMSE: 0.136 Nm/kg), S (mean RMSE: 0.134 Nm/kg)

Hyperparameters

The HPO produced a diverse set of model architectures, some of which differed considerably from the baseline. Full details of the optimized architectures and training parameters for all configurations are provided in the appendix (Tables A2 – A3).

While the merge layer (before the auxiliary layers) was retained in four of the seven cases, all but one optimized model included at least one additional layer before the output, leading to five models with more layers than the baseline. Additionally, the optimized models used a higher number of output channels in the IMU, merge, and Conv1 layers, resulting in substantially larger models overall (Figure 5). Various kernel size distributions emerged across the optimized models. While the baseline model used kernels with decreasing sizes in deeper layers, five out of seven optimized models adopted an ‘hourglass’ like distribution, with wide kernels in the shallower and deeper layers and narrow kernels in the middle. However, two models (FSP and S) showed an inverted pattern, using narrow kernels in the shallower and deeper layers and wide

kernels in the middle. Regarding the training parameters, the optimized models generally favored lower batch sizes (8–16) compared to the baseline (64), together with lower minimum and maximum learning rates. The median number of training epochs was 25.

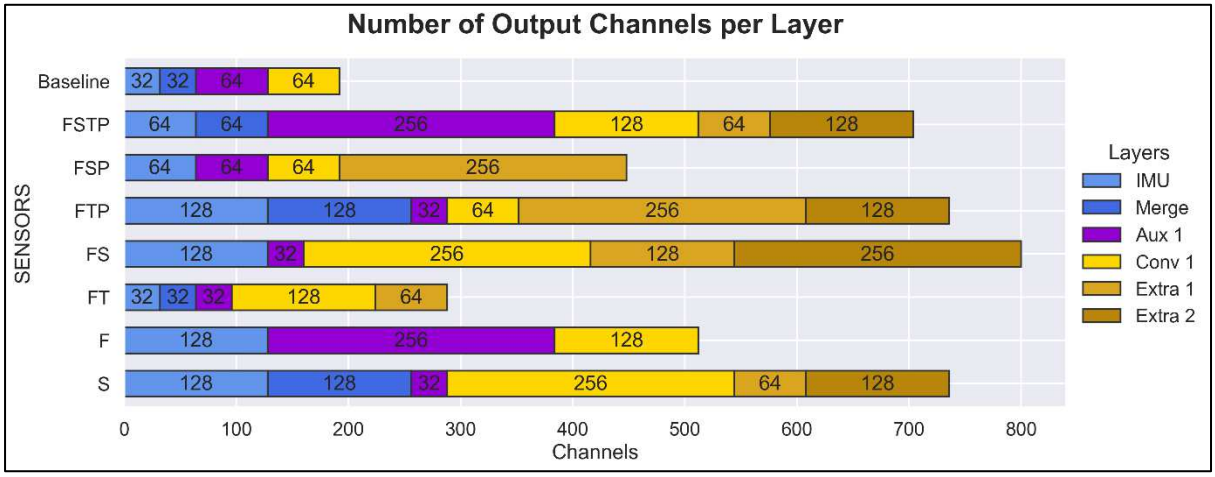


Figure 5. Number of output channels in 1D convolutional layers at baseline and after hyperparameter optimization. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P).

Model Performance

Performance results on the test set showed that the optimized models generally outperformed the baseline configurations in terms of nRMSE (Figure 6, Table A4 Appendix), ICC and RMSE (Figures A1 – A2, Tables A5 – A6 Appendix) across both prediction modes (CONT, PHSS).

The accuracy gains were consistent between CONT and PHSS. For all configurations, the lowest nRMSE and highest ICC values were observed in the sagittal plane (optimized mean PHSS: nRMSE = 0.11, ICC = 0.93), while the frontal plane predictions showed the lowest accuracy (nRMSE = 0.24, ICC = 0.56). Regarding RMSE, the sagittal plane predictions showed larger errors of 0.17 Nm/kg (optimized mean CONT), compared to 0.13 and 0.06 Nm/kg for the frontal and transverse planes, respectively. The configurations with the greatest performance improvements compared to baseline were F, FTP, and FSTP. Among the optimized configurations, FSP achieved the highest accuracy across all planes and both prediction modes, and FT the lowest. Overall, only minor performance differences were found between the optimized configurations.

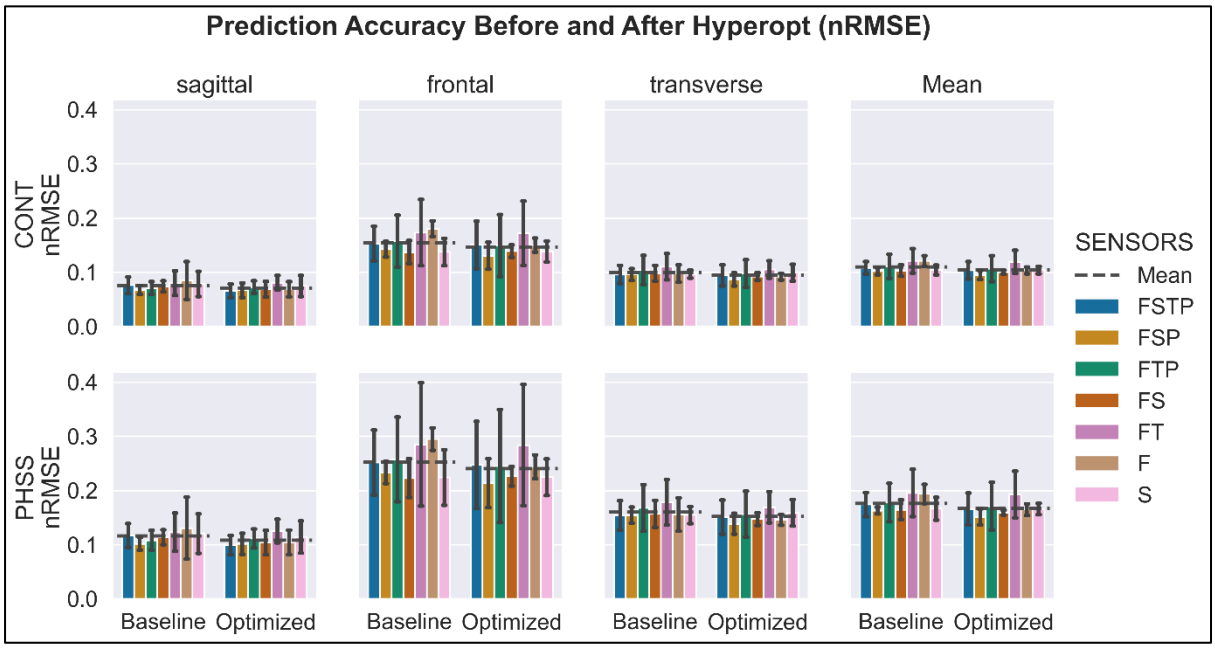


Figure 6. Test set performance (normalized RMSE) of the baseline and optimized models. Displayed are the results in the three planes of motion and the mean across them. Bar height indicates the mean, caps show the standard deviation. The abbreviation of the configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P). The mean across all configurations is displayed as dashed line. CONT – predictions over the entire sample of 10 seconds, PHSS – predictions during stance phases only.

Discussion

The primary goal of this study was to investigate the effect of different sensor configurations on the accuracy of an ML model predicting 3D knee kinetics during running in different conditions. In a second step, the model hyperparameters were optimized to refine accuracy and to support future practical applications. Confirming our first hypothesis, we found that adding PIs significantly improved model accuracy. In contrast to our second hypothesis, increasing n_{IMUs} also improved performance. However, the effect sizes were minimal, and the results of the optimization confirmed that different sensor configurations achieved very similar accuracy levels.

The findings of the baseline comparison have shown that combining multiple IMUs is important to achieve the highest accuracy. In line with foundational ML theory, adding more input information, such as extra sensor data from different body locations, typically improves model performance by allowing the algorithm to capture a broader range of independent data characteristics (Goodfellow et al., 2016). Our results align with findings from Dorschky, Nitschke et al. (2025), who found improved accuracy in their optimal control simulations with additional IMUs. Although significantly lower mean RMSE values were observed, the effect of n_{IMUs} was minimal, indicating that additional sensors contribute, but with limited impact.

After the HPO, performance differences between configurations containing one to four IMUs have become negligible, supporting the findings by Carter et al. (2024) and Moghadam et al. (2023), who observed only small improvements when increasing IMU counts beyond a single sensor. These slightly contrasting results between the two parts of this study may be explained by a potential bias of the baseline model favoring configurations with more IMUs. In the HPO phase, however, the best-fitting architecture and parameters were identified for each sensor configuration. From a physical-informed perspective it seems improbable that knee moments could be estimated accurately using data from only a single foot or shank IMU. Nevertheless,

consistent with other studies using single IMUs (Lim et al., 2019; Long et al., 2023), our results suggest that ML models can capture hidden relationships in the data to produce accurate estimates from minimal inputs when hyperparameters are carefully optimized.

Regarding PIs, our results highlight the importance of ground reaction kinetics data as model input, as configurations incorporating PIs have achieved a 0.01 Nm/kg (approximately 7 %) lower RMSE. This finding aligns with Matijevic et al. (2020), who found that PIs were essential for accurately estimating tibial impact kinetics compared to multiple IMUs alone. Although PIs are less frequently used in ML studies, these results suggest that they provided relevant information to the model. Two recent studies that included PIs in their configurations have shown promising results estimating running kinetics (Carter et al., 2024; van Hooren et al., 2024). A possible explanation is that PIs capture plantar pressure data which is closely related to the vertical ground reaction force and therefore beneficial to kinetic predictions, as GRFs are a main input to conventional inverse dynamics calculations. Together, our findings underscore that in situations where the highest possible accuracy is desired, sensor configurations combining multiple IMUs and PIs should be chosen.

Despite the performance benefits of combining IMUs and PIs, there are notable practical disadvantages of using different sensor types in one configuration. Such configurations are not only more costly and complex to implement, but they also introduce additional risks for data collection issues, particularly from sensor malfunctions and synchronization errors. During the data collection for this study, PI malfunctions led to substantial data loss, limiting their reliability in consistent data capture (Höschler et al., 2025). These drawbacks motivated us to focus on IMU-only configuration for the HPO, as IMUs provide a more robust, lower-cost approach that is better suited for real-world applications. Additionally, since most previous ML studies have relied solely on IMU data, using only IMUs offers easier dataset compatibility, allowing potential model improvements through larger, combined datasets. Furthermore, synthetic IMU data can be generated from marker data, making it possible to utilize motion capture archives to further increase data set sizes (Mundt et al., 2020).

The choice of sensor configuration depends on practical considerations around accuracy, usability, and cost. In most cases, configurations with fewer IMUs are preferable. Our results suggest that foot, shank, and pelvis placements offer the best balance of performance and practicality. The foot IMU, in particular, performs well across nearly all configurations and is minimal obtrusive, making it a common choice (Dorschky, Nitschke et al., 2025; Benson et al., 2022). Its placement near the center of GRF application likely provides valuable kinetic information. Consistent with its frequent use in biomechanical studies (Benson et al., 2022; van den Berghe, Six, Gerlo, Leman, & Clercq, 2019), the shank IMU demonstrates strong predictive capability, likely due to its high signal quality and minimal soft-tissue artifacts (Cereatti et al., 2024). In contrast, thigh IMUs are more susceptible to motion artifacts, which may explain the relatively poor performance of the FT configuration. The pelvis IMU, typically placed at the sacrum, is widely used due to its proximity to the body's center of mass (M. Lee & Park, 2020; Lim et al., 2019). However, alone it is less effective in predicting distal joint kinetics, requiring additional data from distal sensors for accurate estimations (Modghadam et al., 2023). Ultimately, while foot, shank, and pelvis IMUs yield the best results in this study, sensor selection should also consider the specific application, such as prioritizing accuracy in clinical settings. However, optimizing model architecture and hyperparameters remains essential for maximizing prediction accuracy within any chosen configuration.

As discussed in our previous work (Höschler et al., 2025), the proposed ML approach demonstrates that 3D knee moments can be accurately estimated continuously during running based on IMU and PI data. Regardless of sensor configurations, the highest accuracies are

achieved in the sagittal and transverse planes, while the frontal plane shows the greatest errors. The persistent relative error of approximately 15-24 % may be due to high inter-individual variability in this plane (Höschler et al., 2024; Höschler et al., 2025; Stetter et al., 2020), highlighting the need for further improvements before these methods can reliably assess injury risk factors such as elevated knee abduction impulse (Stefanyshyn et al., 2006).

However, some of these errors may be attributed to inherent limitations in marker-based biomechanics. These approaches rely on simplified assumptions regarding joint center and rotational axis definitions – particularly for complex joints like the knee – which makes them highly sensitive to marker placement and anatomical variation (Benoit et al., 2006). Since our models have been trained using marker-based data as a reference, it is possible that the wearable data itself contains additional information enabling the model to capture biomechanical patterns beyond the marker-based estimate. Prior studies from the medical image analysis domain show that ML models can generalize well even with imperfect or noisy labels (Karimi, Dou, Warfield, & Gholipour, 2020). Testing this hypothesis would require a true "gold standard", such as data from instrumented implants or bone-pin markers. However, these methods are highly invasive and raise ethical concerns. Alternatively, increasing dataset size by combining existing datasets or using data augmentation methods – such as synthesizing IMU data from legacy motion capture data (Mundt et al., 2020) – could help to achieve better results.

Interestingly, the HPO produced diverse model architectures across different configurations, without any apparent patterns in terms of merge layer, extra layers, kernel size and output channel distributions. However, most optimized models incorporated more layers, and a substantially larger number of channels compared to baseline, suggesting an increased model capacity. This implies that estimating knee moments from wearable sensor data is a complex task, and that the baseline model may have been too simple to fully capture the relationships in the data. The smaller batch size (median: 8) observed in the optimized models, paired with smaller learning rates compared to the baseline, may have introduced a regularizing effect beneficial to these more complex network architectures (Wilson & Martinez, 2003). Nevertheless, the baseline configuration showed only slightly reduced accuracy, compared to the optimized models. In general, model capacity should be sufficient to capture the hidden relationships in data, yet still simple enough to avoid overfitting (Goodfellow et al., 2016; Halilaj et al., 2018). The Conv1D-based architecture used in this study proved effective, while other studies have found similar results with different architectures such as LSTM (Carter et al., 2024), RF (Long et al., 2023), or GRU (Hossain et al., 2022). This suggests that the model type may be less critical for achieving good performance when architecture and training parameters are optimized. Model design is a time-demanding, iterative process that benefits from interdisciplinary domain knowledge and experience (Goodfellow et al., 2016). Our findings suggest that establishing a solid baseline architecture with a relatively small search space for HPO could be a practical, efficient approach to design successful ML models.

The scope of this study was limited to knee kinetics during running, which are particularly relevant given the high incidence of running induced knee overuse injuries (Francis, Whatman, Sheerin, Hume, & Johnson, 2019). Nevertheless, we cannot be certain that our findings generalize to other joints and movements. Some studies examining multiple joints report consistent results across joints (Carter et al., 2024; Long et al., 2023; Moghadam et al., 2023), whereas others found that performance varies across joints when using different sensor configurations (Dorschky, Nitschke et al., 2025). However, we believe that the established workflow can be applied to other joints and movements, with configurations using foot, shank, and pelvis IMUs likely delivering accurate results. Additionally, due to computational constraints, only a subset of sensor configurations was optimized in this study. As a result, we

cannot conclusively determine whether other configurations might have yielded better performance had they undergone HPO. Researchers interested in testing additional configurations can readily do so, as we provide open access to the data, code and methods used in our analysis. This transparency allows others to further explore and optimize sensor configurations, potentially extending the findings of this study.

Conclusion

This study demonstrates that combining PIs with multiple IMUs yields the highest accuracy for ML-based knee moment estimation in running. However, the performance improvements by additional sensors, particularly PIs, are small and come at the cost of added complexity, higher costs, and increased risks of data collection issues. Our findings suggest that practical configurations, such as foot, shank and pelvis IMUs, offer a balance between accuracy and feasibility. Further optimization of model architectures and hyperparameters can enhance performance even with sparse sensor configurations.

Despite these advances, challenges remain in translating wearable-based joint kinetics estimation into real-world applications. Future research should explore strategies to further improve prediction accuracy, particularly in the frontal plane, and assess the generalizability of models across diverse populations. Expanding dataset size through augmentation techniques or combination of existing datasets may further improve robustness. Ultimately, wearable based methods are expected to complement lab-based assessments of joint kinetics, offering new possibilities for continuous load monitoring and management for runners of all levels during real-world activity.

Open Science

A subset of the dataset of this study is available at: https://zenodo.org/records/11119845

The python code for ML training and HPO is available under:
https://github.com/luceskywalker/TexSense

Acknowledgements

The project TexSense is funded within the context of WISS 2025, der Wissenschafts- und Innovationsstrategie 2025, by the federal state of Salzburg (grant 20102-445 F2101127-FPR).

We acknowledge Jannic Wälde's support during the statistical analysis.

References

Adesida, Y., Papi, E., & McGregor, A. H. (2019). Exploring the Role of Wearable Technology in Sport Kinematics and Kinetics: A Systematic Review. Sensors (Basel, Switzerland), 19(7). https://doi.org/10.3390/s19071597

Benoit, D. L., Ramsey, D. K., Lamontagne, M., Xu, L., Wretenberg, P., & Renström, P. (2006). Effect of skin movement artifact on knee kinematics during gait and cutting motions measured in vivo. *Gait & Posture*, *24*(2), 152–164. https://doi.org/10.1016/j.gaitpost.2005.04.012

Benoit, D. L., Ramsey, D. K., Lamontagne, M., Xu, L., Wretenberg, P., & Renström, P. (2006). Effect of skin movement artifact on knee kinematics during gait and cutting motions measured in vivo. Gait & Posture, 24(2), 152–164. https://doi.org/10.1016/j.gaitpost.2005.04.012

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., others, & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2). https://doi.org/10.1002/widm.1484

Caldas, R., Mundt, M., Potthast, W., Lima Neto Buarque de F., & Markert, B. (2017). A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. *Gait & Posture*, *57*, 204–210. https://doi.org/10.1016/j.gaitpost.2017.06.019

Carter, J. A., Chen, X., Cazzola, D., Trewartha, G., & Preatoni, E. (2024). Estimating joint moments during treadmill running using various consumer based wearable sensor locations. *ISBS Proceedings Archive*, *42*(1), 152. https://commons.nmu.edu/isbs/vol42/iss1/76

Cereatti, A., Gurchiek, R., Mündermann, A., Fantozzi, S., Horak, F., Delp, S., & Aminian, K. (2024). ISB recommendations on the definition, estimation, and reporting of joint kinematics in human motion analysis applications using wearable inertial measurement technology. *Journal of Biomechanics*, *173*, 112225. https://doi.org/10.1016/j.jbiomech.2024.112225

Cohen, J. (1988) Statistical Power Analysis for the behavioral sciences. New York: Psychology Press.

Dorschky, E., Nitschke, M., Mayer, M., Weygers, I., Gassner, H., Seel, T., others, & Koelewijn, A. D. (2025). Comparing sparse inertial sensor setups for sagittal-plane walking and running reconstructions. *Frontiers in Bioengineering and Biotechnology*, *13*. https://doi.org/10.3389/fbioe.2025.1507162

Dorschky, E., Nitschke, M., Martindale, C. F., van den Bogert, A. J., Koelewijn, A. D., & Eskofier, B. M. (2020). CNN-Based Estimation of Sagittal Plane Walking and Running Biomechanics From Measured and Simulated Inertial Sensor Data. Frontiers in Bioengineering and Biotechnology, 8, 604. https://doi.org/10.3389/fbioe.2020.00604

Francis, P., Whatman, C., Sheerin, K., Hume, P., & Johnson, M. I. (2019). The Proportion of Lower Limb Running Injuries by Gender, Anatomical Location and Specific Pathology: A Systematic Review. *Journal of Sports Science & Medicine*, *18*(1), 21–31. https://www.jssm.org/volume18/iss1/cap/jssm-18-21.pdf

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Gurchiek, R. D., Cheney, N., & McGinnis, R. S. (2019). Estimating Biomechanical Time-Series with Wearable Sensors: A Systematic Review of Machine Learning Techniques. *Sensors (Basel, Switzerland)*, *19*(23). https://doi.org/10.3390/s19235227

Halilaj, E., Rajagopal, A., Fiterau, M., Hicks, J. L., Hastie, T. J., & Delp, S. L. (2018). Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. Journal of Biomechanics, 81, 1–11. https://doi.org/10.1016/j.jbiomech.2018.09.009

Höschler, L. (2024). TexSense Example Dataset. Zenodo. https://doi.org/10.5281/zenodo.11119845

Höschler, L., Halmich, C., Schranz, C., Fritz, J., Čigoja, S., Ullrich, M., others, & Schwameder, H. (2025). Wearable-based estimation of continuous 3D knee moments during running using a convolutional neural network. *Sports Biomechanics*, 1–19. https://doi.org/10.1080/14763141.2025.2481164

Höschler, L., Halmich, C., Schranz, C., Fritz, J., Koelewijn, A. D., & Schwameder, H. (2024). Towards Real-Time Assessment: Wearable-Based Estimation of 3D Knee Kinetics in Running and the Influence of Preprocessing Workflows. *ISBS Proceedings Archive*, *42*(1), 406–409. https://commons.nmu.edu/isbs/vol42/iss1/72/

Hossain, M. S. B., Zhishan, G., & Hwan, C. (2022). Estimation of Hip, Knee, and Ankle Joint Moment Using a Single IMU Sensor on Foot Via Deep Learning. In 2022 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (pp. 25–33). IEEE.

Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, *65*, 101759. https://doi.org/10.1016/j.media.2020.101759

Khurelbaatar, T., Kim, K., Lee, S., & Kim, Y. H. (2015). Consistent accuracy in whole-body joint kinetics during gait using wearable inertial motion sensors and in-shoe pressure sensors. *Gait & Posture*, *42*(1), 65–69. https://doi.org/10.1016/j.gaitpost.2015.04.007

Lee, C. J., & Lee, J. K. (2022). Inertial Motion Capture-Based Wearable Systems for Estimation of Joint Kinetics: A Systematic Review. Sensors (Basel. *Switzerland)*, *22*(7), 2507. https://doi.org/10.3390/s22072507

Lee, M., & Park, S. (2020). Estimation of Three-Dimensional Lower Limb Kinetics Data during Walking Using Machine Learning from a Single IMU Attached to the Sacrum. *Sensors (Basel, Switzerland)*, *20*(21). https://doi.org/10.3390/s20216277

Liew, B. X. W., Rügamer, D., Zhai, X., Wang, Y., Morris, S., & Netto, K. (2021). Comparing shallow, deep, and transfer learning in predicting joint moments in running. *Journal of Biomechanics*, *129*, 110820. https://doi.org/10.1016/j.jbiomech.2021.110820

Lim, H., Kim, B., & Park, S. (2019). Prediction of Lower Limb Kinetics and Kinematics during Walking by a Single IMU on the Lower Back Using Machine Learning. *Sensors (Basel, Switzerland)*, *20*(1). https://doi.org/10.3390/s20010130

Long, T., Outerleys, J., Yeung, T., Fernandez, J., Bouxsein, M. L., Davis, I. S., others, & Besier, T. F. (2023). Predicting ankle and knee sagittal kinematics and kinetics using an ankle-mounted inertial sensor. *Computer Methods in Biomechanics and Biomedical Engineering*, 1–14. https://doi.org/10.1080/10255842.2023.2224912

Matijevich, E. S., Scott, L. R., Volgyesi, P., Derry, K. H., & Zelik, K. E. (2020). Combining wearable sensor signals, machine learning and biomechanics to estimate tibial bone force and damage during running. *Human Movement Science*, *74*, 102690. https://doi.org/10.1016/j.humov.2020.102690

McCabe, M. V, van Citters, D. W., & Chapman, R. M. (2022). Developing a method for quantifying hip joint angles and moments during walking using neural networks and wearables. *Computer Methods in Biomechanics and Biomedical Engineering*, 1–11. https://doi.org/10.1080/10255842.2022.2044028

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Moghadam, S. M., Yeung, T., & Choisne, J. (2023). The Effect of IMU Sensor Location, Number of Features, and Window Size on a Random Forest Model's Accuracy in Predicting Joint Kinematics and Kinetics During Gait. *IEEE Sensors Journal*, *23*(22), 28328–28339. https://doi.org/10.1109/JSEN.2023.3317366

Mundt, M. (2023). Bridging the lab-to-field gap using machine learning: A narrative review. *Sports Biomechanics*, 1–20. https://doi.org/10.1080/14763141.2023.2200749

Mundt, M., Johnson, W. R., Potthast, W., Markert, B., Mian, A., & Alderson, J. (2021). A Comparison of Three Neural Network Approaches for Estimating Joint Angles and Moments from Inertial Measurement Units. *Sensors (Basel, Switzerland)*, *21*(13). https://doi.org/10.3390/s21134535

Mundt, M., Koeppe, A., David, S., Witter, T., Bamer, F., Potthast, W., & Markert, B. (2020). Estimation of Gait Mechanics Based on Simulated and Measured IMU Data Using an Artificial Neural Network. *Frontiers in Bioengineering and Biotechnology*, *8*, 41. https://doi.org/10.3389/fbioe.2020.00041

Nakagawa, S., & Schielzeth, H. (2012). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Prasanth, H., Caban, M., Keller, U., Courtine, G., Ijspeert, A., Vallery, H., & Zitzewitz, J. (2021). Wearable Sensor-Based Real-Time Gait Detection: A Systematic Review. *Sensors (Basel, Switzerland)*, *21*(8). https://doi.org/10.3390/s21082727

Stefanyshyn, D. J., Stergiou, P., Lun, V. M. Y., Meeuwisse, W. H., & Worobets, J. T. (2006). Knee angular impulse as a predictor of patellofemoral pain in runners. *The American Journal of Sports Medicine*, *34*(11), 1844–1851. https://doi.org/10.1177/0363546506288753

Stetter, B. J., Krafft, F. C., Ringhof, S., Stein, T., & Sell, S. (2020). A Machine Learning and Wearable Sensor Based Approach to Estimate External Knee Flexion and Adduction Moments During Various Locomotion Tasks. *Frontiers in Bioengineering and Biotechnology*, *8*, 9. https://doi.org/10.3389/fbioe.2020.00009

Van den Berghe, P., Six, J., Gerlo, J., Leman, M., & Clercq, D. (n.d.). de (2019). Validity and reliability of peak tibial accelerations as real-time measure of impact loading during over-ground rearfoot running at different speeds. *Journal of Biomechanics*, *86*, 238–242. https://doi.org/10.1016/j.jbiomech.2019.01.039

Van Hooren, B., van Rengs, L., & Meijer, K. (2024). Predicting Musculoskeletal Loading at Common Running Injury Locations using Machine Learning and Instrumented Insoles. *Medicine & Science in Sports & Exercise*, *56*(10), 2059–2075. https://doi.org/10.1249/MSS.0000000000003493

Watanabe, S. (2023). *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance*. https://doi.org/10.48550/arXiv.2304.11127

Weygers, I., Kok, M., Konings, M., Hallez, H., Vroey, H. de, & Claeys, K. (2020). Inertial Sensor-Based Lower Limb Joint Kinematics: A Methodological Systematic Review. *Sensors (Basel, Switzerland)*, *20*(3). https://doi.org/10.3390/s20030673

Whittlesey, S. N., & Robertson, D. G. E. (2004). Two-Dimensional Inverse Dynamics. In D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, & S. N. Whittlesey (Eds.), Research methods in biomechanics (pp. 103–124). Champaign, IL, USA: Human Kinetics.

Wilson, D. R., & Martinez, T. R. (2003). The general inefficiency of batch training for gradient descent learning. *Neural Networks: The Official Journal of the International Neural Network Society*, *16*(10), 1429–1451. https://doi.org/10.1016/S0893-6080(03)00138-2

Yu, B., Gabriel, D., Noble, L., & An, K.-N. (1999). Estimate of the Optimum Cutoff Frequency for the Butterworth Low-Pass Digital Filter. *Journal of Applied Biomechanics*, *15*(3), 318–329. https://doi.org/10.1123/jab.15.3.318

Zhang, J.-T., Novak, A. C., Brouwer, B., & Li, Q. (2013). Concurrent validation of Xsens MVN measurement of lower limb joint angular kinematics. *Physiological Measurement*, *34*(8), N63-9. https://doi.org/10.1088/0967-3334/34/8/N63

Appendix

Table A1. Results of the stepwise model comparison (ANOVA). Lower AIC and BIC indicate a better model fit. Df: degrees of freedom, AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion, logLik: log likelihood, X^2: Chi-squarred.

Model	Df	AIC	BIC	logLik	Test	X^2	p-value
Model_0	3	-2987	-2975	1498			
Model_1.1	4	-3004	-2987	1508	0 vs 1.1	18.0	< 0.0001
Model_1.2	6	-3001	-2975	1506	1.1 vs.1. 2	1.2	0.54
Model_2	7	-3018	-2988	1516	1.2 vs. 2	18.7	< 0.0001
Model_3	10	-3012	-2969	1516	2 vs. 3	0.3	0.97

Table A2. Model architecture at baseline and after hyperparameter optimization of the 1D convolutional layers. Additionally, the median and mean of the optimized configurations are shown. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P).* IMU kernel size is multiplied by 4 due to higher sampling rate, n.a.: not applicable, layer does not exist.

Layer	**Parameter**	***Baseline***	**FSTP**	**FSP**	**FTP**	**FS**	**FT**	**F**	**S**	**Median**	**Mean**
IMU	In Channels	*24*	24	18	18	12	12	6	6	**12**	**14**
	Out Channels	*32*	64	64	128	128	32	128	128	**128**	**96**
	Kernel Size*	*51*	51	15	51	51	27	51	15	**51**	**37**
Merge	In Channels	*32*	64	n.a.	128	n.a.	32	n.a.	128	**96**	**88**
	Out Channels	*32*	64	n.a.	128	n.a.	32	n.a.	128	**96**	**88**
	Kernel Size	*27*	27	n.a.	7	n.a.	15	n.a.	15	**15**	**16**
Aux 1	In Channels	*34*	66	66	130	130	34	130	130	**130**	**98**
	Out Channels	*64*	256	64	32	32	32	256	32	**32**	**101**
Aux 2	In Channels	*64*	256	64	32	32	32	256	32	**32**	**101**
	Out Channels	*64*	64	64	128	128	32	128	128	**128**	**96**
Conv 1	In Channels	*64*	64	64	128	128	32	128	128	**128**	**96**
	Out Channels	*64*	128	64	64	256	128	128	256	**128**	**146**
	Kernel Size	*15*	7	27	7	15	51	7	51	**15**	**24**
Extra 1	In Channels	*n.a.*	128	64	64	256	128	n.a.	256	**128**	**149**

Table A2. Continued.

Layer	Parameter	*Baseline*	FSTP	FSP	FTP	FS	FT	F	S	Median	Mean
Extra 1	Out Channels	*n.a.*	64	256	256	128	64	n.a.	64	**96**	**139**
	Kernel Size	*n.a.*	51	51	27	27	27	n.a.	27	**27**	**35**
Extra 2	In Channels	*n.a.*	64	n.a.	256	128	n.a.	n.a.	64	**96**	**128**
	Out Channels	*n.a.*	128	n.a.	128	256	n.a.	n.a.	128	**128**	**160**
	Kernel Size	*n.a.*	15	n.a.	27	51	n.a.	n.a.	15	**21**	**27**
Output	In Channels	*64*	128	256	128	256	64	128	128	**128**	**155**
	Out Channels	*3*	3	3	3	3	3	3	3	**3**	**3**
	Kernel Size	*7*	15	15	15	27	27	51	7	**15**	**22**

Table A3. Training parameters of the baseline model and the optimized configurations. Additionally, the median and mean of the optimized configurations are shown. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P), lr_{min}: minimum learning rate, lr_{max}: maximum learning rate.

Parameter	*Baseline*	FSTP	FSP	FTP	FS	FT	F	S	Median	Mean
Epochs	*20*	35	25	25	25	25	35	50	**25**	**31**
Batch Size	*64*	8	8	8	8	8	16	16	**8**	**10**
lr_{min}	1×10^{-3}	2.27×10^{-4}	1.56×10^{-4}	1.60×10^{-4}	4.42×10^{-4}	2.31×10^{-4}	7.96×10^{-4}	3.78×10^{-4}	$\mathbf{2.31\times10^{-4}}$	$\mathbf{3.41\times10^{-4}}$
lr_{min}	4×10^{-3}	7.26×10^{-4}	7.89×10^{-4}	6.49×10^{-4}	4.84×10^{-4}	1.14×10^{-3}	2.01×10^{-3}	1.78×10^{-3}	$\mathbf{7.89\times10^{-4}}$	$\mathbf{1.08\times10^{-3}}$

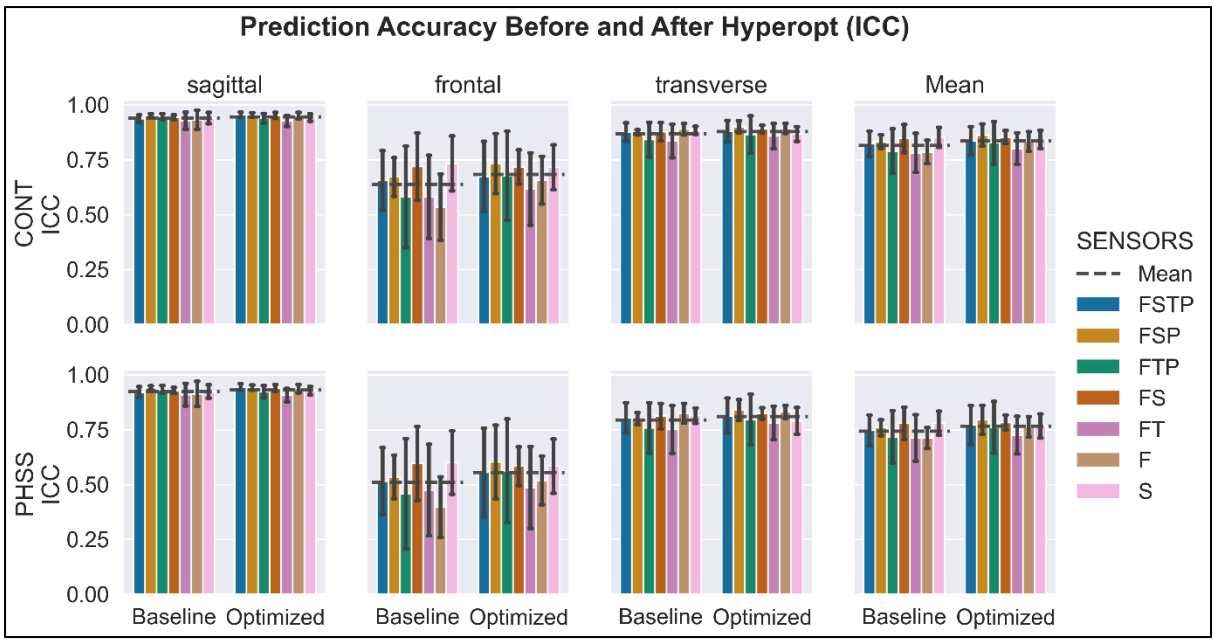


Figure A1. Test set performance (intra-class correlation, ICC) of the baseline and optimized models. Displayed are the results in the three planes of motion and the mean across them. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P). Bar height indicates the mean, caps show the standard deviation. The mean across all configurations is displayed as dashed line. CONT – predictions over the entire samples of 10 seconds, PHSS – predictions during stance phases only.

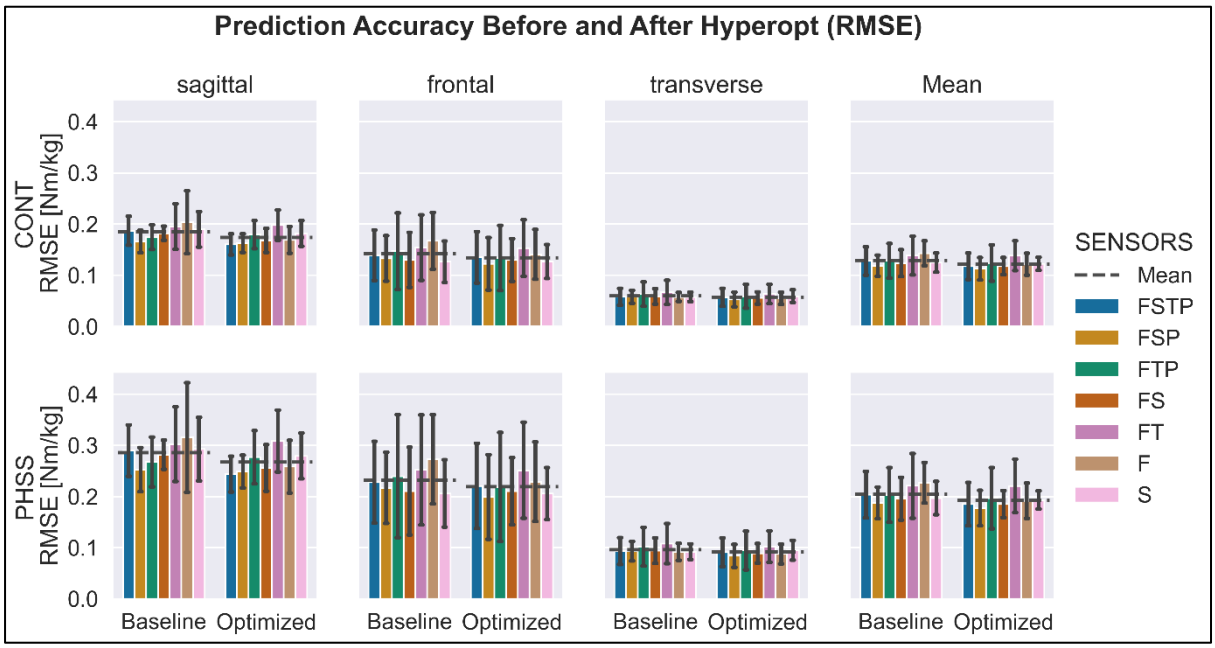


Figure A2. Test set performance (RMSE) of the baseline and optimized models. Displayed are the results in the three planes of motion and the mean across them. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P). Bar height indicates the mean, caps show the standard deviation. The mean across all configurations is displayed as dashed line. CONT – predictions over the entire samples of 10 seconds, PHSS – predictions during stance phases only.

Table A4. Test set performance (normalized root mean squared error – nRMSE) of the baseline and optimized models with the respective mean across configurations. Displayed are the results in the three planes of motion and the mean across them. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P). CONT – predictions over the entire samples of 10 seconds, PHSS – predictions during stance phases only.

nRMSE		**CONT**				**PHSS**			
	Sensors	sagittal	frontal	transverse	*Mean*	sagittal	frontal	transverse	*Mean*
Baseline	FSTP	0.076	0.153	0.096	*0.108*	0.117	0.252	0.154	*0.174*
	FSP	0.067	0.143	0.096	*0.102*	0.102	0.233	0.155	*0.163*
	FTP	0.071	0.158	0.104	*0.111*	0.108	0.258	0.168	*0.178*
	FS	0.074	0.137	0.098	*0.103*	0.114	0.223	0.157	*0.165*
	FT	0.080	0.173	0.110	*0.121*	0.123	0.285	0.178	*0.196*
	F	0.085	0.180	0.098	*0.121*	0.131	0.295	0.156	*0.194*
	S	0.078	0.137	0.097	*0.104*	0.121	0.224	0.155	*0.167*
	Mean	**0.076**	**0.155**	**0.100**	***0.110***	**0.117**	**0.253**	**0.161**	***0.177***
Optimized	FSTP	0.066	0.151	0.094	*0.104*	0.099	0.247	0.151	*0.166*
	FSP	0.067	0.131	0.087	*0.095*	0.102	0.214	0.139	*0.151*
	FTP	0.073	0.149	0.098	*0.107*	0.112	0.245	0.157	*0.171*
	FS	0.069	0.139	0.092	*0.100*	0.104	0.226	0.147	*0.159*
	FT	0.081	0.172	0.105	*0.119*	0.125	0.284	0.169	*0.193*
	F	0.069	0.150	0.092	*0.104*	0.104	0.244	0.146	*0.165*
	S	0.075	0.138	0.099	*0.104*	0.114	0.225	0.159	*0.166*
	Mean	**0.071**	**0.147**	**0.095**	***0.105***	**0.109**	**0.241**	**0.153**	***0.167***

Table A5. Test set performance (intraclass correleation – ICC) of the baseline and optimized models with the respective mean across configurations. Displayed are the results in the three planes of motion and the mean across them. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P). CONT – predictions over the entire samples of 10 seconds, PHSS – predictions during stance phases only.

ICC		**CONT**				**PHSS**			
	Sensors	sagittal	frontal	transverse	*Mean*	sagittal	frontal	transverse	*Mean*
Baseline	FSTP	0.937	0.655	0.876	*0.823*	0.921	0.516	0.804	*0.747*
	FSP	0.952	0.672	0.875	*0.833*	0.942	0.535	0.801	*0.759*
	FTP	0.947	0.581	0.842	*0.790*	0.933	0.459	0.758	*0.717*
	FS	0.944	0.719	0.878	*0.847*	0.929	0.596	0.812	*0.779*
	FT	0.928	0.581	0.836	*0.782*	0.910	0.475	0.751	*0.712*
	F	0.932	0.535	0.890	*0.786*	0.913	0.397	0.825	*0.712*
	S	0.940	0.733	0.884	*0.852*	0.924	0.600	0.814	*0.779*
	Mean	**0.940**	**0.639**	**0.869**	***0.816***	**0.924**	**0.511**	**0.795**	***0.744***
Optimized	FSTP	0.955	0.674	0.880	*0.836*	0.945	0.555	0.814	*0.771*
	FSP	0.954	0.733	0.900	*0.862*	0.942	0.603	0.840	*0.795*
	FTP	0.939	0.677	0.866	*0.827*	0.923	0.563	0.797	*0.761*
	FS	0.952	0.718	0.891	*0.853*	0.940	0.584	0.824	*0.783*
	FT	0.928	0.617	0.859	*0.801*	0.907	0.487	0.781	*0.725*
	F	0.950	0.657	0.893	*0.833*	0.937	0.519	0.832	*0.763*
	S	0.943	0.716	0.868	*0.842*	0.928	0.584	0.790	*0.767*
	Mean	**0.946**	**0.684**	**0.880**	***0.837***	**0.932**	**0.556**	**0.811**	***0.766***

Table A6. Test set performance (root mean squared error – RMSE) of the baseline and optimized models with the respective mean across configurations. Displayed are the results in the three planes of motion and the mean across them. The abbreviation of the optimized configurations corresponds to the combination of sensors used (foot: F, shank: S, thigh: T, pelvis: P). CONT – predictions over the entire samples of 10 seconds, PHSS – predictions during stance phases only.

RMSE		**CONT**				**PHSS**			
[Nm/kg]	Sensors	sagittal	frontal	transverse	*Mean*	sagittal	frontal	transverse	*Mean*
Baseline	FSTP	0.187	0.139	0.058	*0.128*	0.290	0.228	0.093	*0.204*
	FSP	0.166	0.133	0.058	*0.119*	0.253	0.217	0.093	*0.188*
	FTP	0.174	0.147	0.063	*0.128*	0.268	0.240	0.102	*0.203*
	FS	0.182	0.130	0.059	*0.124*	0.282	0.211	0.095	*0.196*
	FT	0.195	0.154	0.067	*0.139*	0.303	0.252	0.108	*0.221*
	F	0.204	0.167	0.058	*0.143*	0.315	0.273	0.092	*0.227*
	S	0.190	0.127	0.058	*0.125*	0.293	0.206	0.092	*0.197*
	Mean	**0.185**	**0.142**	**0.060**	***0.129***	**0.286**	**0.232**	**0.097**	***0.205***
Optimized	FSTP	0.161	0.135	0.057	*0.118*	0.244	0.221	0.091	*0.185*
	FSP	0.163	0.122	0.053	*0.113*	0.249	0.199	0.084	*0.177*
	FTP	0.179	0.134	0.059	*0.124*	0.277	0.219	0.095	*0.197*
	FS	0.168	0.130	0.056	*0.118*	0.256	0.211	0.089	*0.185*
	FT	0.198	0.153	0.064	*0.138*	0.308	0.251	0.102	*0.221*
	F	0.169	0.141	0.055	*0.122*	0.258	0.229	0.088	*0.192*
	S	0.182	0.127	0.059	*0.123*	0.279	0.206	0.095	*0.193*
	Mean	**0.174**	**0.135**	**0.058**	***0.122***	**0.267**	**0.219**	**0.092**	***0.193***