Medical Engineering &
Physics

IPEM
Institute of Physics and Engineering in Medicine

TOPICAL REVIEW • OPEN ACCESS

Application of machine learning techniques in the measurement of ground reaction force (GRF) waveforms using inertial measurement unit (IMU) sensors—a systematic literature review

To cite this article: Malik Naveed Akhter *et al* 2026 *Med. Eng. Phys.* **147** 10TR01

View the article online for updates and enhancements.

You may also like

- MEMS two-axis force plate array used to measure the ground reaction forces during the running motion of an ant
Hidetoshi Takahashi, Nguyen Thanh-Vinh, Uijin G Jung et al.
- ON THE CLASSIFICATION OF GRBs AND THEIR OCCURRENCE RATES
R. Ruffini, J. A. Rueda, M. Muccino et al.
- Concurrent prediction of ground reaction forces and moments and tibiofemoral contact forces during walking using musculoskeletal modelling
Yinghu Peng, Zhifeng Zhang, Yongchang Gao et al.

This content was downloaded from IP address 156.62.3.22 on 03/11/2026 at 02:26

Medical Engineering & Physics

IPEM Institute of Physics and Engineering in Medicine

Check for updates

TOPICAL REVIEW

Application of machine learning techniques in the measurement of ground reaction force (GRF) waveforms using inertial measurement unit (IMU) sensors—a systematic literature review

Malik Naveed Akhter[1,*], **Haider Ali**[2], **Imran Khan Niazi**[3], **Umair Javaid Manj**[1] and **Denise Taylor**[1,4]

[1] Health & Environmental Sciences, Auckland University of Technology, Auckland, New Zealand
[2] Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand
[3] Center of Chiropractic Research, New Zealand College of Chiropractic, Auckland, New Zealand
[4] New Zealand Dizziness & Balance Center, Auckland, New Zealand
[*] Author to whom any correspondence should be addressed.

E-mail: pjt6828@autuni.ac.nz

Keywords: biomechanics, ground reaction force, machine learning, AI, deep learning

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
15 June 2026

REVISED
26 August 2026

ACCEPTED FOR PUBLICATION
30 September 2026

PUBLISHED
13 October 2026

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

Three-dimensional ground reaction forces (GRFs) are important parameters in human motion analysis, yet their measurement is confined to laboratory settings. GRF estimation from wearable inertial measurement unit (IMU) data using machine learning and deep learning has emerged as a promising avenue for kinetic analysis outside the laboratory, but no consensus on standardized methods or best practices has emerged. This review synthesizes existing methods and provides a framework of current practices and recommendations for future work. The review followed PRISMA guidelines (PROSPERO CRD42024578576) and searched PubMed, Scopus, IEEE Xplore, and ScienceDirect from 2011 to 2026. Studies were included if they used IMUs and machine learning or deep learning to estimate GRFs and reported quantitative performance metrics. Data were extracted on participant characteristics, IMU specifications and placement, preprocessing, models, validation protocols, and error metrics. Twenty-six studies met the inclusion criteria, with sample sizes ranging from 5 to 47 (median = 15) and IMU counts ranging from 1–16. Most studies used triaxial accelerometers and gyroscopes sampled at 100–1000 Hz. Vertical GRF was generally estimated more accurately than the anteroposterior and mediolateral components within the same study. A single cross-study accuracy ranking was not defensible because studies used different normalization denominators, evaluation windows, tasks, sensor configurations, and validation strategies. Heterogeneity in validation protocols and demographic samples limited direct cross-study comparisons. This review synthesizes how implementation choices, including model architecture, feature-engineering strategy, sensor configuration, and validation protocol, shape GRF prediction across axes. The variability observed underscores the need for standardized methodologies, larger and more diverse cohorts, consistent performance metrics, and benchmarking on public datasets to enable fair cross-study comparisons and accelerate clinical translation.

1. Introduction

During stepping, the human body exerts force upon the ground. According to Newton's third law, the ground reacts with an equal and opposite force, known as the ground reaction force (GRF). The GRF vector can be decomposed into three anatomical components: Vertical, anteroposterior (AP), and mediolateral (ML). In normal human gait, the vertical GRF exhibits a characteristic M-shaped double-peaked profile. The first peak appears when the leading foot accepts weight, termed as the loading response, and a second peak occurs during push-off as the trailing foot propels the body forward (Giardini *et al* 2023). The AP GRF alternates between a negative phase (the deceleration phase), which occurs when

© 2026 The Author(s). Published on behalf of Institute of Physics and Engineering in Medicine by IOP Publishing Ltd.

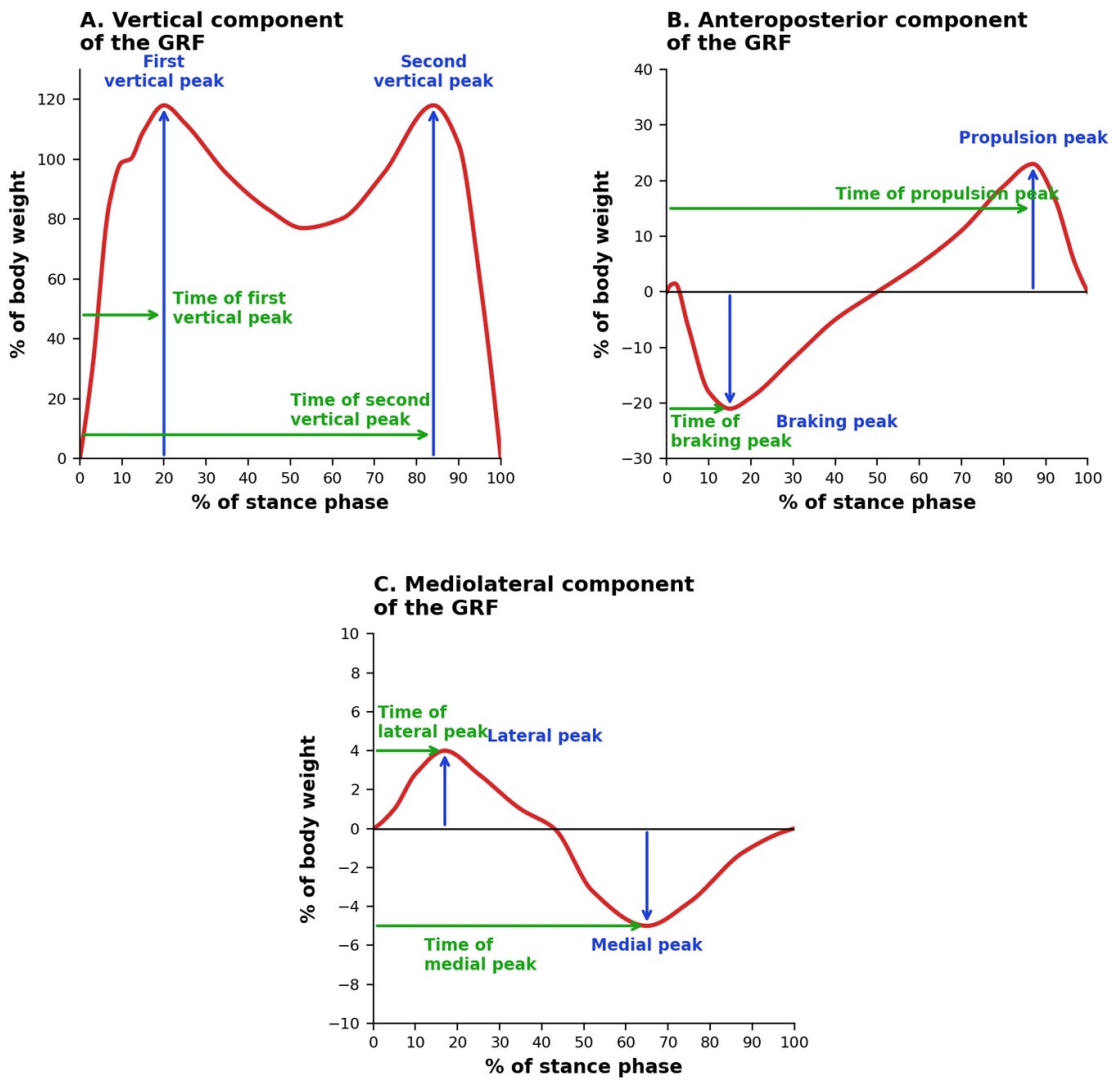


Figure 1. Schematic of the three anatomical GRF components and their characteristic vertical, anteroposterior, and mediolateral waveforms during a typical walking stance. Curves are illustrative and are not empirical data.

the foot initially contacts the ground, and a positive phase (the propulsion phase), which occurs when the foot pushes off (Barela *et al* 2014). The ML GRF is smaller in magnitude and fluctuates as weight shifts from side to side. It is helpful for analyzing lateral balance and foot placement dynamics (McClay and Cavanagh 1994). In healthy gait, these GRF components are well-synchronized and symmetric between limbs, reflecting efficient support of bodyweight, controlled deceleration/acceleration of the center of mass (CoM), and balanced side-to-side stability. Abnormalities in any GRF component can reveal underlying impairments in muscle function, joint stability, or motor control (Van Der Velde *et al* 2025). Figure 1 illustrates the three anatomical GRF components and their characteristic walking-stance waveforms.

From a clinical perspective, GRF metrics provide instrument-derived, quantitative information about gait abnormalities that may not be visible during routine observation. They help identify issues such as inadequate weight acceptance, reduced push-off power, or limb asymmetries (Williams *et al* 2011, Chockalingam *et al* 2018). For instance, in some cases of cerebral palsy (CP), a diminished or absent second vertical GRF peak, known as the 'Ben Lomond' profile, reflects weak push-off and poor limb power (Williams *et al* 2011), often accompanied by a 70% reduction in peak AP GRF (Salehi *et al* 2018). GRF-guided planning is widely used in single-event multilevel surgery in CP with evidence demonstrating long-term improvements in walking efficiency and energy cost (Rutz *et al* 2013). In stroke hemiparesis, the paretic limb typically exhibits reduced AP forces and a lower weight-acceptance peak, leading to asymmetrical stance patterns and increased reliance on the non-paretic limb (Chen *et al* 2005). Knee osteoarthritis patients exhibit reduced dynamic vertical, AP, and ML GRF (Costello *et al* 2019), with a lower first vertical peak and a reduced second peak (Childs *et al* 2004), and a laterally directed ML GRF

to unload the medial knee (Shafizadegan *et al* 2016). Parkinson's disease patients exhibit diminished second peak in vertical GRF (Sharifmoradi *et al* 2015) and prolonged double support (Zanardi *et al* 2021). Multiple sclerosis patients show a significant reduction in AP GRF amplitudes and greater ML GRF as compared to healthy controls (Kelleher *et al* 2009, Jonsdottir *et al* 2020). Age-related gait decline is characterized by a reduced first peak and a notable reduction in the second peak of the GRF during walking (Boyer *et al* 2017). Across these conditions, GRF provides unique insights into the mechanisms underlying pathological gait. Despite the clear diagnostic potential of GRF metrics across these diverse gait conditions, their integration into routine clinical practice remains limited. High equipment costs, restricted access to gait laboratories, and limited evidence directly linking GRF-guided interventions to improved patient outcomes pose significant barriers. Consequently, many clinicians continue to rely on simpler, more accessible functional assessments (simpler here meaning faster setup, less specialist preparation, lower cost, and fewer spatial constraints) for routine evaluation, even though GRF analysis could provide richer, quantitative insights into gait pathology.

To effectively utilize GRF metrics in clinical settings, accurate and accessible measurement methods are essential. The gold-standard device for GRF measurement is the force plate (Pagnacco *et al* 2000), which directly measures the three-dimensional components of the GRF vector at high sampling rates and provides precise magnitude, timing, and center-of-pressure data for each foot strike. In practice, multiple force plates are embedded in the floor of a gait lab to record consecutive steps across a walkway. Although highly accurate, force plates are expensive and restricted to lab environments; their cost reflects multi-axis force transducers, structural installation, amplifiers and acquisition hardware, calibration and maintenance, synchronization with motion-capture systems, and specialist operation (Kloeckner *et al* 2023). Patients must also strike the plates cleanly (one foot per plate), which can alter natural gait (Kloeckner *et al* 2023), compromising data validity. An alternative to force plates is the instrumented treadmill (Sinitski *et al* 2015), which is equipped with force and pressure sensors that can accurately and directly measure the GRF during cyclic activities such as walking or running. However, instrumented treadmills are also expensive and typically limited to research facilities. Moreover, gait on a moving belt differs from overground gait (Ochoa *et al* 2017), which also impacts the validity of collected data. To overcome this, a variety of techniques utilizing wearable sensors have been developed to measure the magnitude and direction of the GRF vector. These include insole pressure sensors (Davis *et al* 1998, Razak *et al* 2012, Huang *et al* 2025), and instrumented shoes (Faivre *et al* 2004). However, the presence of sensors mounted inside or under the shoes can affect natural foot biomechanics by altering shoe fit, plantar contact pressure, sole stiffness, effective foot mass, and plantar sensory feedback, and thereby loading distribution and push-off (Liu *et al* 2010). Overall, GRF measurement using existing methods remains expensive, equipment-dependent, and largely inaccessible for regular clinical or real-world applications.

To overcome the shortcomings of traditional GRF measurement methods, researchers over the last decade have increasingly explored the use of wearable inertial measurement units (IMUs) to estimate GRF waveforms. IMUs are lightweight, affordable sensors that can record human movement outside of specialized laboratories (Liang *et al* 2023). IMUs directly measure linear acceleration and angular velocity of the body segment and can be used to derive kinematic descriptors such as segment orientations, velocities, and positions (Kim *et al* 2025). However, IMUs cannot directly measure kinetic descriptors such as forces, torques, or pressure (Krishnakumar *et al* 2024). Data-driven modeling can be used to correlate raw IMU kinematic data to kinetic parameters. These methods employ machine learning and deep learning models trained on paired datasets, where the input data consist of IMU-derived kinematic signals and the output data correspond to reference GRF measurements obtained from force plates.

Machine learning models learn the mapping between input signals and outputs by optimizing model parameters to minimize prediction error during training. After training, the best-performing model is used for inference, predicting outputs for unseen inputs, and its ability to do so effectively defines its generalization (Wang *et al* 2022). However, raw IMU signals often contain drift, noise, and sensor artifacts, making preprocessing a critical step in traditional machine learning pipelines. Following this, preprocessed signals are transformed into engineered features, such as peak accelerations, temporal dynamics, or frequency-domain characteristics, that are more strongly correlated with GRF and better suited for model learning (Jiang *et al* 2020). In contrast, deep learning models employ multilayer neural architectures that automatically learn hierarchical representations directly from raw or minimally processed IMU data. Using convolutional, recurrent, or fully connected layers, deep learning models reduce reliance on handcrafted features and can capture complex, nonlinear relationships between body motion and GRF (Johnson *et al* 2021). The overall workflow for GRF estimation using both machine learning and deep learning approaches is illustrated in figure 2, highlighting the key differences between traditional feature-engineered machine learning pipelines and modern end-to-end learning deep learning frameworks. Deep

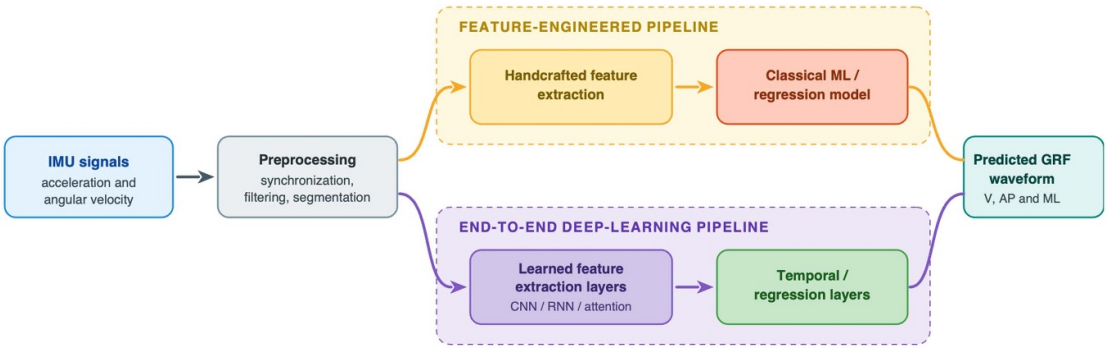


Figure 2. Machine learning and deep learning model training pipeline. Feature extraction is explicit in the classical pipeline and is performed implicitly as representation learning inside the deep-learning pipeline.

learning reduces explicit manual feature engineering, but it does not remove preprocessing or feature extraction from the pipeline; in deep models the feature-extraction stage is performed implicitly as representation learning inside the network.

1.1. Main contributions

Despite significant progress in machine learning and deep learning techniques for estimating GRFs from IMUs, there is currently no methodological consistency across studies, leading to a wide variety of sensor setups, algorithms, and validation protocols in the literature. There is little agreement on which input features (raw accelerations, integrated velocities, frequency-domain features) are most informative, or which model architecture is optimal. Machine learning models also vary, ranging from simpler linear regression models with hand-crafted features to more complex deep neural networks that automatically learn features (Halilaj *et al* 2018). The evaluation metrics and reference standards differ, with some comparing against force plate data for walking, while others compare against instrumented treadmills or instrumented prostheses for various activities. A consequence of this heterogeneity is that no standardized methodology or best practice has emerged yet for IMU-based kinetic estimation.

Methods based on in-shoe pressure insoles, instrumented shoes, and inverse kinematics modeling have been discussed in the existing literature (Razak *et al* 2012, Ancillao *et al* 2018), and earlier reviews have covered wearable inertial estimation of GRF and gait kinetics more broadly (Ancillao *et al* 2018, Liang *et al* 2023, Krishnakumar *et al* 2024). The distinct contribution of the present review is a pipeline-level synthesis of continuous GRF waveform estimation, compared across model families, all three GRF components, sensor configurations, validation levels, and exact error-metric definitions. This review aims to synthesize the methodologies employed across existing studies and provide a coherent framework of current practices and recommendations for future work. The review is structured around five key research questions, following the machine learning and deep learning training pipeline presented in figure 2.

1) What participant populations, experimental protocols, and sensor configurations were used to train and test the neural networks?
2) What preprocessing and feature extraction techniques were applied to the IMU data to improve GRF predictions?
3) What machine learning and deep learning models were implemented? How do traditional machine learning methods compare with advanced deep neural networks in predicting vertical, AP, and ML GRF components?
4) What evaluation metrics were used to assess model performance?
5) What challenges exist in deploying these machine or deep learning solutions for GRF estimation in real-world biomechanical and clinical applications, and what future directions emerge to address these gaps?

2. Review methodology

This systematic review was designed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher *et al* 2015). The protocol was registered with PROSPERO (registration number: CRD42024578576).

2.1. Search strategy and selection criteria

Electronic databases, including PubMed, Scopus, IEEE Xplore, and ScienceDirect, were searched from 2011 to April 2025. The search was subsequently updated to June 2026, and four additional eligible reports identified through the update were included. The search strategy employed a combination of keywords to identify articles that addressed GRF, ground reaction moment, and IMUs. Machine learning and deep learning, or AI, were not included in the search terms because some studies use specific method names in their titles. Instead, the machine learning and deep learning studies were screened manually. The systematic search strategy and yields by database are provided in supplementary table S1.

The titles, abstracts, and full texts of the retrieved documents were evaluated for relevance. Articles were selected based on the participants, intervention, comparisons, and outcomes principle (Schardt *et al* 2007). Only papers published in English were considered. Studies that have not undergone peer review, including conference abstracts, editorials, and other forms of gray literature lacking rigorous scientific evaluation, were not included. The collation of articles, removal of duplicates, and screening process were carried out using Covidence (Kellermeyer *et al* 2018). Two authors (M.N.A. and H.A.) independently conducted the screening process. Disagreements between the reviewers regarding article inclusion were resolved through a third reviewer (IKN). Restricting eligibility to English-language reports is a potential source of language bias. The screening did not code language as a separate exclusion reason, so the number of records excluded solely for language cannot be reconstructed.

2.2. Eligibility criteria

Studies were included if:

- The study focused on GRF estimation
- An IMU was used as a data collection device.
- Data processing and machine or deep learning model development methods were clearly defined
- Measurements obtained from gold-standard reference systems (force plates, instrumented treadmills) were used as the benchmark for evaluating model performance.
- The study reported quantitative performance outcomes.

Studies were excluded if:

- The study focused on estimating parameters other than GRF.
- The studies used modeling approaches other than machine learning and deep learning, such as inverse dynamics or musculoskeletal modeling.
- Non-human studies.

2.3. Quality assessment

The quality of the included studies was assessed using a checklist based on QualSyst Assessment Tool for quantitative studies (Kmet *et al* 2004). This tool was chosen because it offers a standardized, transparent scoring system that can be applied consistently across heterogeneous study designs. The checklist consisted of eight key aspects. Each item was scored as 0 (‘no’), 0.5 (‘partial’), or 1 (‘yes’), yielding a maximum possible score of 8. A threshold of six was selected to ensure that at least 75% of the predefined quality domains were satisfied. This quality check was performed alongside data extraction to ensure that the review incorporated only the most relevant and rigorously reported findings. Evaluation of each study against eight quality-assessment questions is provided in supplementary table S2.

2.4. Data extraction and synthesis

Data from the studies were extracted based on participant demographics, sensor specifications, and machine-learning methodologies. Participant demographics included the total number of participants, gender distribution, age range, population group (healthy adults, clinical populations), and task performed. Sensor specifications covered the number of sensors, their anatomical placement on the body,

the manufacturer, and the sampling frequency. Finally, for each machine-learning and deep-learning approach, we cataloged the specific algorithms employed, the proportions of data allocated to training and testing, cross-validation strategies, data-preprocessing techniques, reference measurement method, evaluation metrics, and reported performance. Metrics are quoted at the precision used by the source. Where a value was converted or derived by the review authors rather than quoted directly, the conversion and its source inputs are stated. Following peer review, the extracted performance values were re-checked against the primary sources, and the page, table, figure, or equation supplying each value was recorded. The re-check was independently confirmed by a second reviewer.

3. Results

The initial electronic database search yielded 628 articles, and 19 additional records were identified through other sources: 15 from reference lists and four eligible reports published after the original search, identified when the search was updated to June 2026 (Yılmazgün *et al* 2025, Gempfer *et al* 2026, Stuchbury-Wass *et al* 2026, Ullah *et al* 2026). After removing 99 duplicates, 548 records underwent title and abstract screening, followed by full-text evaluation and quality assessment, resulting in 26 studies included in the systematic review. The PRISMA Diagram is presented in figure 3.

To ensure direct alignment with the review aims, the results are structured around the five research questions listed above. Specifically, sections 4.1–4.4 report evidence addressing research questions 1–4 (participant protocols, preprocessing/features, models, and evaluation/validation). Question 5 (real-world deployment challenges and future directions) is primarily synthesized in the Discussion section.

3.1. RQ1 participant populations and experimental protocols

This section summarizes participant demographics, tasks/protocols (treadmill, overground, outdoor, dynamic tasks), the reference measurement systems, and the resulting implications for ecological validity across the included studies.

3.1.1. Participant demographics

The majority of participants were healthy adults, except for one study that included individuals with medial knee osteoarthritis (MKOA) (Havashinezhadian *et al* 2023). The sample sizes ranged from 5–47, with a median of 15. Eight studies reported sample sizes less than 10. Johnson *et al* (2021) reported a total of 458 372 motion trials instead of the number of participants. Most cohorts consisted of young, healthy adults with mean ages typically between 21 and 30 years (mean 26 years, median 25 years). Gender distribution was generally mixed, though several studies included only male participants, resulting in a modest male bias across the literature. The number of participants in each study is presented in supplementary table S3.

3.1.2. Types of activities and movements studied

Laboratory-based studies predominantly employed an instrumented treadmill to vary gait speed, incline, and task complexity under controlled conditions. Dorschky *et al* (2020) had participants perform both walking and running trials at variable speeds (0.9–2.0 m s^{-1} and 3.1–4.9 m s^{-1}, respectively) on a treadmill to explore speed-dependent features of the GRF. Jiang *et al* (2020) and Lee and Park (2020) focused exclusively on multi-speed walking protocols (0.4–1.6 m s^{-1} and 0.7–1.6 m s^{-1}, respectively). Scheltinga *et al* (2023) focused on running at fixed speeds while Wouda *et al* (2018) exploring multispeed running. Alcantara *et al* (2022) introduced variable slopes ($\pm 5°$ and $\pm 10°$) while running to assess the effects of task difficulty on lower-limb kinetics. Beyond ambulation, several studies incorporated dynamic or plyometric tasks. Cerfoglio *et al* (2021) examined vertical drop jumps, Sun *et al* (2023) studied single and double-leg drop landings, and Hossain *et al* (2023) included stair-climbing.

Overground studies captured ecologically valid gait behaviors by allowing participants to walk or run freely in real-world settings. Guo *et al* (2017) compared controlled back-and-forth walking along a 50 m walkway with unrestricted urban ambulation to examine how environmental complexity alters IMU-derived GRF estimates. Leporace *et al* (2015) and Wu *et al* (2020) studied level-ground walking. Outdoor running studies by Donahue and Hahn (2023a) assessed gait on level, inclined, and declined surfaces, and their follow-up work Donahue and Hahn (2023b) introduced outdoor running on a 400 m track at different speeds.

A detailed description of the tasks performed by each study is given in table 2.

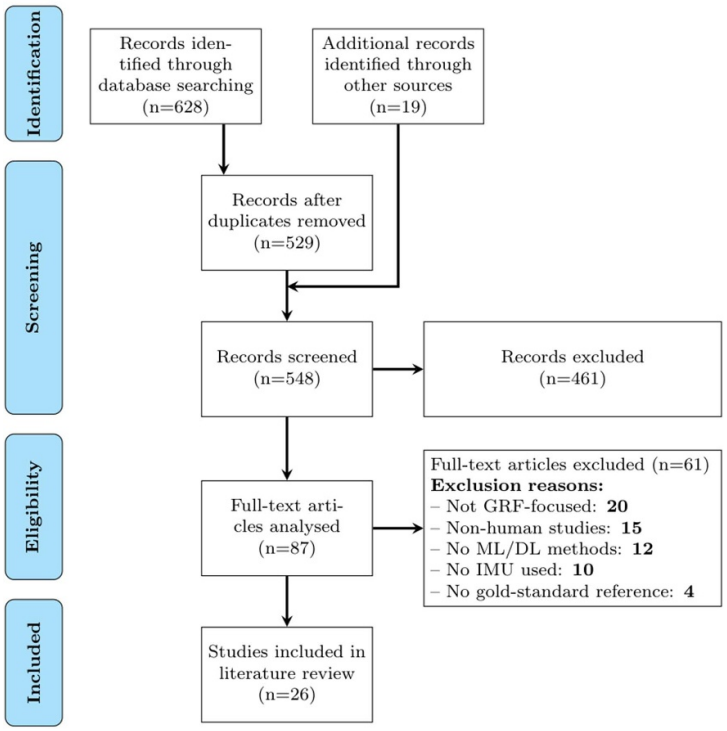


Figure 3. PRISMA Flow diagram.

3.1.3. Reference measurement

Reference systems varied across studies and included force plates, instrumented treadmills, and in-shoe pressure-sensing insoles. Indoor studies mostly used high-precision force plates or instrumented treadmills. Outdoor running studies relied on pressure-sensing insoles. Guo *et al* (2017) used insoles (F-Scan 3000E by Tekscan™) sampled at 128 Hz to capture plantar force in step with 128 Hz IMUs. Donahue and Hahn (2023a), Donahue and Hahn (2023b) used Loadsol insole force sensors (Novel Electronics, St. Paul, MN, USA), sampled at 100 Hz, and downsampled the IMU data to match the sampling rate of insoles. Pressure insoles only record vertical GRF, so studies focusing on AP and ML GRF preferred treadmills or multi-axis force plates.

A detailed description of the reference measurement method used in each study is provided in supplementary table S4.

3.1.4. IMU sensor characteristics and placement

The number of IMU sensors ranged from one to sixteen. Common placements included the pelvis, thigh, shank, tibia, and foot. Table 1 lists the anatomical distribution of IMU sensor placements across the reviewed studies.

Most studies used triaxial IMUs, except Alcantara *et al* (2022) used biaxial IMUs. Sampling rates varied from 100 Hz to 1000 Hz. Eleven studies used only Accelerometer data to train the models, while the remaining studies used both the Accelerometer and Gyroscope data. Although several of the devices used contained magnetometers, no study reported using magnetometer channels as model inputs. In indoor environments and in the proximity of ferromagnetic materials, the magnetic field is highly disturbed and inhomogeneous, which can result in inaccurate orientation estimates from IMUs (Roetenberg *et al* 2005). Notably, only five of the 26 reports used a single physical IMU as their primary configuration (six when the marker-derived input of Kim *et al* is counted), despite single-sensor configurations arguably being more clinically feasible because they reduce setup time, placement complexity, and patient/clinician burden compared with multi-sensor setups. Details of sensor characteristics are provided in supplementary table S4.

Figure 4 shows the study-by-location matrix together with the pairwise co-occurrence of placements, making explicit which anatomical sites are combined in multi-IMU setups. No included study used magnetometer channels as predictive model inputs. This can be attributed to risk of ferromagnetic disturbance in gait laboratories rather than evidence that magnetometers are unhelpful, and it should not be

Table 1. Anatomical distribution of IMU sensor placements across reviewed studies. Count (n) indicates the number of studies that placed at least one IMU at the location (multiple/bilateral sensors counted once per study).

Studies	Anatomical location	*n*	% of studies
Leporace *et al* (2015), Leporace *et al* (2018), Shahabpoor *et al* (2018), Wouda *et al* (2018), Dorschky *et al* (2020), Jiang *et al* (2020), Cerfoglio *et al* (2021), Johnson *et al* (2021), Bach *et al* (2022), Havashinezhadian *et al* (2023), Hossain *et al* (2023), Scheltinga *et al* (2023), Sun *et al* (2023), Yılmazgün *et al* (2025), Gempfer *et al* (2026), Ullah *et al* (2026)	Shank/tibia	16	61.5
Ngoh *et al* (2018), Shahabpoor *et al* (2018), Dorschky *et al* (2020), Jiang *et al* (2020), Alcantara *et al* (2022), Havashinezhadian *et al* (2023), Hossain *et al* (2023), Scheltinga *et al* (2023), Sun *et al* (2023), Donahue and Hahn (2023a, 2023b), Yılmazgün *et al* (2025), Gempfer *et al* (2026), Ullah *et al* (2026)	Foot/shoe	14	53.8
Shahabpoor *et al* (2018), Dorschky *et al* (2020), Jiang *et al* (2020), Cerfoglio *et al* (2021), Johnson *et al* (2021), Hossain *et al* (2023), Scheltinga *et al* (2023), Sun *et al* (2023), Yılmazgün *et al* (2025)	Thigh	9	34.6
Lim *et al* (2019), Lee and Park (2020), Cerfoglio *et al* (2021), Alcantara *et al* (2022), Donahue and Hahn (2023a, 2023b), Gempfer *et al* (2026), Ullah *et al* (2026)	Sacrum	8	30.8
Wouda *et al* (2018), Johnson *et al* (2021), Scheltinga *et al* (2023), Sun *et al* (2023), Yılmazgün *et al* (2025)	Pelvis/waist (ASIS region)	5	19.2
Guo *et al* (2017), Shahabpoor *et al* (2018), Dorschky *et al* (2020)	Lower back (L5/unspecified lower back)	3	11.5
Guo *et al* (2017), Shahabpoor *et al* (2018)	Upper trunk (C7)	2	7.7
Scheltinga *et al* (2023), Sun *et al* (2023), Gempfer *et al* (2026)	Trunk/chest (thoracic)	3	11.5
Guo *et al* (2017), Stuchbury-Wass *et al* (2026)	Ear	2	7.7
Shahabpoor *et al* (2018)	Upper arm	1	3.8
Shahabpoor *et al* (2018)	Forearm	1	3.8

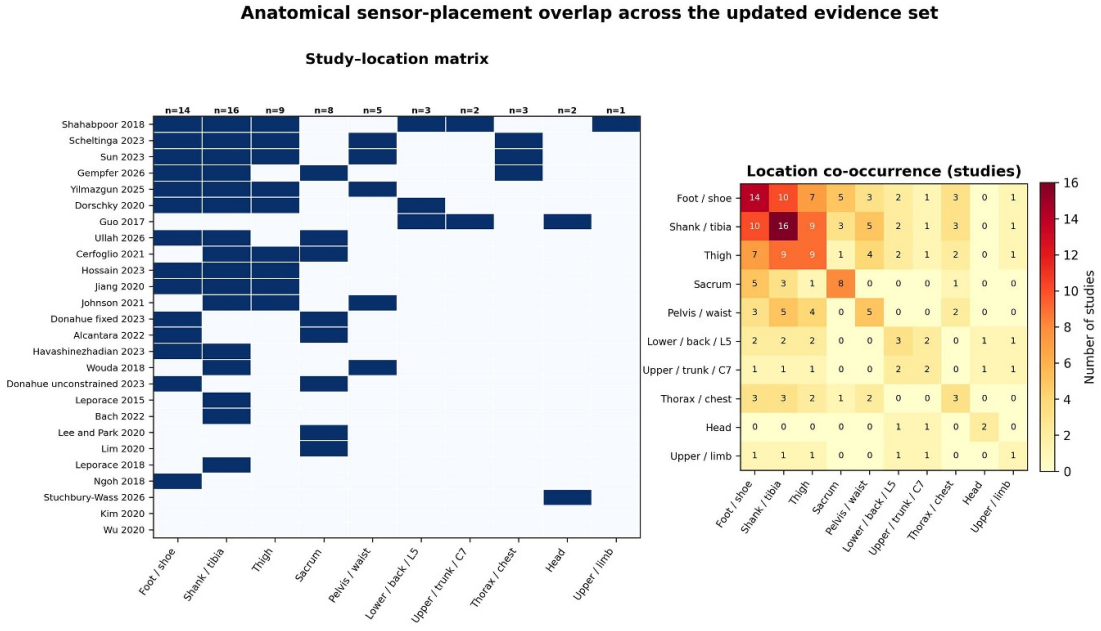


Figure 4. Study-by-location matrix and pairwise co-occurrence of anatomical inertial sensor placements across the included reports. The co-occurrence panel shows how often location pairs appear together in multi-IMU setups. Frequency of use does not establish that a location is more accurate.

read as a recommendation to exclude them permanently. *Direct six-Axis versus nine-Axis comparisons, with hard- and soft-iron calibration, field-magnitude checks, and adaptive* anomaly rejection, are needed in both magnetically clean and disturbed environments before their contribution can be judged.

3.2. RQ2—preprocessing and feature extraction

The findings below synthesize preprocessing techniques, including synchronization, filtering, normalization, segmentation, and feature engineering choices reported across studies.

3.2.1. Data synchronization

Data Synchronization is the process of aligning IMU data with ground truth measurements to ensure temporal correspondence. Most studies used foot stomps or jumps before each trial and then employed cross-correlation between IMU signals and GRF data to identify temporal offsets. Some studies reported using external hardware to start data acquisition simultaneously (Ngoh *et al* 2018, Jiang *et al* 2020, Sun *et al* 2023). Donahue and Hahn (2023b) developed an iterative correction algorithm to resolve asynchronization caused by variations in the device clocks of the IMU and the ground truth device. Lim *et al* (2019) applied drift removal in velocity and displacement calculations.

3.2.2. Filtering

Filtering removes noise and unwanted frequency components from signals, enhancing data quality. A Butterworth low-pass filter was used in most of the reviewed studies, with cutoff frequencies ranging from 10–35 Hz. Wu *et al* (2020) used discrete wavelet transform, which filters out the noise caused by motion artifacts that arise due to improper fixation of the IMU sensor (Avrutov *et al* 2022).

3.2.3. Normalization

Since GRF amplitudes depend on bodyweight, most studies normalized GRF outputs to bodyweight to help machine learning models generalize better to new data by focusing on relative changes rather than absolute values. Beyond normalizing the GRF targets to bodyweight, several studies additionally scaled the predicted GRFs by each participant's body mass at the output stage to recover absolute force values (Dorschky *et al* 2020, Jiang *et al* 2020, Lee and Park 2020, Hossain *et al* 2023).

3.2.4. Segmentation

Some studies isolated relevant portions of the data, such as individual gait cycles or stance phases, to exclude non-informative segments, thereby preventing the machine learning model from learning noise (Wouda *et al* 2018). Sun *et al* (2023) built an IMU-based event detection module to automatically identify the start of a drop landing in real-time. They applied a threshold to the IMU acceleration magnitude to segment the region of interest within the continuous signal.

3.2.5. Feature engineering

Features transform raw sensor data into meaningful features, enhancing the ability of machine learning models to learn patterns and relationships within the data. Most studies that used deep learning models did not extract features, instead, they used raw IMU data as input. However, studies that used classical machine learning models performed explicit feature engineering on the IMU signals. Jiang *et al* (2020) extracted 10 temporal domain statistical features, including root mean square (RMS), variance, and mean absolute deviation (MAD) from 110 ms windows. Two studies (Leporace *et al* 2015, 2018) extracted a total of 14 features, including kinematic features such as velocity and displacement by integrating acceleration data and temporal information from the accelerometer signals. However, they observed non-normal residual distributions (high kurtosis/Skewness), indicating that these 14 features, while useful, might not capture every nuance of the GRF waveform. Guo *et al* (2017) used iterative orthogonal forward regression (iOFR) to automatically select features from raw acceleration components. iOFR iteratively picks the most significant features from a large pool, ensuring that only the most informative inputs are used for modeling. Scheltinga *et al* (2023) used a physics model based on Newton's second law as inputs to a hybrid model. Kim *et al* (2020) incorporated forces derived from a spring-loaded inverted pendulum (SLIP) model as input to the model. SLIP features led to an improved Normalized root mean square error (NRMSE) of 5.28% compared with 6.48% for vertical GRF estimation. Bach *et al* (2022) applied principal component analysis (PCA) to triaxial acceleration data. PCA reduces the dimensionality of the data by transforming it into a set of orthogonal components that capture the most significant variance while filtering out noise (Jolliffe and Cadima 2016). Johnson *et al* (2021) mapped accelerometer data onto RGB color space images, representing sensor locations and time frames, and used these images as input to pretrained CNNs CaffeNet and ResNet-50.

3.3. RQ3—machine and deep learning models and performance by GRF component

This section summarizes model families and reports performance patterns across V/AP/ML components and tasks.

A comprehensive review of the machine and deep learning models used and reported accuracies of the 26 included studies is presented in table 2, and a detailed description is provided in the subsections

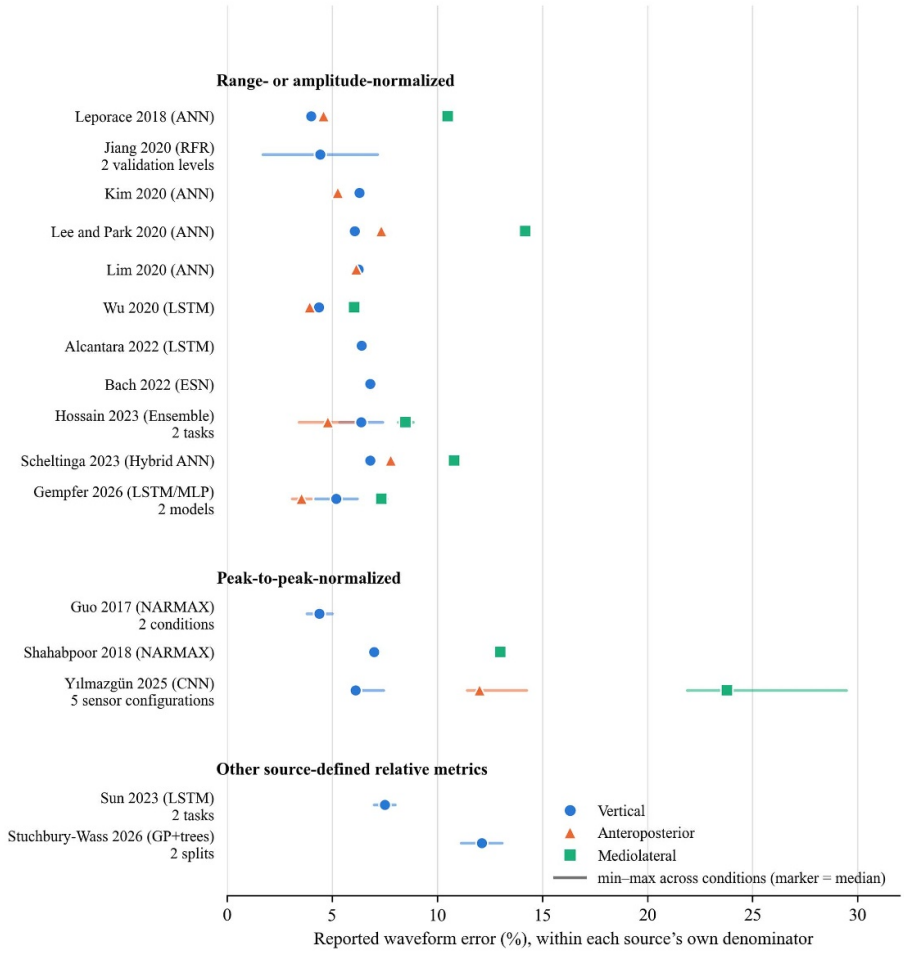


Figure 5. Reported percentage waveform errors for the vertical, anteroposterior, and mediolateral GRF components, by study, grouped by normalization family. Markers are medians; horizontal segments span the minimum–maximum across a study's reported conditions (sensor configurations, tasks, speeds, models, or validation levels). Values are comparable only within a study, since denominators differ between sources; families are never pooled. Reports whose primary error is in force units or %BW (Dorschky *et al* (2020), Ullah *et al* (2026), Donahue and Hahn (2023a, 2023b), Cerfoglio *et al* (2021), Havashinezhadian *et al* (2023), Johnson *et al* (2021), Ngoh *et al* (2018), Wouda *et al* (2018), Leporace *et al* (2015)) appear in table 2 and supplementary table S7.

below. Notably, table 2 shows that early work emphasized interpretable, low-parameter methods, including NARMAX, feedforward neural networks (FFNN), multilayer perceptrons (MLPs), and random forest. More recent studies favor sequence-modeling approaches, such as long short-term memory (LSTM), echo state networks (ESN), convolutional neural networks (CNN), and hybrid or ensemble frameworks. Implementation details of each study, including layer dimensions and hyperparameter settings, are provided in supplementary table S6. Figure 5 shows the distribution of percentage-based waveform errors across the included reports, separated by GRF component and by normalization family so that values computed against different denominators are never pooled. The distribution is nuanced rather than a clean separation between classical and deep models: within the range-normalized subset vertical errors cluster lower than ML errors while AP values overlap both, the spread within a single model family is often wider than the difference between families, and validation level materially affects where a study falls. No single model family emerges as uniformly superior from these data. Table 3 summarizes how the model families compare in inputs, components estimated, validation pattern, and current limitation.

3.3.1. System identification models

Nonlinear auto regressive models with exogenous input (NARX) and their generalization to include moving average error terms (NARMAX) provide a powerful, data-driven framework for modeling time-series data. Guo *et al* (2017) utilized the NARMAX framework to estimate vertical GRF during walking. The authors first decomposed raw acceleration data into left-dominant and right-dominant components

Table 2. Participant characteristics, number of IMUs used, Tasks performed, machine learning and deep learning Model used, and reported accuracies per study. 'N/A' indicates that GRF estimation in that direction was not performed. This table directly addresses RQ3 by pairing each study's sensor setup, task, model class, and reported accuracy for ML/AP/V GRF.

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Guo *et al* (2017)	2017	9 healthy (6 M, 3 F) age: 28 ± 3 years	3	Outdoor controlled walking: participants walked back and forth along a 50-meter walkway Outdoor free walking: participants walked freely in a city center along public footpaths, without restrictions on the route or walking speed	NARMAX	N/A	N/A	**Outdoor Controlled Walking (OCW):** rRMSE: 3.8% $r = 0.993$ **Outdoor free walking (OFW):** rRMSE: 5.0% $r = 0.990$
Shahabpoor *et al* (2018)	2018	6 healthy males Age: 21 ± 1 years Height: 182 ± 8 cm Weight 77 ± 16 kg	12	Treadmill walking	NARMAX	NRMSE: 14%	NRMSE: 12%	NRMSE: 7%
Cerfoglio *et al* (2021)	2021	11 healthy (6 M, 5 F) Age: 24.5 ± 1.7 years Height: 170.5 ± 10.3 cm Weight: 65 ± 11.3 kg	3	Vertical drop jumps	NARX	RMSE: 0.0081 N kg^{-1}	RMSE: 0.0350 N kg^{-1}	RMSE: 0.0184 N kg^{-1}
Jiang *et al* (2020)	2020	9 healthy males Age: 27 ± 8.0 years Height: 176 ± 7 cm Weight: 72 ± 9 kg	4	Treadmill walking for 1 min at five different speeds: 0.4 m s^{-1}, 0.7 m s^{-1}, 1.0 m s^{-1}, 1.3 m s^{-1}, 1.6 m s^{-1}	RFR	N/A	N/A	**Intra-participant testing:** *r*: 1.00 RMSE: 0.02 BW NRMSE: 1.7% **Inter-participant testing:** *r*: 0.97 RMSE: 0.10 BW NRMSE: 7.15%

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Lee and Park (2020)	2020	17 analyzed participants (20 recruited: 12 F, 8 M; 3 excluded due to equipment failure) Age: 24.7 ± 3.2 years Height: 166.7 ± 7.8 cm Weight: 60.1 ± 8.9 kg	1	Treadmill walking at four different speeds (0.7, 1.0, 1.3, and 1.6 m s^{-1}) for 2 min at each speed.	ANN	**0.7 m s^{-1} walking speed:** NRMSE: 17.75 ± 6.41%R^2: 0.54 ± 0.40 **1.0 m s^{-1} walking speed:** NRMSE: 15.23 ± 5.50% R^2: 0.66 ± 0.25 **1.3 m s^{-1} walking speed:** NRMSE: 14.19 ± 3.91% R^2: 0.70 ± 0.16 **1.6 m s^{-1} walking speed:** NRMSE: 15.04 ± 2.78% R^2: 0.63 ± 0.14	**0.7 m s^{-1} walking speed:** NRMSE: 13.52 ± 4.38% R^2: 0.77 ± 0.15 **1.0 m s^{-1} walking speed:** NRMSE: 8.31 ± 2.89% R^2: 0.91 ± 0.07 **1.3 m s^{-1} walking speed:** NRMSE: 7.34 ± 2.46% R^2: 0.92 ± 0.05 **1.6 m s^{-1} walking speed:** NRMSE: 7.67 ± 2.73% R^2: 0.91 ± 0.06.	**0.7 m s^{-1} walking speed:** NRMSE: 6.86 ± 2.21% R^2: 0.94 ± 0.04 **1.0 m s^{-1} walking speed:** NRMSE: 6.17 ± 1.81% R^2: 0.94 ± 0.03 **1.3 m s^{-1} walking speed:** NRMSE: 6.08 ± 1.88% R^2: 0.94 ± 0.04 **1.6 m s^{-1} walking speed:** NRMSE: 7.68 ± 3.15% R^2: 0.89 ± 0.09.
Wouda *et al* (2018)	2018	8 healthy males Age: 25.1 ± 5.2 years Height: 183.7 ± 4.5 cm Weight: 77.7 ± 9.4 kg	3	Treadmill running at three different speeds—10, 12, and 14 km h^{-1} for 3 min at each speed.	ANN	N/A	N/A	**Single-subject training:** RMSE < 0.27 BW, $r > 0.99$ **Multi-subject training:** $r > 0.9$ for most subjects; accuracy decreased relative to the single-subject condition

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Lim *et al* (2019)	2020	7 healthy males Age: 25.0 ± 2.9 years. Height: 168.8 ± 7.5 cm. Weight: 72.1 ± 7.7 kg	1	Treadmill walking 1-min trials at three customized gait speeds for each subject: Slow speed: approximately 20% slower than the participant's preferred speed. Moderate speed: the participant's preferred speed. Fast speed: a midpoint between the preferred and maximum walking speeds.	ANN	N/A	**Moderate-speed walking:** NRMSE: 6.16% ± 1.76 R^2: 0.98 ± 0.01	**Moderate-speed walking:** NRMSE: 6.26% ± 1.24 R^2: 0.96 ± 0.03
Ngoh *et al* (2018)	2018	7 healthy males Age: 21.3 ± 0.5 years Height: 174.9 ± 6.6 cm Weight: 63 ± 6.1 kg	1	Treadmill walking and running	ANN (FFN)	N/A	N/A	**At 8 km h^{-1}:** RMSE: 0.017 ± 0.010 BW *r*: 0.99 **At 9 km h^{-1}:** RMSE: 0.015 ± 0.0072 BW *r*: 0.99 **At 10 km h^{-1}:** RMSE: 0.017 ± 0.0091 BW *r*: 0.99

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Leporace *et al* (2015)	2015	17 healthy (11 M, 6 F) Age: 27.1 ± 3.4 years Weight: 84.3 ± 4.5 kg	1	Overground walking—8 m	ANN (MLP)	Single network with three outputs: MAD: 1.4 ± 0.5%BW *r*: 0.801 Three separate networks with one output each: MAD: 1.35 ± 0.46%BW *r*: 0.792	**Single network with three outputs:** MAD: 1.9 ± 0.6% bodyweight (BW) Correlation (r): 0.969 **Three separate networks with one output each:** MAD: 1.83 ± 0.68%BW *r*: 0.971	Single Network with Three Outputs: MAD: 5.2 ± 1.7%BW *r:* 0.968 Three Separate Networks with One Output Each: MAD: 5.1 ± 1.64%BW *r*: 0.969
Leporace *et al* (2018)	2018	17 healthy (11 M, 6 F) Age: 27.1 ± 3.4 years Weight: 84.3 ± 4.5 kg	1	8 m walk overground	ANN (MLP)	MAD: 1.4% BW Normalized MAD: 10.5% *r*: 0.80	MAD: 1.8% BW Normalized MAD: 4.6% *r*: 0.97	MAD: 4.5% BW NMAD: 4.0% *r*: 0.98

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Kim *et al* (2020)	2020	7 healthy males Age: 25.0 ± 2.9 years Height: 168.8 ± 7.5 cm Weight: 72.1 ± 7.7 kg	1	Treadmill walking at three speeds (1.0–2.3 m s^{-1})	ANN	N/A	**Method 1 (CoM kinematics):** NRMSE: 6.48 ± 2.93% *r*: 0.98 ± 0.01 **Method 2 (SLIP Transform):** NRMSE: 5.28 ± 0.99% *r*: 0.98 ± 0.01	**Method 1 (CoM kinematics):** NRMSE: 6.93 ± 4.02% *r*: 0.98 ± 0.01 **Method 2 (SLIP Transform):** NRMSE: 6.30 ± 1.31% *r*: 0.98 ± 0.01
Scheltinga *et al* (2023)	2023	12 healthy (8 M, 4 F) Age: 31.6 ± 9.0 years Height: 178 ± 11 cm Weight: 73.7 ± 17.5 kg	8	Treadmill running	ANN and hybrid ANN	**Direct** RMSE = 0.05 ± 0.01 rRMSE = 10.9 ± 2.5% $r = 0.57 \pm 0.21$ **Hybrid** RMSE = 0.05 ± 0.01 rRMSE = 10.8 ± 2.5% $r = 0.58 \pm 0.22$	**Direct** RMSE =0.08 ± 0.03 rRMSE = 7.8 ± 3.1% $r = 0.91 \pm 0.09$ **Hybrid** RMSE = 0.07 ± 0.03 rRMSE = 7.8 ± 3.2% $r = 0.91 \pm 0.10$	**Direct** RMSE = 0.19 ± 0.04 rRMSE = 7.3 ± 1.8% $r = 0.97 \pm 0.02$ **Hybrid** RMSE =0.18 ± 0.04 rRMSE = 6.8 ± 1.7% $r = 0.97 \pm 0.01$

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Havashinezhadian *et al* (2023)	2023	Healthy: 27 Age: 36.66 ± 4.34 years Height: 172 ± 8 cm Weight: 73.11 ± 16.45 kg **MKOA**:18 Age: 65.63 ± 7.32 years Height: 165 ± 12 cm Weight: 80.51 ± 15.27 k	5	Treadmill walking	ESN	**Healthy** MAE: 0.016 **MKOA** MAE: 0.014	**Healthy** MAE:0.032 **MKOA** MAE: 0.025	**Healthy** MAE: 0.123 **MKOA** MAE: 0.107
Bach *et al* (2022)	2022	21 healthy (13 M, 8 F) Age: 20.8 years ± 1.0 year Height: 181.7 cm ± 10.3 cm Weight: 71.1 kg ± 9.8 kg.	2	Treadmill walking and runningwalking speed: 1.24 ± 0.12 m s^{-1}running speed: 2.20 ± 0.14 m s^{-1}	ESN	N/A	N/A	NRMSE: 6.8% ± 0.3% R^2: 0.96 ± 0.00

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Donahue and Hahn (2023a)	2023	13 healthy (8 M, 5 F) Age: 23.2 years Height: 167.8 cm Weight: 65.0 kg	3	Outdoor running: level ground, incline and decline	LSTM	N/A	N/A	**Level ground running** **2.25 m s^{-1}:** RMSE: 0.315 ± 0.102 BW **3.00 m s^{-1}:** RMSE: 0.294 ± 0.065 BW **4.00 m s^{-1}:** RMSE: 0.339 ± 0.085 BW **4.50 m s^{-1}:** RMSE 0.397 ± 0.180 BW **5.00 m s^{-1}:** RMSE: 0.427 ± 0.010 BW **Decline running** **3.25 m s^{-1}:** RMSE: 0.277 ± 0.065 BW **4.00 m s^{-1}:** RMSE: 0.309 ± 0.082 BW **4.50 m s^{-1}:** RMSE: 0.576 ± 0.371 BW **5.00 m s^{-1}:** RMSE: 0.635 ± 0.000 BW **Incline running** **2.25 m s^{-1}:** RMSE: 0.230 ± 0.081 BW **3.25 m s^{-1}:** RMSE: 0.277 ± 0.065 BW **4.00 m s^{-1}:** RMSE: 0.309 ± 0.082 BW **4.50 m s^{-1}:** RMSE: 0.269 ± 0.007 BW
Donahue and Hahn (2023b)	2023	15 healthy (9 M, 6 F) Age: 23.6 years Height: 178.3 cm Weight: 73.5 kg	3	Outdoor running—400 m. 5 trials with increasing velocity (2.33–5.36 m s^{-1})	BD-LSTM	N/A	N/A	**Stance phase RMSE:** 0.189-0.288 BW **Whole waveform RMSE:** 0.191-0.309 BW

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Wu *et al* (2020)	2020	5 healthy (3 M, 2 F) Age: 24 ± 2 years Height: 168 ± 2 cm Weight: 64 ± 11 kg	4	Overground Walking	LSTM	NRMSE 6.05 ± 1.11%	NRMSE 3.93 ± 0.58%	NRMSE 4.37 ± 0.89%
Alcantara *et al* (2022)	2022	19 healthy (10 M, 9 F) Age: 29 ± 9 years Height: 173 ± 9 cm Weight: 68.1 ± 9.9 kg	3	Treadmill Runningfive different slopes: 0°, +5°, −5°, +10°, and −10°. three speeds: 2.5 m s^{-1}, 3.33 m s^{-1}, and 4.17 m s^{-1} for each slope.	LSTM	N/A	N/A	RMSE: 0.16 ± 0.04 BW rRMSE: 6.4 ± 1.5%
Sun *et al* (2023)	2023	16 healthy males Age: 23.2 ± 1.4 years. Height: 177 ± 4 cm Weight: 72.8 ± 9.7 kg	8	Single-leg and double-leg drop landing tasks from a height of 30 cm	LSTM	N/A	N/A	**Single-leg drop landings:** RMSE: 0.25 ± 0.13 BW rRMSE: 0.07 ± 0.04 R^2 = 0.88 ± 0.12 **Double-leg drop landings:** RMSE: 0.18 ± 0.06 rRMSE: 0.08 ± 0.03

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Dorschky *et al* (2020)	2020	10 healthy males Age: 27 ± 2.6 years Height: 182 ± 5 cm Weight: 76.9 ± 8.6 kg	7 recorded; 4 used	Overground walking and running at three speeds each	CNN	N/A	**Only measured data** RMSE: 4.29-4.52% BW $r = 0.970$-0.976 **With simulated data** RMSE:3.76-5.10% BW $r = 0.966$-0.980	**Only measured data** RMSE: 11.74-13.24% BW $r = 0.979$-0.988 **With simulated data** RMSE:9.99-16.24% BW $r = 0.974$-0.991
Johnson *et al* (2021)	2021	458 372 motion trials from healthy subjects 59.9% M, 40.1% F Height: 177 ± 10.1 cm Weight: 74.9 ± 34.1 kg	5	Overground running and 90-degree sidestep	CNN	**CaffeNet:** r= 0.87 ± 0.009 **ResNet-50:** r= 0.87 ± 0.001	**CaffeNet:** r= 0.90 ± 0.005 **ResNet-50:** r= 0.90 ± 0.008	**CaffeNet** r= 0.89 ± 0.009 **ResNet-50** r= 0.85 ± 0.013
Hossain *et al* (2023)	2023	20 healthy (12 M, 8 F) Age: 21.7 ± 3.65 years Height: 1.70 ± 0.07 m Weight: 68.21 ± 11.52 kg	3	Treadmill walking, level-ground walking, ramp walking, and Stair Climb	Ensemble-Net	**Treadmill walking** NRMSE: 8.14 ± 3.55 *r*: 0.913 ± 0.045 **Level ground walking:** NRMSE: 8.85 ± 1.77 *r*: 0.833 ± 0.101 **Ramp walking** NRMSE: 8.32 ± 1.74 r: 0.847 ± 0.066 **Stair walking** NRMSE: 8.16 ± 1.56 *r*: 0.885 ± 0.042	**Treadmill walking:** NRMSE: 3.42 ± 1.05 *r*: 0.962 ± 0.031 **Level ground walking:** NRMSE: 6.18 ± 3.87 *r*: 0.883 ± 0.177 **Ramp walking** NRMSE: 3.99 ± 1.03 *r*: 0.951 ± 0.040 **Stair walking** NRMSE: 4.64 ± 0.87 *r*: 0.926 ± 0.030	**Treadmill walking:** NRMSE: 5.36 ± 3.20 *r*: 0.976 ± 0.037 **Level ground walking:** NRMSE: 7.40 ± 6.45 *r*: 0.957 ± 0.087 **Ramp walking** NRMSE: 4.60 ± 2.40 *r*: 0.978 ± 0.043 **Stair walking** NRMSE: 4.51 ± 2.24 *r*: 0.983 ± 0.024

(Continued.)

Table 2. (Continued.)

Study	Year	Participants characteristics	IMUs used	Task performed	Model used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
Yılmazgün *et al* (2025)	2025	20 healthy (10 M/10 F); age 24.61 ± 3.67 years; height 172 ± 9 cm; mass 68.26 ± 11.45 kg	16 recorded; configurations of 7, 4, 2, and 1 used	Walking, stairs, running, step turn and spin turn	CNN	Overall relRMSE 21.89–29.46%; r 0.74–0.81	Overall relRMSE 11.43–14.24%; r 0.92–0.95	Overall relRMSE 6.02–7.44%; r 0.98
Gempfer *et al* (2026)	2026	32 healthy (20 M/12 F); age 30.4 ± 8.6 years; height 176 ± 9 cm; mass 72.8 ± 12.4 kg	10 accelerometers recorded; 6 used per leg-side model	Barefoot self-selected overground laboratory walking	LSTM and MLP	nRMSE 7.6 ± 2.5% (LSTM)/7.1 ± 2.0% (MLP)	nRMSE 4.0 ± 1.1% (LSTM)/3.1 ± 0.8% (MLP)	nRMSE 6.2 ± 1.6% (LSTM)/4.2 ± 1.2% (MLP)
Ullah *et al* (2026)	2026	47 (29 healthy; 18 osteoarthritis, predominantly knee); 25 F/22 M; age 40.98 ± 16.04 y	8 recorded; bilateral feet, shanks and sacrum (5) used	Ten treadmill walking conditions	Transformer with low-rank multi-stream fusion	N/A	N/A	Mixed-augmentation test RMSE: H2K 0.38 ± 0.10 N kg^{-1} vs 0.42 ± 0.11 baseline; K2H 0.47 ± 0.17 vs 0.53 ± 0.17 baseline
Stuchbury-Wass *et al* (2026)	2026	30 healthy runners (18 M/12 F); age 28.2 ± 9.2 years; height 176.2 ± 8.8 cm; mass 73.8 ± 13.2 kg	2	Treadmill and indoor overground running	Gaussian-process scalars with pointwise boosted trees	N/A	N/A	Treadmill unseen participants: RMSE 0.154 BW; source-defined NRMSE 11.14%

Table 3. Comparison of model families used for IMU-based GRF estimation. The table is a qualitative taxonomy: performance is not compared across families because the underlying studies differ in dataset, task, validation split, and metric denominator.

Model family	Typical inputs	Components; validation pattern	Reported advantage (+) and current limitation (−)
System identification/linear (Guo *et al* 2017, Shahabpoor *et al* 2018)	Handcrafted, decomposed accelerations	V; within-participant holdout	+ interpretable; low data requirement − assumes near-periodic steady-state gait
Random forest (Jiang *et al* 2020)	Summary and temporal features	V, AP, ML; participant-level splits	+ robust to outliers; little tuning − limited temporal modeling
Feed-forward ANN/MLP (Leporace *et al* 2015, Leporace *et al* 2018, Lim *et al* 2019, Gempfer *et al* 2026)	Engineered features; segment accelerations	V, AP, ML; LOSO and holdout	+ simple and fast; competitive on small data − no temporal memory; often needs per-subject calibration
Recurrent/reservoir (Wu *et al* 2020, Bach *et al* 2022, Donahue and Hahn 2023a,Donahue and Hahn 2023b)	Raw or lightly processed sequences	V, AP, ML; mixed, some sample-level splits	+ captures temporal dependence − Error can compound at speed; split rigor varies
Convolutional (Dorschky *et al* 2020, Alcantara *et al* 2022, Yılmazgün *et al* 2025)	Raw multi-channel IMU windows	V, AP, ML; LOSO and fixed held-out participants	+ strong waveform accuracy from raw data − Data-hungry; sensitive to placement
Transformer (Ullah *et al* 2026)	Multi-stream raw IMU	V; participant-disjoint cross-cohort	+ cross-cohort transfer with augmentation − single component; high data and compute cost
Hybrid/physics-informed (Kim *et al* 2020, Scheltinga *et al* 2023)	IMU plus physical model terms	V, AP, ML; participant-level	+ reduces data need; interpretable structure − gains component-specific, not general
Ensemble (Hossain *et al* 2023, Stuchbury-Wass *et al* 2026)	Feature sets across sensors	V, AP, ML; participant-disjoint	+ variance reduction − added complexity; limited external validation

based on gait-event-derived membership functions: $w_{left}(t)$, $w_{right}(t)$ which represent the relative contribution of each foot's GRF at any given time as:

$$a_{left}(t) = a(t).w_{left}(t), \quad a_{right}(t) = a(t).w_{right}(t).$$

These decomposed acceleration values were then fed to a NARMAX model, which is defined as:

$$v\mathrm{GRF}(k) = F(\mathrm{asensor}(k), \mathrm{asensor}(k-1), \ldots, \mathrm{asensor}(k-L))$$

where F is a multivariate polynomial function representing the system's nonlinear dynamics, and L is the maximum time lag considered. iOFR was used to identify a sparse and efficient set of features. The model was validated on data from nine young healthy adults performing controlled and free outdoor walking trials. The reported vertical force Root mean square error (RMSE) was 3.8% during controlled outdoor walking and 5.0% during free outdoor walking, with correlations of 0.993 and 0.990, respectively.

Shahabpoor *et al* (2018) proposed two linear models based on the correlations between segmental accelerations and GRFs. Model 1 utilized PCC to weigh the contributions of each IMU, whereas Model 2 selected the IMU with the highest correlation for each GRF direction. Model 2 outperformed Model 1, achieving NRMSE values of 14%, 12%, and 7% for ML, AP, and V GRFs, respectively. The authors further introduced second-order polynomial terms and lagged inputs into the NARMAX framework, which yielded marginal improvements of 2%–3% in NRMSE across all directions.

Cerfoglio *et al* (2021) used a feedforward NARX network to estimate GRFs during the vertical drop jump task from IMUs. The model reported mean RMSE values of ≈0.0081-0.0350 N kg^{-1}, indicating high precision in its predictions. However, the authors acknowledged certain limitations in accuracy, particularly boundary errors observed in the initial frames of the landing phase, which may stem from the model's inability to capture rapid changes at the onset of movement.

3.3.2. Random forest

The random forest algorithm is an ensemble learning method that constructs multiple decision trees by training each tree on a randomly sampled subset of the input data, using bootstrapping, and on a random subset of features at each split. For prediction, the algorithm aggregates outputs from all trees to produce a final estimate. This approach is robust to data outliers and enhances generalizability. Jiang *et al* (2020) employed a random forest ensemble-based model for estimating vertical GRF using data from a single shank-mounted IMU placed on various lower limb locations. Ten temporal domain features, including RMS, variance, and wavelet coefficients, were extracted. The model yielded high accuracy, with PCC values of 1.00 and 0.97 and low RMSE values of 0.02 BW and 0.10 BW in intra- and inter-participant testing, respectively. The study did not compare the performance against other machine learning models for their dataset.

3.3.3. Artificial neural networks (ANNs)

ANNs are computational models inspired by the structure and function of the human brain. They consist of interconnected nodes which are organized into layers: an input layer, one or more hidden layers, and an output layer. Each node receives input, processes it using an activation function, and passes the result to the next layer. The connections between neurons have associated weights that are adjusted during training to improve the network's accuracy. Lee and Park (2020) used ANN to estimate 3D lower limb kinetics during walking on a split-belt treadmill at different speeds via a single IMU attached to the sacrum. Inputs to the ANN included time and 3D components of sacral position and velocity (obtained by integrating acceleration data), as well as raw acceleration data. LOOCV was applied to train and validate the model. The study demonstrated promising results with NRMSE ranging from 6.7% to 15.6%.

Lim *et al* (2019) used ANN to predict lower limb kinetics during walking using a single IMU placed on the lower back near the sacrum. Input to the model consisted of 7 input nodes, including time, acceleration, velocity (integral of acceleration), and position (double integral of acceleration) in x and y directions. The model achieved an average NRMSE of 7%.

Wouda *et al* (2018) used two concatenated ANNs to estimate vertical GRF during running using three IMUs placed on the pelvis and lower legs. The first ANN mapped the relative orientations of the lower legs to lower-body joint angles from IMU data. The output of the first ANN served as input for the second network, which estimated the vertical GRF. The model's performance was evaluated under two scenarios: single-subject training where models were trained on 10 and 14 km h^{-1} and tested on 12 km h^{-1} within the same participant, and multiple-subject training, where models were trained on multiple participants and evaluated on a different participant. For single-subject training, the paper reported RMSE $<$ 0.27 BW and correlation coefficients above 0.99 for most subjects. Under multi-subject training, accuracy decreased, though good agreement (correlation coefficient $>$0.9) was still achieved for most subjects.

FFNN are a class of ANNs in which information flows strictly from inputs through one or more hidden layers to outputs without any cycles or feedback loops. Ngoh *et al* (2018) used a two-layer FFNN to estimate vertical GRF during treadmill running using a single-axis accelerometer on the right shoe. The network architecture comprised 10 sigmoid neurons in the hidden layer and 100 linear neurons in the output layer, trained using the Levenberg–Marquardt backpropagation algorithm with 280 samples for training, 120 for validation, and testing. The model achieved an average RMSE of less than 0.017 BW across different running speeds, accompanied by a cross-correlation coefficient exceeding 0.99. However, the sample size was relatively small ($n = 7$), which may limit the model's generalizability.

The MLP is the canonical FFNN architecture, featuring one or more fully connected hidden layers whose neurons apply nonlinear activation functions and are trained end-to-end via backpropagation. Leporace *et al* (2015) evaluated two MLP configurations: a single network with three outputs for the AP, ML, and V GRF, and three separate networks for each component from an accelerometer attached to the distal tibia of ($n = 17$) healthy participants. Input to the machine learning model consisted of 14 features, including 3D accelerometer data; relative velocity and displacement (first and second integrations of the accelerometer data); the first derivative of the accelerometer signal; stance duration; and time

points expressed as a percentage of stance time. The study reported high PCC for the AP (0.97) and vertical GRF (0.97). The ML component, while lower (0.80), still demonstrated moderate predictive capability. A further study by Leporace *et al* (2018) demonstrated high predictive accuracy, achieving PCCs of 0.98, 0.97, and 0.80 and MADs of 4.5% BW, 1.8% BW, and 1.4% BW for V, AP, and ML GRF, respectively. These findings demonstrate that even a simple FFNN, when paired with a well-chosen feature set, can capture kinetic parameters using data from just a single IMU.

3.3.4. Hybrid ANN

Hybrid or Physics-informed neural networks combine traditional neural networks with explicit knowledge of physical laws. Scheltinga *et al* (2023) developed a physics-informed hybrid ANN to estimate GRF waveforms during running based on Newton's second law of motion. The vertical GRF was calculated using the formula:

$$v\mathrm{GRF} = (m_b \cdot g) + i = \sum_{i=1}^{3} (m_\mathrm{b} \cdot \mathrm{WF}_i \cdot a_{z,i})$$

where:
m_b represents the body mass.
g is the gravitational acceleration.
WF_i are predefined weight factors for each sensor placement (0.55 for the pelvis and 0.23 for each tibia)
$a_{z,i}$ denotes the vertical component of the free acceleration measured by each IMU.

The authors trained seven separate ANNs, each with different splits of training and validation data ($n = 4$ subjects for validation and $n = 7$ subjects for training). The final prediction was obtained by averaging the outputs of these seven ANNs. The hybrid ANN achieved relative root mean square error (rRMSE) of 6.8% BW for vertical GRF, outperforming the direct ANN, which had an rRMSE of 7.3% BW

Kim *et al* (2020) integrated the SLIP model with ANNs to predict GRFs during walking from marker-derived sacral CoM kinematics emulating a single sacrum-mounted sensor. Two methods were evaluated. Method 1 utilized direct CoM kinematics, including displacement, velocity, acceleration, and temporal information of sacrum marker. Method 2, termed the SLIP hybrid method, transformed the CoM kinematics into approximate horizontal and vertical forces using the SLIP model, and then fed them into the ANN. Both methods demonstrated high correlation ($r \approx 0.98$) between predicted and measured GRFs. Method 1 achieved an average NRMSE of 6.93% for vertical GRF and 6.48% for AP GRF. In contrast, Method 2 showed improved performance with an average NRMSE of 6.30% for vertical GRF and a significantly lower 5.28% for AP GRF.

3.3.5. Recurrent neural networks (RNNs)

RNNs include loops or feedback connections that allow information to persist and be passed from one step of a sequence to the next. This means that at each timestep, an RNN takes in the current input as well as the hidden state (or memory) from the previous timestep, enabling the network to remember previous information and use it to influence current and future outputs

An ESN is a type of RNN in which the hidden layer, known as the reservoir, consists of a large, sparsely and randomly connected network of neurons whose internal weights remain fixed after random initialization. Only the output weights, which connect the reservoir to the output layer, are trained. This minimizes computational overhead and training time compared to traditional RNNs. Bach *et al* (2022) used ESNs to predict vertical GRFs from two triaxial accelerometers placed bilaterally on the shanks. For validation, 100-fold cross-validation and leave-M-out cross-validation were applied. The model achieved a coefficient of determination (R^2) of 0.96 and an NRMSE of 6.8 $\pm$ 0.3%. This indicates an excellent fit between predicted and measured GRFs. Furthermore, authors reported that the model maintained satisfactory performance even when trained on a limited subset of data ($\sim$17% of total strides), highlighting the data efficiency of ESNs compared to deep learning models

Havashinezhadian *et al* (2023) conducted a study to identify the optimal placement of IMUs for predicting GRF using ESN. Five synchronized IMUs were placed at different lower-limb locations. The authors reported that the IMU placed on top of the shoe consistently yielded the lowest mean absolute error (MAE) across the healthy (72.2%) and MKOA (41.7%) populations.

One of the main drawbacks of RNNs is the vanishing gradient problem, a phenomenon where gradients used to update network weights diminish exponentially during backpropagation (Hu *et al* 2021). This issue arises because RNNs process sequential data by unfolding over time, effectively creating deep architectures with multiple hidden layers. Each backward pass through these layers involves repeated

multiplication of gradients, particularly when using activation functions like sigmoid or hyperbolic tangent (tanh), whose derivatives are bounded. For long input sequences, these multiplicative effects cause gradients to shrink toward zero, preventing meaningful weight updates in earlier layers and degrading performance (Gers *et al* 2000). To address this, gated architectures like LSTM networks were introduced.

Donahue and Hahn (2023b) evaluated the performance of LSTM for estimation of vertical GRF waveforms during fixed pace running outside of laboratory settings using three IMU sensors placed on each foot and clipped to the waistband near the sacrum. RMSE for the stance phase of the GRF waveform ranged from 0.189–0.288 BW. This range aligns with the error margins found in similar studies.

Sun *et al* (2023) used LSTM to estimate vertical GRF in real-time during single- and double-leg drop landings using ($n = 8$) IMUs placed on chest, waist, thighs, shanks, and feet. The model reported high R^2 values of 0.88 $\pm$ 0.12 during single-leg drop landings, and 0.85 $\pm$ 0.11 during double-leg drop landings.

Wu *et al* (2020) used LSTM to predict 3D GRFs during gait using IMU sensors mounted beneath the walking surface, rather than attaching them to the body. Accuracy was assessed using NRMSE, yielding average values of 6.05% for ML, 3.93% for AP, and 4.37% for vertical GRF.

While a standard LSTM reads the input sequence in one direction, BD LSTMs consist of two separate LSTM layers that process the sequence in both directions: one forward (past to future) and one backward (future to past). This dual processing allows BD LSTMs to capture context from both preceding and succeeding elements in the sequence. The outputs from the forward and backward layers are then combined to form the final output. Alcantara *et al* (2022) used BD-LSTM complemented by an MLP to predict vertical GRF waveforms during uphill and downhill running using 3 biaxial accelerometers placed on the feet and sacrum. The model architecture comprised a BD-LSTM layer with a tanh activation function. Following the LSTM layer, the model incorporated an MLP comprising three fully connected layers. The model achieved RMSE of 0.16 $\pm$ 0.04 BW and rRMSE of 6.4 $\pm$ 1.5% across all conditions and subjects. This signifies a substantial improvement compared to other studies.

Donahue and Hahn (2023a) used BD-LSTM to estimate vertical GRF during running in an unconstrained environment. LOOCV was used to evaluate the model's performance. Authors reported whole-waveform RMSEs that varied substantially across speed and grade conditions and noted systematic underestimation at higher velocities and steeper grades. Lower agreement was observed for individuals not included in training, highlighting limited transferability across participants.

3.3.6. CNN

The CNN include convolutional layers, which allow the network to automatically learn spatial hierarchies of features by sliding filters across the input and generating activation maps that highlight the presence of specific patterns. However, CNNs are very data hungry. Dorschky *et al* (2020) augmented measured IMU data with simulated data generated through biomechanical simulations. They developed seven planar musculoskeletal models, one corresponding to each training subject. Each model consisted of seven rigid segments (trunk, thighs, shanks, feet) connected by six hinge joints (hip, knee, and ankle in each limb), yielding nine degrees of kinematic freedom. The models were scaled according to each subject's weight and height to ensure that the simulated movements were anatomically accurate (Winter 2009). The simulated biomechanical data was translated into IMU signals by virtually placing sensors at the same anatomical locations as in the measured dataset. Accelerometer data was simulated from segmental accelerations, while Gyroscope data was derived from the global trunk orientation and relative joint angular velocities. These synthetic data points were then combined with the measured data to create an expanded training dataset with varying ratios of simulated to measured samples (ranging from 1$\times$ to 16$\times$ the original measured data). The simulated data was iteratively evaluated alongside the measured data to assess their impact on the CNN's performance. The effect of adding simulated data was mixed rather than uniformly beneficial: across the three held-out participants, AP RMSE ranged from 3.76–5.10%BW and vertical RMSE from 9.99 to 16.24%BW across the tested ratios, improving for two participants and worsening for the third, with performance degrading as the simulated-to-measured ratio rose.

Johnson *et al* (2021) used CNNs to estimate GRFs using wearable accelerometer data. The authors used two pretrained CNN architectures, CaffeNet (Jia *et al* 2014) and ResNet-50 (He *et al* 2016), and modified them to perform multivariate regression by replacing the traditional SoftMax classification layer with a Euclidean loss layer. This allowed the networks to output continuous GRF waveforms instead of discrete class labels. Training was conducted on 458 372 trials of accelerometer data simulated from a motion capture dataset. To ensure the models could generalize beyond laboratory conditions, the test set consisted of accelerometer recordings from an independent inter-laboratory session involving ($n = 5$) athletes performing running and sidestepping tasks. The accelerometer data was transformed into 2D

images to fit the input requirements of the CNNs. The models demonstrated high correlations with ground-truth measurements of 0.85, 0.90, and 0.87 for V, AP, and ML GRF, respectively. These results indicate strong potential for CNNs to predict GRFs, although the variability in accuracy across components highlights areas for further improvement.

3.3.7. Ensemble net

Ensemble Nets combine the predictions of multiple individual models. Hossain *et al* (2023) introduced an ensemble framework, Kinetics fusion module (FM)-double loss regression (DLR)-Ensemble-Net, to estimate lower extremity joint moments and 3D GRFs using three IMU sensors placed on the thigh, shank, and foot. The model architecture consisted of gated recurrent units, one and two-dimensional convolutional layers, and fully connected dense layers to capture both temporal and spatial features from the IMU data. Finally, a FM integrates outputs from multiple primary models by learning optimal weights. The DLR strategy employed RMSE and PCC error functions to simultaneously minimize prediction errors and maximize correlation with the ground-truth data. To further improve performance, an ensemble learning technique using bagging was implemented, training multiple instances of the network on different bootstrap samples and aggregating the predictions. The model was evaluated using a publicly available dataset (Camargo *et al* 2021) that includes diverse walking conditions (treadmill, level ground, stairs, ramp) and speeds. LOOCV was used to validate the model's performance. Kinetics FM-DLR-Ensemble-Net demonstrated substantial improvements over existing deep learning models, achieving a mean NRMSE of 4.13% and a PCC of 0.931.

3.4. RQ4—evaluation metrics and validation protocols

This section synthesizes cross-validation approaches and the performance metrics used across studies, highlighting constraints on cross-study comparability.

3.4.1. Cross-validation techniques

Some studies used leave-one-subject-out (LOSO) validation, which ensures models generalize across different individuals by training on all but one subject and testing on the excluded subject (Lee and Park 2020, Hossain *et al* 2023, Sun *et al* 2023), while other studies used k-Fold cross-validation, which divides data into k subsets, training on k-1 and testing on the remaining subset, repeated k times (Leporace *et al* 2015). Hold-out validation was used by some studies, which involves splitting data into distinct training and testing sets, often with a portion reserved for validation (Wu *et al* 2020, Johnson *et al* 2021).

3.4.2. Evaluation metrics

A brief introduction to the evaluation metrics used across the studies included in this review is presented below. The specific metrics employed by each study are summarized in supplementary table S7.

RMSE: RMSE quantifies the average magnitude of the errors between predicted and actual values, with errors squared before averaging and then square-rooted. It measures how well a model's predictions match the observed data, expressed in the same units as the target variable. Lower RMSE values indicate better predictive accuracy, but RMSE is sensitive to outliers and is not scale-invariant.

rRMSE: rRMSE is the RMSE expressed relative to a reference value, typically as a percentage of the mean, peak, or peak-to-peak value of the observed data. This normalization allows comparison across datasets or variables with different scales. The term is incomplete without its denominator: rRMSE values computed against different reference quantities are not interchangeable, and this review therefore records the denominator, evaluation window, and aggregation level reported by each source in supplementary table S7.

NRMSE: NRMSE is RMSE divided by a normalization factor, such as the range, mean, or standard deviation of the observed data. It provides a scale-independent measure, but for the same reason as rRMSE it cannot be compared across studies unless the normalization factor is stated explicitly.

MAE: MAE calculates the average of the absolute differences between predicted and actual values. Unlike RMSE, MAE treats all errors equally and is less sensitive to large outliers.

MAD: as used by the included Leporace studies (Leporace *et al* 2015, 2018), MAD is the mean absolute prediction error, which is mathematically equivalent to MAE for the equation those sources define. The

abbreviation is used elsewhere in the literature for the MAD about a sample mean, which is a different quantity; the present review uses it only in the prediction-error sense defined here.

Normalized mean absolute deviation (NMAD): NMAD is the MAD normalized to a reference value, such as the mean or range, enabling scale-independent interpretation.

Pearson's correlation coefficient (r): Pearson's r measures the linear correlation between predicted and actual values, ranging from -1 (perfect negative correlation) to $+1$ (perfect positive correlation).

Cross-correlation coefficient (p): Cross-correlation is a generalized form of Pearson's correlation coefficient, especially when analyzing time series data. While Pearson's correlation measures the linear relationship between two variables at corresponding time points, cross-correlation assesses how well the values of one series at time t align with the values of another series at different time lags. For example, consider two time series $\bar{x}, \bar{y}$ and standard deviations σ_x, σ_y. The sample cross-correlation matrix R is defined by

$$R_{t,s} = \frac{(x_t - \bar{x})\,(y_s - \bar{y})}{\sigma_x \sigma_y} \text{for } t, s = 1, \ldots, T,$$

so that each entry (t, s) compares x_t with y_s. Because we allow all possible pairs of lags between the two series, R is $T \times T$.

Pearson's correlation coefficient appears on the main diagonal (where $s = t$):

$$r_{xy} = \frac{\sum_{t=1}^{T} (x_t - \bar{x})\,(y_t - \bar{y})}{\sigma_x \sigma_y}.$$

Coefficient of determination (R^2): R^2 measures the fraction of the total variance in the observed data that the model captures.

4. Discussion

This systematic review critically examined 26 reports that employed machine learning and deep learning techniques to estimate GRFs from IMU sensors. The review examined how these methods are currently used and highlighted gaps that must be addressed before machine learning-based GRF estimation methods can be clinically useful. The analysis was structured around five research questions.

4.1. Participant populations and experimental protocols

The first research question examined which populations were included in machine learning-based GRF studies and how experimental design choices influence the generalizability of their findings. Across the literature, studies typically recruited young, healthy adults, with median ages around 25 years and sample sizes ranging from 5–47 participants. Eight reports included fewer than 10 participants. Such narrow and homogeneous datasets increase the risk of overfitting and limit the statistical power and clinical applicability of learned models (Halilaj *et al* 2018). Age-related variations in biomechanics are particularly significant. For instance, older adults exhibit a 14.5% decrease in plantar-flexor strength compared to younger individuals (DeVita and Hortobagyi 2000), and they tend to compensate for these deficits by increasing hip extension during the initial vertical GRF peak (DeVita and Hortobagyi 2000). Deep models trained on young adults may therefore misinterpret signals from older populations.

Similar challenges arise in pathological gait. After neural injury, reduced push-off force output is common and correlates with slower walking speeds, typically manifesting as diminished AP GRF in late stance (Chen *et al* 2005). Individuals with Parkinson's disease, when walking at matched speeds, exhibit shorter single-support phases, reduced vertical GRF peaks, and reduced stride times (∼10%–15%), with increased variability in force profiles (Morris *et al* 1994, Sharifmoradi *et al* 2015). These subtle deviations are difficult for machine learning and deep learning algorithms trained exclusively on healthy datasets to interpret reliably. The limitations extend to pediatric and developmental conditions. Ambulatory children with spastic CP, for example, frequently exhibit an absence of the second vertical GRF peak ('Ben Lomond' profile), with up to 87% of cases showing a flattened late-stance force component (Rodda *et al* 2004). Despite the clinical relevance, such pathological data are largely absent from existing datasets.

Consequently, the exclusion of older adults, children, and individuals using assistive devices leaves a critical gap in understanding how IMU-derived features map to GRFs across the lifespan and under pathological conditions, representing a key limitation for model generalizability.

The variation in sensor placement ($n = 1$ to $n = 12$) across studies is also substantial. A subgroup analysis or meta-regression on optimal sensor configurations would strengthen the practical recommendations.

Experimental protocols also varied widely. Laboratory protocols often constrain participants to walk or run on treadmills, whereas others allow over-ground ambulation. Treadmills impose steady-state gait with low stride variability (Dingwell and Cusumano 2000). Models built under these conditions often report lower rRMSE for vertical GRF however, vertical GRF peaks can differ by 2%–5% between treadmill and overground conditions due to altered propulsion dynamics and reduced push-off forces on treadmills (Riley *et al* 2007, Lee and Hidler 2008). In contrast, free-living protocols introduce ecological variability, including turning, obstacle avoidance, and fluctuations in walking speed. These factors increase biomechanical variability, with coefficients of variation for peak forces up to 50% higher outdoors than on a treadmill (Lee and Hidler 2008). Dynamic tasks such as jumping and drop-landing pose additional challenges. Impact transients contain high-frequency components (>40 Hz), which standard IMUs sampled at 100–200 Hz may not capture, and accelerations during landings can saturate sensors with ±16 g ranges (Cerfoglio *et al* 2021). These factors emphasize that protocols influence what frequency content must be preserved and the degree of sensor noise or saturation that models must withstand.

Collectively, these findings highlight that both population selection and experimental design strongly influence model performance. Datasets collected from narrow demographic groups under tightly controlled laboratory conditions underestimate the variability observed in real-world settings. As a result, machine learning-based GRF models trained on such datasets risk poor generalizability when applied to diverse populations, pathological conditions, or free-living environments.

4.2. Preprocessing and feature extraction

IMU-GRF studies rely on careful preprocessing to align IMU data with the ground truth. Synchronization is crucial because IMUs and force plates operate on separate clocks. Foot-stomp cross-correlation is often used to temporally align the signals (Kerns *et al* 2023). However, even small residual errors cause phase shifts between predicted and measured GRF, artificially inflating correlation coefficients. Some studies iteratively corrected clock drift to reduce timing errors to a few milliseconds, but such approaches are seldom used and warrant broader adoption (Alcantara *et al* 2022).

Filtering is another important preprocessing step. Butterworth low-pass filters with cut-off frequencies between 10 and 35 Hz are commonly applied to attenuate high-frequency noise. This range introduces a trade-off between signal preservation and noise reduction. During running tasks, heel-strike transients can contain energy up to 20 Hz (Gruber *et al* 2011), therefore cut-offs below 20 Hz remove meaningful information, whereas cut-offs above 30 Hz allow noise from sensor mounting compliance. Zero-phase filtering removes phase distortion but is non-causal, complicating real-time deployment. Wavelet-based methods have been proposed as an alternative, offering improved noise rejection while preserving transient features (Avrutov *et al* 2022).

Studies also differ in their choice of features. Classical machine learning approaches often extract summary statistics (mean, variance, RMS), temporal features (stance time, stride frequency), or integrate accelerations to obtain velocities and displacements. Physics-informed features, such as estimates of CoM acceleration, can reduce training data requirements and improve interpretability. Deep learning approaches, in contrast, operate on raw accelerometer data and automatically learn hierarchical features. While end-to-end learning eliminates manual feature engineering, it demands large training datasets. Hybrid methods that incorporate domain knowledge into neural networks have shown promise, suggesting that combining physics-based features with learned representations may offer an optimal balance. These preprocessing choices cannot yet be linked causally to model performance. The reviewed literature rarely varies a single preprocessing decision while holding the dataset, model architecture, and validation split constant, so differences in reported error cannot be attributed to a particular filter, normalization, or feature family. Matched ablation studies under participant-independent splits are needed.

Whether hybrid or physics-informed models outperform conventional data-driven models is not yet settled by the reviewed evidence. Scheltinga *et al* (2023) found that adding a physical model reduced vertical rRMSE from 7.3% to 6.8%, while the AP component was unchanged at 7.8% and the ML component moved only from 10.9% to 10.8%; the statistically significant advantage was confined to the vertical component. Kim *et al* (2020) reported that a SLIP transform improved AP NRMSE from 6.48% to 5.28%, with vertical NRMSE changing from 6.93% to 6.30% and only the AP improvement reaching

Table 4. Positioning of the present review against earlier reviews.

Review	Primary scope	Methods and outcomes	Relationship to the present review
Ancillao *et al* (2018)	Wearable inertial estimation of GRF and moments	Direct biomechanical models and early machine learning	Broad methodological taxonomy; predates most deep-learning work reviewed here
Liang *et al* (2023)	IMU-based biomechanical analysis generally	Wide range of mechanical targets and applications	Much broader scope; GRF waveform estimation not treated as the primary outcome
Krishnakumar *et al* (2024)	Kinetic estimation for ACL rehabilitation	Clinical decision-making focus within one pathway	Single clinical population; not a cross-model-family synthesis
Present review	Continuous GRF waveform estimation from IMUs	Model families, V/AP/ML components, sensor configurations, validation levels, metric definitions	Pipeline-level synthesis with explicit metric-definition auditing

significance, but that experiment used marker-derived rather than physical IMU input. Dorschky *et al* (2020) found that augmenting measured data with physics-simulated data improved performance for two of three held-out participants and degraded it for the third. Physics-informed structure is therefore best described as a promising inductive bias that reduces data requirements and improves interpretability, not as an established performance winner; matched comparisons on identical datasets, splits, and metrics are required to settle the question.

4.3. Machine-learning and deep-learning model comparison

Traditional machine learning models, such as random forests and system-identification methods, remain competitive when training data are limited. Random forests are effective at capturing nonlinear relationships, are robust to outliers, and require minimal hyperparameter tuning. System-identification approaches, such as NARMAX, provide interpretable polynomial relationships but rely on assumptions of periodic and steady-state behavior, making them less suitable for irregular or non-periodic tasks.

The present review is positioned against earlier syntheses as follows. Ancillao *et al* (2018) reviewed indirect measurement of GRFs and moments from wearable inertial sensors, covering direct biomechanical modeling and early machine learning, and predates most of the deep-learning literature examined here. Liang *et al* (2023) surveyed IMU-based biomechanical analysis far more broadly, across many mechanical targets rather than GRF waveforms specifically. Krishnakumar *et al* (2024) organized kinetic estimation around anterior cruciate ligament rehabilitation and clinical decision-making, and so addresses a single clinical pathway. None of these reviews compares continuous GRF waveform estimation across model families, all three components, sensor configurations, validation levels, and metric definitions simultaneously, which is the gap the present review addresses. Table 4 sets out this positioning.

In contrast, ANNs have been the most widely adopted due to their ability to model complex, nonlinear relationships. However, ANNs process each input vector independently and lack inherent temporal memory. They also rely heavily on hand-engineered features, struggle to capture long-range temporal dependencies, and often require per-subject calibration (Sharma *et al* 2021).

CNNs demonstrated superior performance when trained on raw IMU data. CNNs are biased toward discerning spatial patterns in data, giving them the remarkable capacity to achieve translational invariance. Some studies utilized an LSTM, which is specialized in identifying temporal relationships. However, the reduced performance of LSTMs over CNNs justifies that:

1. The relationship between the quaternions evolving over time and the corresponding change in GRF depends not only on temporal relationships but also on spatiotemporal relationships. Hence, it is recommended that a combination of LSTM in CNN layers should be used to test this assertion.
2. CNN 1-D layers can filter out noise, similar to an adaptive filter that learns noise from a signal based on the output performance, whereas the LSTM may compound that error. Hence, at higher velocities, LSTMs' error increases, which was also reported by Donahue and Hahn (2023a), giving credence to our view that LSTMs on their own tend to compound the noise.

Overall model performance shows a clear split. Vertical GRF prediction is reliable ($r > 0.9$). By contrast, AP and ML errors remain higher. The type of activity also impacted model performance. Models developed for treadmill walking at controlled speeds (Lim *et al* 2019, Lee and Park 2020) generally showed higher accuracy than those tested under overground or free-walking conditions (Guo *et al* 2017).

Activities involving more dynamic movements, such as running or jumping, posed additional challenges due to greater gait variability and higher impact forces (Cerfoglio *et al* 2021).

Validation of models in real-world environments is essential to assess their practical applicability and robustness. Several studies (Guo *et al* 2017, Donahue and Hahn 2023a, 2023b) extended their methods beyond controlled laboratory settings, demonstrating that their models could estimate GRFs accurately in unconstrained environments. However, in outdoor settings, these studies used pressure insoles, which can only provide GRF data in the vertical direction; hence, they estimated only vertical GRF.

IMU-GRF estimation can be framed as a sequence-to-sequence regression problem, where an input IMU window is mapped to a GRF waveform over the same interval. A key limitation of sequence-to-sequence models based on RNNs and CNNs is that they require the entire input sequence to be compressed into a single fixed-length context vector. This approach makes it challenging for the model to capture all the necessary information when dealing with long or complex input sequences. Cho *et al* (2015) demonstrated that the performance of RNN-based models deteriorates rapidly as the length of the input sequence increases. To address this issue, Bahdanau *et al* (2016) introduced the attention mechanism. Their innovation was to eliminate the reliance on a single fixed-length context vector and instead compute a unique context vector for each output element. This is achieved by creating a weighted combination of all encoder hidden states, allowing the model to dynamically focus on different parts of the input sequence as needed. As a result, information does not need to be compressed into a single vector. Future research should incorporate attention mechanisms into RNNs, CNNs, and other neural architectures to further improve model accuracy and performance.

4.4. Evaluation metrics and validation protocols

The evaluation of performance across studies is impeded by inconsistent methodologies and metrics. Studies adopt diverse normalization techniques such as peak force, mean force, or bodyweight, each with distinct implications. Normalization to peak force enhances precision in predicting peak values but neglects waveform morphology (Jlassi and Dixon 2024), whereas mean-force normalization unevenly penalizes errors across the stance and swing phases. Bodyweight normalization facilitates inter-subject comparisons but masks individual variability. Furthermore, the inclusion or exclusion of the flight phase affects RMSE calculations. Including the zero-force swing phase artificially reduces RMSE. Temporal waveform misalignments can inflate correlation coefficients (temporal misalignment here denoting residual time or phase offset between predicted and reference GRF signals recorded on separate clocks), allowing high correlations to coexist with poor waveform congruence when phase shifts occur (Iosa *et al* 2014). Validation strategies also differ. LOSO cross-validation improves generalizability assessment but demands larger datasets; *k*-fold cross-validation provides greater statistical power but risks data leakage; and simple train/test splits yield less robust evaluations. Reporting practice should therefore be standardized on a single primary metric. We recommend that stance-phase RMSE normalized to body weight, $\mathrm{RMSE_\%BW} = 100 \times \mathrm{RMSE}/(\mathrm{m} \times \mathrm{g})$, where m is body mass and g is gravitational acceleration, be adopted as the mandatory primary magnitude metric for future studies. Body-weight normalization is preferred over peak or range normalization because it preserves an interpretable force scale and supports comparison between participants without dividing by a quantity that is itself altered by pathology: in cohorts with attenuated waveforms, peak or range normalization makes an identical absolute error appear artificially different across axes and populations. Because a magnitude metric alone does not capture waveform morphology, it should be accompanied by waveform correlation or R-squared, peak-magnitude error, and peak-timing error, which together preserve shape and timing information. Peak-, range-, or mean-normalized metrics may be reported as secondary outcomes only when the exact denominator, evaluation window, and aggregation level are stated. Alongside this, future work should standardize evaluation procedures by (1) normalizing GRF to bodyweight, (2) computing error metrics only over the stance phase, (3) correcting for temporal misalignment using cross-correlation or time normalization, and (4) using LOSO or nested k-fold cross-validation rather than simple hold-out splits.

4.5. Deployment challenges and future directions

- Machine learning-based estimation of GRF from IMUs has progressed from proof-of-concept demonstrations to consistently high vertical-GRF accuracy under controlled conditions. It has not, however, reached clinical validation. Judged against the criteria that clinical use would require, namely decision-linked magnitude and timing accuracy, a reproducible calibration and setup procedure, subject-independent validation, prospective external validation at an independent site or on independent hardware, evidence in clinically representative populations, robustness to sensor reattachment and free-living conditions, and causal real-time operation, no approach in this review satisfies the full set (table 5). Nor did the reviewed literature establish a universal minimum accuracy threshold for

Table 5. Clinical-readiness criteria assessed against the current synthesis.

Readiness criterion	What clinical use would require	Evidence in the current synthesis	Required next step
Decision-linked accuracy	Error below the smallest clinically meaningful difference	No consensus threshold identified; acceptability never tied to a stated decision	Define target decisions and their minimum detectable change
Calibration and setup burden	Reproducible, brief, non-specialist setup	1–16 sensors recorded; setup and calibration rarely reported in full	Report setup time, calibration steps, and sensors actually used
Subject-independent validation	Performance on unseen participants	Present in many studies, but some used sample- or step-level splits	Mandate participant-disjoint splits and report the split explicitly
External validation	Independent site, cohort, or hardware	Rare; mostly single-laboratory, single-device	Prospective multi-site validation on independent hardware
Clinical populations	Evidence in the intended patient group	Predominantly young healthy adults; one osteoarthritis cohort (Ullah *et al* 2026)	Recruit clinically representative and pathological cohorts
Real-world robustness	Stable across reattachment, footwear, and free living	Reattachment and placement-offset sensitivity seldom quantified	Quantify sensitivity to reattachment and placement error
Real-time operation	Causal, low-latency inference	Zero-phase filtering and windowing often non-causal	Report causal-pipeline latency and online feasibility

clinical use: no consensus value exists, and acceptable error is not a single number but depends on the clinical decision being supported, the GRF component, the task, the population, and the smallest difference that would change management. Setup complexity can be summarized by proxy, and table 2 therefore now reports the number of sensors recorded separately from the number actually entering each model, since the two differ in several studies.

- Most datasets involve young, healthy adults in treadmill/lab protocols. Overground/outdoor locomotion can shift the IMU-GRF relationship, so treadmill-trained models may not generalize without real-world training/validation.
- Many outdoor studies use pressure insoles, which are often vertical-only and sensitive to drift and footwear, constraining AP/ML model training and reducing comparability across force plates, instrumented treadmills, and insoles.
- Studies used 1–16 IMUs with heterogeneous placements. Multi-sensor setups may improve observability but reduce clinical practicality due to setup time and placement error; single-sensor approaches are more deployable but demand robust placement choices and soft-tissue artifact handling. Future work should quantify sensitivity to placement offsets and reattachment. As a practical starting point, sacrum or lower-back placement is recommended for single-sensor estimation of vertical and AP GRF in level walking, because the site is close to the CoM, is quick to attach, and is less obtrusive than distal arrays. This recommendation is conditional rather than exclusive: shank and foot placements retain impact and gait-event information that matters in running, and ML estimation during turning or asymmetric tasks may require additional lower-limb sensors. Few studies compared placements under identical participants, tasks, models, and validation, so the frequency with which a location appears in the literature is not evidence of its superiority.
- Cross-study comparisons are hindered by inconsistent normalization, evaluation windows (stance-only vs full-waveform), and validation strategies. Future studies should standardize by reporting BW-normalized forces, using clearly defined windows and reporting stance-only and flight phase separately, addressing temporal misalignment, and applying subject-independent validation with transparent splits.

Moreover, future research should focus on expanding participant diversity to include varied populations, including different groups and those with gait abnormalities. Incorporating synthesized data from physics-based models, as proposed by Dorschky *et al* (2020), can reduce the need for extensive experimental data collection. Future research should also focus on testing their model on some standardized public datasets, such as:

1. The comprehensive, open-source dataset of lower-limb biomechanics across multiple conditions of stairs, ramps, level ground ambulation, and transitions (Camargo *et al* 2021).

2. The dataset of asymptomatic human gait and movements obtained from markers, IMUs, insoles, and force plates (Grouvel *et al* 2023).
3. The High-density EMG, IMU, kinetic, and kinematic open-source data for comprehensive locomotion activities (Dimitrov *et al* 2023).
4. The comprehensive dataset on biomechanics and motor control during human walking with discrete mechanical perturbations (Lorenz and van den Bogert 2024).
5. The gait data from 51 healthy participants, collected using motion capture, IMUs, and computer vision (Lavikainen *et al* 2024).

Several limitations of this review should be noted. Eligibility was restricted to peer-reviewed English-language reports, which introduces potential language bias, and because language was not coded as a separate exclusion reason during the original screen the size of that effect cannot be quantified. The literature search was updated to June 2026; the four eligible reports identified through the update entered the review as additional records, as shown in figure 3. Finally, the heterogeneity of metric definitions across the included studies means that this review reports distributions and within-study comparisons rather than a pooled quantitative synthesis. The re-checked extraction was independently confirmed by a second reviewer.

5. Conclusion

This systematic review identified 26 reports that used machine learning or deep learning methods to estimate GRF waveforms from IMU data. Overall, the literature shows strong performance for vertical GRF, particularly in controlled treadmill protocols, while AP and ML components remain more challenging and less consistently reported. Most studies relied on small, homogeneous cohorts of young, healthy adults and employed highly variable experimental protocols, sensor counts, and placements, thereby limiting cross-study comparability and clinical translation. Across pipelines, key recurring methodological choices included signal synchronization, low-pass filtering, and bodyweight normalization. Classical machine learning approaches typically depended on hand-engineered features, whereas deep learning approaches used raw IMU inputs. Finally, the review highlights that inconsistent validation strategies and error metrics hinder benchmarking, and that broader datasets and more standardized evaluation are needed before IMU-based GRF estimation can be deployed reliably in real-world clinical and biomechanical settings.

Acknowledgments

The authors wish to acknowledge use of the eResearch Infrastructure Platform hosted by the Crown company, Research and Education Advanced Network New Zealand (REANNZ) Ltd, and funded by the Ministry of Business, Innovation & Employment.

Data availability statement

No new data were created or analysed in this study.

Supplementary Tables S1-S7 available at https://doi.org/10.1088/1873-4030/aeae3c/data1.

Funding

The authors received no funding for this work.

Ethics statement

This article is a systematic review of previously published studies. No new data involving human participants or animals were collected by the authors, and ethical approval was therefore not required.

Author contributions

Haider Ali
Conceptualization (equal), Writing – original draft (equal), Writing – review & editing (equal)

Imran Khan Niazi
Conceptualization (equal), Resources (equal), Writing – review & editing (equal)

Umair Javaid Manj 0009-0008-7085-3662
Formal analysis (equal), Writing – review & editing (equal)

Denise Taylor
Methodology (equal), Supervision (equal), Writing – review & editing (equal)

References

Alcantara R S, Edwards W B, Millet G Y and Grabowski A M 2022 Predicting continuous ground reaction forces from accelerometers during uphill and downhill running: a recurrent neural network solution *PeerJ* **10** e12752
Ancillao A, Tedesco S, Barton J and O'flynn B 2018 Indirect measurement of ground reaction forces and moments by means of wearable inertial sensors: a systematic review *Sensors* **18** 2564
Avrutov V *et al* 2022 Wavelet De-noising of mems inertial measurement unit for autonomous latitude determination *SSRN Electron. J.* **12** 344–51
Bach M M, Dominici N and Daffertshofer A 2022 Predicting vertical ground reaction forces from 3D accelerometry using reservoir computers leads to accurate gait event detection *Front. Sports Act Living* **4** 1037438
Bahdanau D, Cho K and Bengio Y 2016 Neural machine translation by jointly learning to align and translate (arXiv:1409.0473)
Barela A M, de Freitas P B, Celestino M L, Camargo M R and Barela J A 2014 Ground reaction forces during level ground walking with body weight unloading *Braz J. Phys. Ther.* **18** 572–9
Boyer K A, Johnson R T, Banks J J, Jewell C and Hafer J F 2017 Systematic review and meta-analysis of gait mechanics in young and older adults *Exp. Gerontol.* **95** 63–70
Camargo J, Ramanathan A, Flanagan W and Young A 2021 A comprehensive, open-source dataset of lower limb biomechanics in multiple conditions of stairs, ramps, and level-ground ambulation and transitions *J. Biomech.* **119** 110320
Cerfoglio S, Galli M, Tarabini M, Bertozzi F, Sforza C and Zago M 2021 Machine learning-based estimation of ground reaction forces and knee joint kinetics from inertial sensors while performing a vertical drop jump *Sensors* **21** 7709
Chen G, Patten C, Kothari D H and Zajac F E 2005 Gait differences between individuals with post-stroke hemiparesis and non-disabled controls at matched speeds *Gait Posture* **22** 51–56
Childs J D, Sparto P J, Fitzgerald G K, Bizzini M and Irrgang J J 2004 Alterations in lower extremity movement and muscle activation patterns in individuals with knee osteoarthritis *Clin. Biomech.* **19** 44–49
Cho J, Lee K, Shin E, Choy G and Do S 2015 How much data is needed to train a medical image deep learning system to achieve necessary high accuracy (arXiv:1511.06348)
Chockalingam N *et al* 2018 Interpreting ground reaction forces in gait ed B Müller *Handbook of Human Motion* (Springer) pp 609–23
Costello K E, Felson D T, Neogi T, Segal N, Lewis C E, Gross K D, Nevitt M, Lewis C L and Kumar D 2019 Ground reaction force patterns in knees with and without pain and radiographic osteoarthritis: descriptive analyses from a large cohort study *Osteoarthr. Cartil.* **27** S142–3
Davis B L, Perry J E, Neth D C and Waters K C 1998 A device for simultaneous measurement of pressure and shear force distribution on the plantar surface of the foot *J. Appl. Biomech.* **14** 93–104
DeVita P and Hortobagyi T 2000 Age causes a redistribution of joint torques and powers during gait *J. Appl. Physiol.* **88** 1804–11
Dimitrov H, Bull A M J and Farina D 2023 High-density EMG, IMU, kinetic, and kinematic open-source data for comprehensive locomotion activities *Sci. Data* **10** 789
Dingwell J B and Cusumano J P 2000 Nonlinear time series analysis of normal and pathological human walking *Chaos* **10** 848–63
Donahue S R and Hahn M E 2023a Estimation of gait events and kinetic waveforms with wearable sensors and machine learning when running in an unconstrained environment *Sci. Rep.* **13** 2339
Donahue S R and Hahn M E 2023b Estimation of ground reaction force waveforms during fixed pace running outside the laboratory *Front. Sports Act Living* **5** 974186
Dorschky E, Nitschke M, Martindale C F, van den Bogert A J, Koelewijn A D and Eskofier B M 2020 CNN-based estimation of sagittal plane walking and running biomechanics from measured and simulated inertial sensor data *Front. Bioeng. Biotechnol.* **8** 604
Faivre A, Dahan M, Parratte B and Monnier G 2004 Instrumented shoes for pathological gait assessment *Mech. Res. Commun.* **31** 627–32
Gempfer C, Santos G F, Hobein A, Kuhlgatz T, Hurschler C and Jakubowitz E 2026 Predicting gait kinetics using 3-degrees of freedom acceleration data and artificial neural networks *Clin. Biomech.* **134** 106775
Gers F A, Schmidhuber J and Cummins F 2000 Learning to forget: continual prediction with LSTM *Neural Comput.* **12** 2451–71
Giardini M, Turcato A M, Arcolin I, Corna S and Godi M 2023 Vertical ground reaction forces in parkinson's disease: a speed-matched comparative analysis with healthy subjects *Sensors* **24** 179

Grouvel G, Carcreff L, Moissenet F and Armand S 2023 A dataset of asymptomatic human gait and movements obtained from markers, IMUs, insoles and force plates *Sci. Data* **10** 180
Gruber A H, Davis I S and Hamill J 2011 Frequency content of the vertical ground reaction force component during rearfoot and forefoot running patterns *Med. Sci. Sports Exerc.* **43** 60
Guo Y, Storm F, Zhao Y, Billings S, Pavic A, Mazzà C and Guo L-Z 2017 A new proxy measurement algorithm with application to the estimation of vertical ground reaction forces using wearable sensors *Sensors* **17** 2181
Halilaj E, Rajagopal A, Fiterau M, Hicks J L, Hastie T J and Delp S L 2018 Machine learning in human movement biomechanics: best practices, common pitfalls, and new opportunities *J. Biomech.* **81** 1–11
Havashinezhadian S, Chiasson-poirier L, Sylvestre J and Turcot K 2023 Inertial sensor location for ground reaction force and gait event detection using reservoir computing in gait *Int. J. Environ. Res. Public Health* **20** 3120
He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 770–8
Hossain M S B, Guo Z and Choi H 2023 Estimation of lower extremity joint moments and 3D ground reaction forces using IMU sensors in multiple walking conditions: a deep learning approach *IEEE J. Biomed. Health Inform.* **27** 2829–40
Hu Z, Zhang J and Ge Y 2021 Handling vanishing gradient problem using artificial derivative *IEEE Access* **9** 22371–7
Huang P, Mostovov A, Cohen R, Cadilhac C and Pionnier R 2025 Comparison of feetme insoles with a motion capture system coupled to force plates for assessing gait and posture *Sci. Rep.* **15** 13476
Iosa M, Cereatti A, Merlo A, Campanini I, Paolucci S and Cappozzo A 2014 Assessment of waveform similarity in clinical gait data: the linear fit method *Biomed. Res. Int.* **2014** 214156
Jia Y *et al* 2014 *Caffe: Convolutional Architecture for Fast Feature Embedding* (Association for Computing Machinery, in MM '14) pp 675–8
Jiang X, Napier C, Hannigan B, Eng J J and Menon C 2020 Estimating vertical ground reaction force during walking using a single inertial sensor *Sensors* **20** 4345
Jlassi O and Dixon P C 2024 The effect of time normalization and biomechanical signal processing techniques of ground reaction force curves on deep-learning model performance *J. Biomech.* **168** 112116
Johnson W R, Mian A, Robinson M A, Verheul J, Lloyd D G and Alderson J A 2021 Multidimensional ground reaction forces and moments from wearable sensor accelerations via deep learning *IEEE Trans. Biomed. Eng.* **68** 289–97
Jolliffe I T and Cadima J 2016 Principal component analysis: a review and recent developments *Phil. Trans. A* **374** 20150202
Jonsdottir J, Lencioni T, Gervasoni E, Crippa A, Anastasi D, Carpinella I, Rovaris M, Cattaneo D and Ferrarin M 2020 Improved gait of persons with multiple sclerosis after rehabilitation: effects on lower limb muscle synergies, push-off, and toe-clearance *Front. Neurol* **11** 668
Kelleher K J, Spence W D, Solomonidis S E and Apatsidis D P 2009 The effect of impaired plantar sensation on gait in people with multiple sclerosis *Int. J. MS Care* **11** 25–31
Kellermeyer L, Harnke B and Knight S 2018 Covidence and rayyan *J. Med. Libr Assoc.* **106** 580
Kerns J A, Zwart A S, Perez P S, Gurchiek R D and McBride J M 2023 Effect of IMU location on estimation of vertical ground reaction force during jumping *Front. Bioeng. Biotechnol.* **11** 1112866
Kim B, Lim H and Park S 2020 Spring-loaded inverted pendulum modeling improves neural network estimation of ground reaction forces *J. Biomech.* **113** 110069
Kim D, Jin Y, Cho H, Jones T, Zhou Y M, Fadaie A, Popov D, Swaminathan K and Walsh C J 2025 Learning-based 3D human kinematics estimation using behavioral constraints from activity classification *Nat. Commun.* **16** 3454
Kloeckner J, Visscher R M S, Taylor W R, Viehweger E and De Pieri E 2023 Prediction of ground reaction forces and moments during walking in children with cerebral palsy *Front. Hum. Neurosci.* **17** 1127613
Kmet L M, Cook L S and Lee R C 2004 Standard quality assessment criteria for evaluating primary research papers from a variety of fields (https://doi.org/10.7939/R37M04F16)
Krishnakumar S, van Beijnum B F, Baten C T M, Veltink P H and Buurke J H 2024 Estimation of kinetics using IMUs to monitor and aid in clinical decision-making during ACL rehabilitation: a systematic review *Sensors* **24** 2163
Lavikainen J, Vartiainen P, Stenroth L, Karjalainen P A, Korhonen R K, Liukkonen M K and Mononen M E 2024 Gait data from 51 healthy participants with motion capture, inertial measurement units, and computer vision *Data Brief* **56** 110841
Lee M and 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* **20** 6277
Lee S J and Hidler J 2008 Biomechanics of overground vs. treadmill walking in healthy individuals *J. Appl. Physiol.* **104** 747–55
Leporace G, Batista L A, Metsavaht L and Nadal J 2015 Residual analysis of ground reaction forces simulation during gait using neural networks with different configurations *2015 37th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* pp 2812–5
Leporace G, Batista L A and Nadal J 2018 Prediction of 3D ground reaction forces during gait based on accelerometer data *Res. Biomed. Eng.* **34** 211–6
Liang W, Wang F, Fan A, Zhao W, Yao W and Yang P 2023 Extended application of inertial measurement units in biomechanics: from activity recognition to force estimation *Sensors* **23** 4229
Lim H, Kim B and Park S 2019 Prediction of lower limb kinetics and kinematics during walking by a single IMU on the lower back using machine learning *Sensors* **20** 130
Liu T, Inoue Y and Shibata K 2010 A wearable ground reaction force sensor system and its application to the measurement of extrinsic gait variability *Sensors* **10** 10240–55
Lorenz D L and van den Bogert A J 2024 A comprehensive dataset on biomechanics and motor control during human walking with discrete mechanical perturbations *PeerJ* **12** e17256
McClay I S and Cavanagh P R 1994 Relationship between foot placement and mediolateral ground reaction forces during running *Clin. Biomech.* **9** 117–23
Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P and Stewart L A 2015 Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement *Syst. Rev.* **4** 1
Morris M E, Iansek R, Matyas T A and Summers J J 1994 The pathogenesis of gait hypokinesia in Parkinson's disease *Brain* **117** 1169–81
Ngoh K J, Gouwanda D, Gopalai A A and Chong Y Z 2018 Estimation of vertical ground reaction force during running using neural network model and uniaxial accelerometer *J. Biomech.* **76** 269–73
Ochoa J, Sternad D and Hogan N 2017 Treadmill vs. overground walking: different response to physical interaction *J. Neurophysiol.* **118** 2089–102

Pagnacco G, Silva A, Oggero E and Berme N 2000 Inertially compensated force plate: a means for quantifying subject's ground reaction forces in non-inertial conditions *Biomed. Sci. Instrum.* **36** 397–402
Razak A H, Zayegh A, Begg R K and Wahab Y 2012 Foot plantar pressure measurement system: a review *Sensors* **12** 9884–912
Riley P O, Paolini G, Della Croce U, Paylo K W and Kerrigan D C 2007 A kinematic and kinetic comparison of overground and treadmill walking in healthy subjects *Gait Posture* **26** 17–24
Rodda J M, Graham H K, Carson L, Galea M P and Wolfe R 2004 Sagittal gait patterns in spastic diplegia *J. Bone Joint Surg. Br.* **86** 251–8
Roetenberg D, Luinge H J, Baten C T and Veltink P H 2005 Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation *IEEE Trans. Neural Syst. Rehabil. Eng.* **13** 395–405
Rutz E, Donath S, Tirosh O, Graham H K and Baker R 2013 Explaining the variability improvements in gait quality as a result of single event multi-level surgery in cerebral palsy *Gait Posture* **38** 455–60
Salehi A, Khandan A, Arab Baniasad M, Baghdadi S, Farahmand F and Zohoor H 2018 P 043—Center of pressure progression and ground reaction forces are altered in cerebral palsy crouch gait *Gait Posture* **65** 307–8
Schardt C, Adams M B, Owens T, Keitz S and Fontelo P 2007 Utilization of the PICO framework to improve searching PubMed for clinical questions *BMC Med. Inform. Decis. Mak* **7** 16
Scheltinga B L, Kok J N, Buurke J H and Reenalda J 2023 Estimating 3D ground reaction forces in running using three inertial measurement units *Front. Sports Act Living* **5** 1176466
Shafizadegan Z, Karimi M T, Shafizadegan F and Rezaeian Z S 2016 Evaluation of ground reaction forces in patients with various severities of knee osteoarthritis *J. Mech. Med. Biol.* **16** 1650003
Shahabpoor E, Pavic A, Brownjohn J M W, Billings S A, Guo L Z and Bocian M 2018 Real-life measurement of tri-axial walking ground reaction forces using optimal network of wearable inertial measurement units *IEEE Trans. Neural Syst. Rehabil. Eng.* **26** 1243–53
Sharifmoradi K, Farahpour N, Bahram A, Karimi M T and Mazdeh M 2015 An assessment of gait spatiotemporal and ground reaction force characteristics of patients with Parkinson compared with normal elderly *J. Res. Rehabil. Sci.* **10** 676–86
Sharma D, Davidson P, Muller P and Piche R 2021 Indirect estimation of vertical ground reaction force from a body-mounted INS/GPS using machine learning *Sensors* **21** 1553
Sinitski E H, Lemaire E D and Baddour N 2015 Evaluation of motion platform embedded with dual belt treadmill instrumented with two force plates *J. Rehabil. Res. Dev.* **52** 221–34
Stuchbury-Wass J *et al* 2026 Ear-worn inertial sensors can predict gait metrics and reconstruct vertical ground reaction force curves during running *J. Appl. Biomech.* **25**
Sun T, Li D, Fan B, Tan T and Shull P B 2023 Real-time ground reaction force and knee extension moment estimation during drop landings via modular LSTM modeling and wearable IMUs *IEEE J. Biomed. Health Inform.* **27** 3222–33
Ullah Z *et al* 2026 Pathology-informed augmentation improves cross-cohort IMU-to-vGRF estimation between healthy adults and adults with osteoarthritis *IEEE Trans. Neural Syst. Rehabil. Eng.* **34** 3164–77
Van Der Velde G, Laloyaux H and Ronsse R 2025 Inducing asymmetric gait in healthy walkers: a review *Front. Rehabil. Sci.* **6** 1463382
Wang J *et al* 2022 Generalizing to unseen domains: a survey on domain generalization *IEEE Trans. Knowl Data Eng.* **35** 8052–72
Williams S E, Gibbs S, Meadows C B and Abboud R J 2011 Classification of the reduced vertical component of the ground reaction force in late stance in cerebral palsy gait *Gait Posture* **34** 370–3
Winter D A 2009 Kinetics: forces and moments of force *Biomechanics and Motor Control of Human Movement* 4th edn (Wiley) ch 5 pp 107–38
Wouda F J, Giuberti M, Bellusci G, Maartens E, Reenalda J, van Beijnum B-J F and Veltink P H 2018 Estimation of vertical ground reaction forces and sagittal knee kinematics during running using three inertial sensors *Front. Physiol.* **9** 218
Wu C-C, Wen Y-T and Lee Y-J 2020 IMU sensors beneath walking surface for ground reaction force prediction in gait *IEEE Sens. J.* **20** 9372–8
Yılmazgün B, Weber J, Stein T, Sell S and Stetter B J 2025 Predicting 3D ground reaction forces across various movement tasks: a convolutional neural network study comparing different inertial measurement unit configurations *J. Biomech.* **192** 112888
Zanardi A P J, da Silva E S, Costa R R, Passos-monteiro E, Dos Santos I O, Kruel L F M and Peyré-tartaruga L A 2021 Gait parameters of Parkinson's disease compared with healthy controls: a systematic review and meta-analysis *Sci. Rep.* **11** 752