Kinetic Instrumentation of Sit-to-Walk Component of Timed Up and Go Test

Malik Naveed Akhter

19092748

School of Allied Health, Faculty of Health and Environmental Sciences

A thesis submitted to Auckland University of Technology

In fulfillment of the requirements for the degree of

Doctor of Philosophy

2026

Supervisors

Prof. Denise Taylor, Dr. Imran Khan Niazi

Abstract

Functional mobility assessment is crucial for identifying fall risk and guiding rehabilitation in older adults. The Timed Up and Go (TUG) test is a widely used clinical measure of mobility, but its primary metric, total time to complete, only offers a basic overview of the complex movement process. Wearable inertial measurement units (IMUs) used in instrumented versions of the TUG provide valuable kinematic and spatiotemporal data. However, kinetic variables such as ground reaction forces (GRF), ground reaction moments (GRM), and center of pressure (CoP) are mostly measurable only in laboratory settings because they require fixed force plates. These kinetic measures provide important clinical insights into lower-limb loading, postural control, and balance strategies that kinematics alone cannot reveal. Therefore, developing methods to estimate kinetic variables from wearable sensors during the TUG could enhance the clinical usefulness of this assessment without requiring specialized lab equipment. This thesis aims to develop and evaluate machine learning techniques to estimate these kinetic parameters during the initial transition of the TUG, the sit-to-walk (STW).

A PRISMA-compliant systematic review was conducted to outline the current landscape of machine learning and deep learning methods for GRF estimation from IMU data. The review included 22 studies, examining participant groups, sensor setups, preprocessing methods, model types, and evaluation approaches. Although individual studies compared selected architectures, differences in normalization, validation protocols and performance metrics limited meaningful comparisons across studies.

To address these gaps, an experimental study tested 13 model configurations to predict continuous three-dimensional GRF waveforms during STW from a single sacrum-mounted IMU, using a publicly available dataset of 65 participants. All models were evaluated with a consistent pipeline. Results showed that no single architecture dominated across all three GRF axes. However, the DS-Conformer was the most consistently high-performing, achieving the best vertical GRF prediction (RMSE 4.05 %BW, R^2 = 0.985) and the highest waveform agreement on the anteroposterior and mediolateral axes.

A follow-up study expanded the DS-Conformer to estimate all eight kinetic components during STW: three GRF, three GRM, and two CoP channels. Both experimental studies used the same dataset. Testing on held-out participants and age groups supports prediction for new people

within that dataset, but does not establish performance with a different sensor, laboratory or acquisition setup. Leave-one-out cross-validation across 65 subjects showed normalized RMSEs from 3.78% to 11.48% and Pearson correlations between 0.787 and 0.990. Predictions for GRF and GRM were consistent across ages, but CoP accuracy varied significantly between age groups in the mediolateral direction, highlighting the limited ability of a single midline sensor to detect inter-limb lateral weight-shifting strategies that differ among individuals and age groups.

This thesis demonstrates the potential for wearable kinetic assessment during the STW transition of the TUG test using a single body-worn IMU and deep learning. The proposed framework is a step towards providing mechanistic kinetic information during routine TUG assessments that time alone cannot, without requiring laboratory infrastructure. However, several questions remain open. Validation in clinical populations with asymmetric or impaired movement has not yet been undertaken, and normative reference ranges and outcome-linked thresholds for STW kinetics must be established before predicted waveforms can inform clinical decisions. The mediolateral bottleneck also warrants targeted work on sensor expansion or on the fusion of sensing modalities to recover lateral balance information that a single midline sensor cannot capture. The framework currently covers only the first transition of TUG. Future work should extend the framework to the remaining TUG subtasks.

Table of Contents

ABSTRACT ..... **II**

LIST OF FIGURES ..... **VII**

LIST OF TABLES ..... **IX**

LIST OF APPENDICES ..... **X**

ATTESTATION OF AUTHORSHIP ..... **XI**

ARTIFICIAL INTELLIGENCE DECLARATION ..... **XII**

CO-AUTHOR CONTRIBUTIONS ..... **XIII**

ACKNOWLEDGEMENTS ..... **XVI**

LIST OF ABBREVIATIONS ..... **XVII**

CHAPTER 1. INTRODUCTION **19**

1.1 CONTEXT 19

1.2 MOTIVATION 26

1.3 RESEARCH OBJECTIVE 27

1.4 THESIS STRUCTURE 29

CHAPTER 2. APPLICATION OF MACHINE LEARNING TECHNIQUES IN THE MEASUREMENT OF GROUND REACTION FORCE (GRF) WAVEFORMS USING INERTIAL MEASUREMENT UNIT (IMU) SENSORS - A SYSTEMATIC LITERATURE REVIEW **32**

2.1 PREFACE 32

2.2 ABSTRACT 33

2.3 INTRODUCTION 34

2.3.1 Main Contributions *38*

2.4 REVIEW METHODOLOGY 39

2.4.1 Search Strategy and Selection Criteria *39*

2.4.2 Eligibility Criteria *40*

2.4.3 Quality Assessment *40*

2.4.4 Data Extraction and Synthesis *41*

2.5 RESULTS 41

2.5.1 RQ1 Participant populations and experimental protocols *42*

2.5.2 RQ2 - Preprocessing and feature extraction *46*

2.5.3 RQ3 - Machine and Deep Learning models and performance by GRF component *48*

2.5.4 RQ4 - Evaluation metrics and validation protocols *70*

2.6 DISCUSSION 72

2.6.1 Participant populations and experimental protocols *73*

2.6.2 Preprocessing and feature extraction *74*

2.6.3 Machine-learning and deep-learning model comparison *76*

2.6.4 Evaluation metrics and validation protocols *77*

2.6.5 Deployment challenges and future directions *79*

2.7 CONCLUSION 80
2.8 CHAPTER SUMMARY 81
CHAPTER 3. PREDICTING GROUND REACTION FORCE DURING SIT TO WALK: TOWARDS KINETICALLY INSTRUMENTING TIMED UP AND GO TEST 83
3.1 PREFACE 83
3.2 ABSTRACT 84
3.3 INTRODUCTION 85
3.4 MATERIALS AND METHODS 87
3.4.1 Dataset 87
3.4.2 Preprocessing 88
3.4.3 Model Architectures 88
3.4.4 Loss Functions 95
3.4.5 Validation Protocol 95
3.4.6 Age-Group Generalization Experiment 97
3.5 RESULTS 98
3.5.1 5-Fold Cross-Validated Waveform Prediction 98
3.5.2 Leave-One-Out Cross-Validation Results 102
3.5.3 Age-Group Generalization 104
3.6 DISCUSSION 104
3.6.1 Axis-Dependent Architectural Preferences 105
3.6.2 Architectural Component Analysis 106
3.6.3 Comparison with Prior IMU-to-GRF Studies 108
3.6.4 Generalizability 108
3.6.5 Limitations 110
3.7 CONCLUSION 110
3.8 CHAPTER SUMMARY 112
CHAPTER 4. DEEP LEARNING-BASED KINETICS AND CENTER OF PRESSURE ESTIMATION DURING SIT-TO-WALK TRANSITIONS FROM A SINGLE SACRUM-MOUNTED IMU 115
4.1 PREFACE 115
4.2 ABSTRACT 116
4.3 INTRODUCTION 116
4.4 MATERIALS AND METHODS 118
4.4.1 Dataset 118
4.4.2 Data Processing 120
4.4.3 Center of Pressure Computation 120
4.4.4 Model Architecture 122
4.4.5 Training Strategy 124
4.4.6 Model Evaluation 124
4.5 RESULTS 126

4.5.1 Performance Across Age Groups *129*
4.5.2 Comparison with Existing Methods *131*
4.6 DISCUSSION 132
4.6.1 Limitations and Future Work *134*
4.7 CONCLUSION 134
4.8 CHAPTER SUMMARY 135
CHAPTER 5. DISCUSSION AND CONCLUSION 138
5.1 PROLOGUE 138
5.2 OVERVIEW OF FINDINGS 138
5.3 WHAT THIS THESIS ADDS 143
5.4 IMPLICATIONS FOR CLINICAL PRACTICE 144
5.5 STRENGTHS OF THIS THESIS 146
5.6 LIMITATIONS OF THIS THESIS 147
5.7 IMPLICATIONS FOR FUTURE WORK 149
5.8 CONCLUSION 150
REFERENCES 152
APPENDICES 167

List of Figures

Figure 1-1 Overview of the Timed Up and Go (TUG) test framework and its kinetic instrumentation. (Top left) The four sequential stages of the forward portion of the TUG test. (Bottom left) Experimental setup illustrating the sit-to-walk transition, which is the first transition of the TUG test. (Center) The three kinetic outputs recorded by the force plate. (Right) Clinical applications of these kinetic measures. 22

Figure 1-2 Thesis Structure with objectives 31

Figure 2-1 Machine learning and deep learning model training pipeline 37

Figure 2-2 PRISMA Flow Diagram 42

Figure 2-3 Sensor placements across the 22 reviewed studies. (a) Filled cells indicate reported anatomical placements; column totals count studies using each location. (b) The co-occurrence matrix counts studies using each pair of locations; diagonal entries give the location totals. Bilateral or multiple sensors at one location are counted once per study. Wu et al. (2020) used sensors beneath the walking surface, and Kim et al. (2020) used marker-derived sacral/CoM input; these studies are retained as rows but are not counted as anatomical IMU placements. Frequency of use does not indicate predictive accuracy. 46

Figure 3-1 Cumulative FFT power for GRF used to determine low-pass filter cutoff frequency. The cumulative percentage of total signal power is plotted against frequency. 88

Figure 3-2 Architectural progression for 3D GRF prediction from a single representative fold. Each row represents a stage in the investigation, showing representative predicted (orange) versus true (blue) waveforms for vertical, AP, and ML GRF. Annotations between stages indicate the observed limitation motivating each transition. Border colors denote per-axis R^2 performance: green (≥ 0.95), blue (0.80-0.95), grey (0.50-0.80), and red (< 0.50). 90

Figure 3-3 Model Rank flow across GRF axes based on RMSE. Rank lines cross extensively between axes, with DS-Conformer and ResCorr (joint) being the only models remaining in the top three on at least two axes. 102

Figure 3-4 Per-subject concordance correlation coefficient (CCC) distributions for the top four models on each GRF axis under the age-group holdout protocol. Box plots show median line, interquartile range (box), and 1.5×IQR whiskers across subjects. 104

Figure 3-5 Graphical Summary of Chapter 3. 114

Figure 4-1 Sagittal-view representation of the experimental setup. Skeleton color indicates movement phase. Right leg segments are shown as bold lines. Red dots denote the sacrum-mounted IMU location at each pose. Three serially aligned BERTEC force plates (600 × 400 mm each) are shown at floor level. The participant began seated on a backless, height-adjustable stool with both feet on Force plate 1, then performed the STW transition progressing across all three force plates. 119

Figure 4-2 Cumulative FFT power for GRF and GRM channels used to determine low-pass filter cutoff frequency. For each axis, the cumulative percentage of total signal power is plotted against frequency. 120

Figure 4-3 DS-Conformer neural network architecture. (a) Overall model architecture comprising a skip-connection branch, a strided downsampling encoder, a bottleneck, and a transposed-convolution upsampling decoder. (b) Detail of a single Conformer block:

feedforward half-step, multi-head self-attention, depth-wise convolution, and feedforward half-step, each wrapped in residual connections with layer normalization. 122

Figure 4-4 Mean predicted (red) and ground truth (blue) waveforms across all 325 trials (65 subjects × 5 windows) from leave-one-out cross-validation using the DS-Conformer model. Solid lines represent mean, shaded regions represent ± 1 SD 126

Figure 4-5 Representative single-trial prediction, showing ground truth (blue) and predicted (red) waveforms. RMSE and Pearson r are reported for each component. 127

Figure 4-6 CoP prediction from IMU data using the DS-Conformer model. (a) Mean CoP trajectory ± 1 SD across all 325 trials, showing close agreement between ground truth (blue) and predicted (red) paths with largely overlapping variability bands. (b) Representative single-trial CoP trajectory demonstrating high prediction accuracy for both AP (r = 0.998) and ML (r = 0.977) components. CoP coordinates are expressed in meters. 127

Figure 4-7 Age-group generalization performance (Pearson r) of the DS-Conformer model across all eight kinetic output components, organized by biomechanical axis. Each panel displays per-subject distributions for GRF, GRM, and CoP predictions under three leave-one-age-group-out holdout scenarios: train on Young+Mid, test on Elder (YM→E, n=18); train on Mid+Elder, test on Young (ME→Y, n=32); and train on Young+Elder, test on Mid (YE→M, n=15). Dashed and dotted horizontal lines indicate r = 0.9 and r = 0.8, respectively. The vertical axis has no associated CoP component. CoP(ML) exhibits the lowest prediction accuracy across all splits, while CoP(AP), GRM(Sagittal), and GRF(V) consistently achieve r > 0.95. 131

Figure 4-8 Graphical Summary of Chapter 4 137

List of Tables

Table 2-1 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).....45

Table 2-2 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. 50

Table 3-1 Five-fold subject-wise cross-validation vertical GRF results for 13 model configurations 99

Table 3-2 Five-fold subject-wise cross-validation AP GRF results for 13 model configurations. 100

Table 3-3 Five-fold subject-wise cross-validation ML GRF results for 13 model configurations. 101

Table 3-4 Leave-one-out cross-validation results for the top two performing models for each GRF axis. 103

Table 3-5 Comparison of IMU-to-GRF prediction accuracy across studies, movement tasks, and sensor configurations. Results for the present study (Ours) are reported under leave-one-out cross-validation to ensure subject-independent evaluation consistent with other studies. A dash (-) indicates no metric was reported for that axis. "r" is Pearson Correlation Coefficient. 108

Table 4-1 Leave-One-Out Cross-Validation Results for GRF, GRM, and CoP Prediction (N = 65) 128

Table 4-2 Normalized RMSE (%) by Age Group for All Output Components (N = 65).....130

Table 4-3 Comparison of reported GRF, GRM and CoP prediction errors across studies....132

Table 5-1 Summary of research questions and main findings by chapter. 142

List of Appendices

APPENDIX A. SUPPLEMENTARY MATERIAL TO CHAPTER 2 167

A1. SYSTEMATIC SEARCH STRATEGY AND YIELD BY DATABASE 167

A2. QUALITY ASSESSMENT QUESTIONS 168

A3. PARTICIPANT CHARACTERISTICS OF INCLUDED STUDIES 170

A4. SENSOR CHARACTERISTICS OF INCLUDED STUDIES 172

A5. PREPROCESSING AND FEATURE EXTRACTION METHODS USED IN THE INCLUDED STUDIES 177

A6. DETAILS OF MACHINE/DEEP MODELS USED IN THE INCLUDED STUDIES 189

A7. TASKS PERFORMED, VALIDATION STRATEGIES, AND GRF WAVEFORM EVALUATION METRICS USED IN THE INCLUDED STUDIES 199

APPENDIX B. SUPPLEMENTARY MATERIAL TO CHAPTER 3 205

B1. PER-FOLD STABILITY ANALYSIS 205

B2. MEDIAN AND INTERQUARTILE RANGE SUMMARY OF GRF PREDICTION PERFORMANCE 207

B3. STATISTICAL COMPARISONS BETWEEN THE MODELS 211

B4. BIOMECHANICAL EVENT METRICS 213

B5. ROBUSTNESS EXPERIMENTS 215

B6. MODEL SIZE AND INFERENCE COSTS 219

APPENDIX C. SUPPLEMENTARY MATERIAL TO CHAPTER 4 222

C1. MEDIAN AND INTERQUARTILE RANGE SUMMARY 222

C2. MODEL SIZE AND INFERENCE COSTS 223

C3. EXPLORATORY GRF ESTIMATION ON EXTERNAL DATASET 224

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor used artificial intelligence tools or generative artificial intelligence tools (unless it is clearly stated, and referenced, along with the purpose of use), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

_____

Malik Naveed Akhter
19092748

Artificial Intelligence Declaration

Specify the chapter number(s) in your research proposal/research component where AI was used.	
Chapter number(s):	**Chapter 1-5**
Briefly explain why the AI tool was used (e.g., idea generation, summarization, data analysis, image creation, etc.).	
Purpose of AI Use:	ChatGPT was used for grammar correction and language refinement. Its use was limited to improving spelling, punctuation, sentence structure, clarity, and readability of text written by the author. It was not used to generate original research ideas, analyse data, interpret results, or create figures.
Clearly state the AI tools used.	
AI Tool(s) Used:	**ChatGPT (OpenAI)**
Prompts or queries entered into the AI tool (only the instructive portion of the prompt is required). Include all relevant variations.	
Prompts Used:	**Check this chapter for grammar, punctuation, and readability.** **Edit this paragraph for grammar, clarity, and flow.** **Improve sentence structure and remove repetition.**
A summary or description of the AI-generated output.	
Output Received:	**The AI tool returned edited text with suggested corrections to grammar, spelling, punctuation, sentence structure, and overall phrasing. In some cases, it also suggested clearer and more concise wording while preserving the original meaning.**
Describe how the AI output was used, edited, or transformed.	
Post-AI Processing Methods:	**All AI-generated suggestions were manually reviewed by the author before use. The author checked the edited text to ensure that the meaning, interpretation, argument, and technical accuracy remained unchanged. Suggestions were selectively accepted, modified, or rejected as appropriate. Final responsibility for the wording, content, and accuracy of the thesis remained with the author.**

STUDENT DECLARATION			
By signing you are confirming that the AI use stated in the table(s) above are accurate and follow AUT's recommended guidelines detailed in the AI Hub and Postgraduate Handbook.			
Student Name:	**Malik Naveed Akhter**	**Signature:**	**Date:** 23/04/2026

Co-Author Contributions

STUDENT AND SUPERVISOR APPROVALS

By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.

Student Name	Malik Naveed Akhter	**Signature**		**Date**	23/04/2026
Supervisor Name	Prof. Denise Taylor	**Signature**		**Date**	24/4/2026
Secondary Supervisor Name	Dr. Imran Khan Niazi	**Signature**		**Date**	23/04/2026

Chapter Number:	**2**
Manuscript Title:	**Application of Machine Learning Techniques in the Measurement of Ground Reaction Force waveforms using IMU Sensors**
Publication Status:	**Submitted for Publication**
Reference if published:	

AUTHOR SURNAME: (order as per manuscript)	**CONTRIBUTION** (May copy from the guidelines above)
Akhter	**Concept and design of the study, Data Analysis and interpretation, and Writing**
Niazi	**Concept and design of the study, Reviewing and editing**
Ali	**Validation**
Manj	**Validation**
Taylor	**Concept and design of the study, Reviewing and editing**

STUDENT AND SUPERVISOR APPROVALS

By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.

Student Name	Malik Naveed Akhter	**Signature**		**Date**	23/04/2026
Supervisor Name	Prof. Denise Taylor	**Signature**		**Date**	24/4/2026
Secondary Supervisor Name	Dr. Imran Khan Niazi	**Signature**		**Date**	23/04/2026

Chapter Number:	**3**
Manuscript Title:	**Predicting Ground Reaction Force during Sit to Walk: Towards kinetically instrumenting Timed Up and Go Test**
Publication Status:	**Submitted for Publication**
Reference if published:	

AUTHOR SURNAME: (order as per manuscript)	**CONTRIBUTION** (May copy from the guidelines above)
Akhter	**Concept and design of the study, Data Analysis and interpretation, and Writing**
Niazi	**Concept and design of the study, Reviewing and editing**
Ali	**Validation**
Manj	**Validation**
Taylor	**Concept and design of the study, Reviewing and editing**

STUDENT AND SUPERVISOR APPROVALS

By signing you are confirming that the co-author contributions stated in the table(s) below are accurate.

Student Name	Malik Naveed Akhter	**Signature**		**Date**	23/04/2026
Supervisor Name	Prof. Denise Taylor	**Signature**		**Date**	24/4/2026
Secondary Supervisor Name	Dr. Imran Khan Niazi	**Signature**		**Date**	23/04/2026

Chapter Number:	**4**
Manuscript Title:	**Deep Learning-Based Kinetics and Center of Pressure Estimation During Sit-to-Walk Transitions from a Single Sacrum-Mounted IMU**
Publication Status:	**Submitted for Publication**
Reference if published:	

AUTHOR SURNAME: (order as per manuscript)	**CONTRIBUTION** (May copy from the guidelines above)
Akhter	**Concept and design of the study, Data Analysis and interpretation, and Writing**
Niazi	**Concept and design of the study, Reviewing and editing**
Ali	**Validation**
Manj	**Validation**
Taylor	**Concept and design of the study, Reviewing and editing**

Acknowledgements

This thesis marks the end of a long and exciting journey, and it would not have been possible without the guidance and support of many people.

First and foremost, I want to express my gratitude to my supervisors, Prof. Denise Taylor and Dr. Imran Khan Niazi, for their guidance throughout my doctoral journey. Their clinical and technical expertise, along with their careful attention to detail influenced every phase of this thesis.

I would also like to acknowledge the New Zealand College of Chiropractic for funding this PhD.

My thanks go to Haider Ali, Usman Sheikh, Usman Ghani, and Umair Manj for many spirited discussions along the way. Some of the clearest ideas in this thesis were sharpened in their company.

Finally, to my mum, thank you for your unwavering love and belief in me. I also honor the memory of my late father, whose guidance continues to shape me in his absence. So much of what this thesis required, including patience, persistence, and the quiet faith that difficult problems do eventually yield, I learned first from him.

With gratitude.

List of Abbreviations

ANN	Artificial Neural Network
AP	Anteroposterior
BiGRU	Bidirectional Gated Recurrent Unit
BiLSTM	Bidirectional Long Short-Term Memory
BW	Bodyweight
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
CoM	Center of Mass
CoP	Center of Pressure
CV	Cross-Validation
1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
DLR	Double Loss Regression
DS	Downsampled
DWT	Discrete Wavelet Transform
EMG	Electromyography
ESN	Echo State Network
FC	Fully Connected
FFN	Feedforward Network
FFNN	Feedforward Neural Network
FFT	Fast Fourier Transform
FM	Fusion Module
GRF	Ground Reaction Force
GRM	Ground Reaction Moment
GRU	Gated Recurrent Unit
IMU	Inertial Measurement Unit
iOFR	Iterative Orthogonal Forward Regression
IQR	Interquartile Range
iTUG	Instrumented Timed Up and Go
KNN	k-Nearest Neighbors
LOOCV	Leave-One-Out Cross-Validation
LOSO	Leave-One-Subject-Out
LSTM	Long Short-Term Memory
MAD	Mean Absolute Deviation
MAE	Mean Absolute Error
MHSA	Multi-Head Self-Attention
MKOA	Medial Knee Osteoarthritis
ML	Mediolateral
MLP	Multilayer Perceptron
MRU	Multi-Resolution Unit
MSE	Mean Squared Error
NARMAX	Nonlinear Auto-Regressive Moving Average with Exogenous Input
NARX	Nonlinear Auto-Regressive with Exogenous Input
NMAD	Normalized Mean Absolute Deviation
NN	Neural Network
NRMSE	Normalized Root Mean Square Error

PCA	Principal Component Analysis
r	Pearson Correlation Coefficient
PD	Parkinson's Disease
PICO	Participants, Intervention, Comparisons, and Outcomes
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
R^2	Coefficient of Determination
ReLU	Rectified Linear Unit
RMS	Root Mean Square
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
rRMSE	Relative Root Mean Square Error
SD	Standard Deviation
SGD	Stochastic Gradient Descent
SLIP	Spring-Loaded Inverted Pendulum
SLR	Systematic Literature Review
STW	Sit-to-Walk
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
TUG	Timed Up and Go

Chapter 1. Introduction

1.1 Context

Functional mobility, the ability to stand up, walk, turn, and sit down safely and efficiently, is a prerequisite for independent living and participation in everyday activities. In older adults, mobility impairment and functional decline are associated with adverse outcomes, including reduced quality of life and increased risk of hospitalization (Freiberger et al., 2020). One of the most clinically relevant outcomes of impaired mobility is an increased risk of falls. Falls have been identified as a significant contributor to injury, mortality, and disability, particularly among individuals aged 65 years and above (Montero-Odasso et al., 2022).

In New Zealand, population aging is expected to amplify the burden of fall-related injury and disability. With over 950,000 people currently aged 65 and over in New Zealand, the older adult cohort represents 16.6% of the total population and is projected to reach 1.55 million by 2051 (Stats NZ-Tatauranga Aotearoa, 2024), accounting for 25% of the total population. Approximately 30% of older adults report experiencing at least one fall annually (Pillay et al., 2024). Falls can lead to severe injuries and prolonged recovery, and some individuals do not fully regain functional capacity, increasing the likelihood of requiring residential care (Hartholt et al., 2011). Rates of hospitalization for fall-related injuries among adults aged 65 and older have increased significantly over the past two decades, with hip fracture and head injury among the most serious outcomes in older adults (Bradley & Harrison, 2011; Cummings & Melton, 2002). In New Zealand, fall-related hospital admissions between 2012 and 2022 are estimated to have cost NZ$9,792 per event, with an average length of stay of 5.5 days (Simpkins et al., 2024). Over this 11-year period, the Midland trauma region of New Zealand alone recorded 13,142 fall-related injury admissions (Simpkins et al., 2024), highlighting a substantial economic and health burden.

Many mobility limitations that elevate fall risk reflect underlying deficits in postural control and balance during everyday tasks (Lektip et al., 2024; Robinovitch et al., 2013; Schrager et al., 2008). Clinical assessment of mobility-related fall risk commonly involves tests that directly or indirectly evaluate balance. The primary objectives of balance assessment are to detect whether an impairment exists and to identify its underlying cause (Goethel et al., 2025;

Mancini & Horak, 2010; Noamani et al., 2023). Mancini and Horak (2010) divided these tests into three categories: Functional, Systems, and Objective assessment tests. Functional tests evaluate stability during tasks that resemble everyday activities. They are quick to administer and do not require any specialized equipment. However, the outcomes of these tests are observer-dependent, and the results generally provide limited insight into the physiological mechanisms underlying the balance deficit. Systems tests help localize the source of impairment by separately probing sensory inputs (vestibular, visual, and somatosensory), motor responses, and central integration. Yet, their technical complexity makes them time-consuming, reducing feasibility in routine clinical workflows. Objective tests, such as force-platform posturography, quantify balance through objective measures such as the magnitude of center-of-pressure (CoP) sway during static or dynamic tasks, but their reliance on specialized equipment, staff training, and space limits scalability and broad clinical uptake. In summary, each category of balance test involves a trade-off between clinical practicality and diagnostic value. Among the three types of assessments, functional assessments are the fastest to administer and place the least demand on clinician time, making them the most practical foundation on which to build.

Among functional assessments, the Timed Up and Go (TUG) is one of the most widely used tests in older adult care and rehabilitation (Park, 2018). TUG integrates multiple mobility subtasks within a single protocol, including standing up from a chair, walking 3 meters, turning around, walking back, and sitting down again (Podsiadlo & Richardson, 1991). It is widely used in clinical settings because it requires minimal equipment and can be administered rapidly, typically within a few minutes, including instructions and safety setup (Weiss et al., 2010). The primary assessment metric is the time it takes the participant to complete it. The TUG test has demonstrated good test-retest and inter-rater reliability across a range of populations and is widely used as a functional mobility and fall-risk screening tool (Barry et al., 2014; Park, 2018; Schoene et al., 2013; Shumway-Cook et al., 2000). However, total time is a coarse summary of a multi-component task and can mask clinically relevant differences in movement strategy. Two individuals can achieve the same completion time while using very different movement strategies. One may rise quickly using high trunk flexion and aggressive momentum transfer, then pause before initiating gait, whereas another may rise slowly and cautiously with armrest support and transition smoothly into walking, producing similar total times through very different strategies. Consequently, the single time score may fail to distinguish impairment mechanisms or sensitively track change over time, particularly when a faster completion time

results from altered movement strategies, such as generating greater trunk momentum to assist rising, relying on arm support for stability, or widening the base of support during turning, rather than recovery of balance capacity (Chandler et al., 2021). Moreover, systematic review and meta-analysis evidence has concluded that TUG time has limited ability to predict future falls in community-dwelling older adults and should not be used in isolation for fall risk identification (Barry et al., 2014; Schoene et al., 2013). These limitations have motivated the development of instrumented approaches that preserve the clinical practicality of the TUG while extracting more information about movement quality.

The instrumented Timed Up and Go (iTUG) refers to a variant of the TUG protocol in which wearable or environmental sensors are used to quantify kinematics and spatiotemporal parameters during each subtask (Ortega-Bastidas et al., 2023). Common sensing modalities include inertial measurement units (IMUs) (Alcantara et al., 2022), motion capture systems (Faisal et al., 2023), pressure sensors (Carter et al., 2024), and instrumented walkways (Barela et al., 2014). Among these, IMUs have emerged as a pragmatic choice because they are low-cost, portable, and embedded in consumer wearables. IMU-based iTUG systems typically place sensors on the lower back, shanks, feet, and thighs. From these signals, algorithms can compute features such as durations of each TUG subtask, step counts, gait speed, turning angular velocity, and stride-to-stride variability metrics (Ortega-Bastidas et al., 2023). These features can uncover deficits masked by the single-time score and support rehabilitation monitoring and clinical decision-making.

Kinematics describes motion, such as joint angles, velocity and acceleration. Kinetics describes the forces and moments associated with that motion. For example, trunk acceleration describes how quickly the trunk's velocity changes, whereas vertical GRF describes the upward force exerted by the ground on the body. To date, most iTUG work predominantly focuses on kinematic and temporal features (Ortega-Bastidas et al., 2023). While these provide valuable information about movement execution, they do not directly quantify the forces and moments exchanged between the body and ground. As illustrated in Figure 1-1, three kinetic measures are central to understanding balance during functional tasks: ground reaction forces (GRFs), ground reaction moments (GRMs), and Center of Pressure (CoP). GRFs represent the net external force vector applied by the ground to the body, which is typically resolved into vertical (V), anteroposterior (AP), and mediolateral (ML) components. GRMs represent the net moment vector about a reference point on the ground, capturing the rotational effects of the

ground reaction. The CoP represents the location on the support surface where the resultant GRF vector acts and provides an intuitive measure of the control of the base of support and distribution of plantar loading.

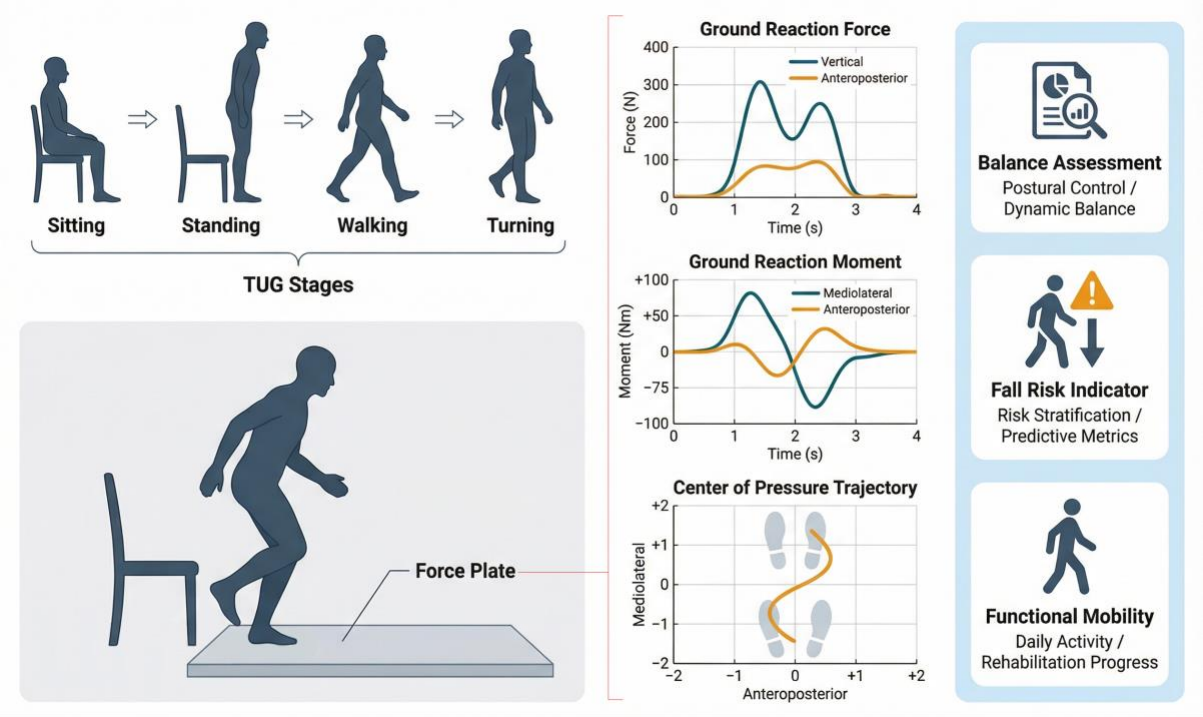


Figure 1-1 Overview of the Timed Up and Go (TUG) test framework and its kinetic instrumentation. (Top left) The four sequential stages of the forward portion of the TUG test. (Bottom left) Experimental setup illustrating the sit-to-walk transition, which is the first transition of the TUG test. (Center) The three kinetic outputs recorded by the force plate. (Right) Clinical applications of these kinetic measures.

The mechanistic value of kinetics is especially pronounced for transitional movements, such as sit-to-walk (STW). The STW transition involves standing up and initiating gait without pausing between the two components. This transition entails rapid changes in the base of support, momentum generation, and balance constraints. The movement requires coordinated trunk flexion, lower-limb extension, and timely foot unloading as the center of mass transitions from seated support to a gait pattern (Buckley et al., 2009; Magnan et al., 1996). Kinetic signals make these phase transitions explicit in ways that kinematics alone cannot, because the boundaries between phases are defined by changes in load transfer, propulsion, and base-of-support control rather than by large visible changes in joint angles (Kerr et al., 2004). The rate of vertical GRF rise during seat unloading directly reflects how rapidly load is accepted by the

lower limbs (Buckley et al., 2009; Miller et al., 2024). AP GRF captures the balance between the forward propulsion needed to initiate gait and the braking force that prevents forward loss of balance (Magnan et al., 1996). The CoP trajectory during step initiation indexes the coordination of lateral weight transfer and anticipatory postural adjustments that precede swing-limb unloading (Bestaven et al., 2013). Kinematics describe the visible motion pattern, but kinetics reveal when support is transferred to the feet, when forward momentum is generated or braked, and how balance is controlled as the body transitions toward single-limb support (Janssen et al., 2002).

The STW transition warrants special attention because it is common in daily life and challenging to execute safely, particularly with declining mobility (Perera et al., 2023). Unlike sit-to-stand, which ends in quiet standing, STW requires simultaneous stabilization and forward progression. Biomechanically, STW can be described as a sequence of phases: (i) flexion momentum, during which trunk flexion generates forward momentum, (ii) momentum transfer, during which support shifts from the seat to the feet, (iii) extension, during which lower-limb extension elevates the Center of Mass, and (iv) gait initiation, during which CoP shifts and one foot unloads to initiate the first step. These phases have distinct kinetic signatures (Perera et al., 2023). For example, vertical GRF rises sharply during flexion momentum, AP GRF reflects forward propulsion and braking, and CoP typically first moves toward the limb about to be unloaded, helping generate the lateral momentum required for step initiation, then rapidly shifts toward the stance limb as the other foot is lifted. In impaired populations, these signatures can change. Individuals might produce lower peak vertical GRF rates, which may indicate reduced power, show prolonged weight transfer, which could reflect a cautious strategy, or demonstrate increased ML GRF variability, which may suggest balance challenges, or show left–right GRF asymmetry, which may reflect unilateral weakness or pain in a lower limb, hip, or back. Kinetic deficits during STW have been associated with reduced lower limb power, impaired balance, and altered strategies in older adults and clinical populations (Buckley et al., 2009). Because STW is embedded within the TUG, accurate kinetic characterization of STW has direct implications for improving the sensitivity of iTUG for fall-risk screening and rehabilitation monitoring. This motivates a focus on STW as a core subtask for kinetic instrumentation within the iTUG.

Despite the clinical and scientific value of GRF, GRM, and CoP, these measures are typically captured using laboratory-grade force plates or instrumented treadmills. Such equipment is

costly, requires specialized installation and calibration, and often constrains testing to laboratory environments. Even where force plates are available, routine clinical use may be limited by space, time, workflow constraints, and staff training requirements. There are also protocol-related limitations. Participants must contact the plates accurately, which can alter natural behavior, particularly during turning or among individuals with mobility impairment (Huang et al., 2025). Capturing multiple steps often requires multiple plates or long instrumented walkways. Kinetic analysis is also frequently paired with optical motion capture to estimate joint kinematics and moments; however, marker-based systems require time for marker placement and calibration, are sensitive to the precision of anatomical marker placement, and are affected by soft-tissue artifacts, whereby skin-mounted markers move relative to the underlying bone (Della Croce et al., 2005; Leardini et al., 2005). Collectively, these barriers limit the translation of kinetic biomechanics into widespread clinical practice and motivate approaches to estimate kinetic variables from wearable sensors outside the laboratory.

Estimating external kinetics from wearable inertial sensing is fundamentally an inverse problem. IMUs measure linear acceleration and angular velocity at the sensor locations, which reflect segmental motion but do not directly encode net external forces. Physical modeling approaches can estimate GRFs from kinematics using inverse dynamics, subject to assumptions about segment inertial parameters, joint constraints, and contact models. However, in unconstrained clinical environments, these assumptions are often violated (Karatsidis et al., 2016). Machine learning provides an alternative approach. Rather than explicitly modeling body segments and contact mechanics, machine learning models can learn a mapping from IMU time series to kinetic outputs when trained on synchronized data. This learning-based strategy can capture complex nonlinear relationships and generalize across movement strategies if trained appropriately (Ancillao et al., 2018; Wouda et al., 2018).

Machine learning methods span feature-based regression and end-to-end sequence modeling. In feature-based approaches, IMU signals are segmented into windows and transformed into engineered features like mean, variance, spectral power, peak counts, and event timings, which are then mapped to scalar targets using models such as linear regression, random forests, and support vector machines. These approaches can be interpretable and data-efficient but may discard temporal structures and may not support trajectory-level prediction. In end-to-end approaches, models consume raw or minimally processed time-series data and produce an entire GRF trajectory, learning relevant temporal features internally (Donahue & Hahn,

2023b). Prior work on GRF estimation from wearables spans a wide spectrum of tasks, including walking (Guo et al., 2017), running (Donahue & Hahn, 2023a), jumping (Kerns et al., 2023), stair climbing (Fukushi et al., 2019), and a wide spectrum of modeling families, including, random forests (Jiang et al., 2020), shallow neural networks (Lee & Park, 2020), recurrent neural networks (Alcantara et al., 2022), encoder-decoder convolutional architectures (Faisal et al., 2023), and Convolution models (Johnson et al., 2021). Many studies have focused primarily on the vertical GRF component because it is strongly coupled to bodyweight support and typically has a greater force magnitude than the anteroposterior and mediolateral GRF components during walking (Chockalingam et al., 2016; Guo et al., 2017; Wouda et al., 2018). However, the AP and ML components carry critical information about propulsion, braking, and lateral balance control, and they are often the most sensitive to stability deficits and compensatory strategies (Peterson et al., 2010; Rawal & Singer, 2021). Therefore, 3D GRF estimation is a more informative target for functional mobility assessment compared to vertical only estimation, but also a more difficult learning problem due to lower signal-to-noise ratios and greater inter-subject variability in AP and ML components (Mundt et al., 2020; Shahabpoor et al., 2018).

In addition to GRFs, estimating GRMs and CoP provides further interpretability and clinical utility. GRMs provide a direct measure of net external rotational effects at ground contact and can complement GRFs in describing stability and turning mechanics. CoP trajectories show where a person's weight is centered on the ground at each moment. As balance is maintained by keeping this point within the base of support, the CoP trajectory directly reflects how postural control is maintained throughout a movement. In the context of TUG, GRM, and CoP can help quantify the turning strategy, asymmetry in weight transfer, and balance control during sit-to-walk initiation. However, fewer studies have addressed these outputs compared with GRFs, and methodological gaps remain regarding robust estimation during transitional tasks rather than steady-state walking.

The central aim of this thesis is to enable kinetic-level assessment of the TUG during the STW phase, including 3D GRF, GRM, and CoP, using wearable IMU data and machine learning. The work is organized as a progression from the synthesis of the state of the art to rigorous model comparison for GRF prediction, and finally to an extension toward broader kinetic outcomes (GRM and CoP) that better support a stability-focused interpretation.

1.2 Motivation

The motivation for this research comes from the need to connect laboratory biomechanics with scalable clinical assessments. On the one hand, GRF, GRM, and CoP measurements enable a mechanistic understanding of movement and are useful for clinicians to understand the underlying cause of mobility impairment. However, the most common clinical tools for mobility assessment, such as the TUG, only provide a single summary score and do not reveal the root causes of mobility impairment. This gap hinders clinical application: clinicians typically rely on time scores and visual assessments, whereas researchers develop kinetic markers typically measured through force plates, but are difficult to use outside the lab. The rise of wearable sensing and machine learning offers an opportunity to bridge this divide by developing a portable kinetic estimation system that infers kinetic parameters from wearable data. If successful, kinetic iTUG could deliver force and stability measures to enable more precise tracking of functional changes. This kind of monitoring is important in many settings, such as fall-risk detection in older adults, stroke or joint-replacement rehabilitation, Parkinsonian mobility assessment, and evaluation of interventions for strength, balance, or gait training. However, several challenges must be addressed for this opportunity to translate into clinical value. First, machine learning models must be evaluated using rigorous protocols that reflect the requirements for clinical generalization. Second, the model choice must be justified, as different architectures embed distinct inductive biases for temporal dependencies and multiscale GRF structures. Third, kinetic outcomes beyond GRF, especially GRM and CoP, must be included to provide clinically interpretable stability measures rather than force magnitudes. These considerations motivate the three core lines of work in this thesis:

- A systematic literature review to map existing machine learning methods for 3D GRF estimation,
- A comparative study of different machine learning architectures for GRF prediction during STW trials, and
- An extension of the framework to estimate GRM and CoP during the STW.

1.3 Research Objective

The TUG test comprises several biomechanically distinct transitions, including sit-to-walk, steady-state walking, walk-to-turn, turning, turn-to-walk, and walk-to-sit. A complete kinetic instrumentation of the TUG would require estimation across all these transitions. Among them, the STW transition is the first and the most mechanically demanding, as it requires simultaneous generation of momentum, postural stabilization, and gait initiation, and has demonstrated established sensitivity to age-related and pathological decline (Buckley et al., 2009; Buckley et al., 2008; Perera et al., 2023). This thesis, therefore, focuses on the STW subtask as a first step toward full kinetic instrumentation of the TUG.

The primary objective of this research is to develop and evaluate machine learning methods for kinetic instrumentation of the TUG test by estimating 3D GRF, GRM, and CoP trajectories using wearable IMU data during the STW transition.

This research aims to assess the feasibility and performance of different machine learning models for estimating kinetic variables traditionally measured with force plates, based on wearable IMU signals collected during the STW portion of the TUG test. The work is structured around the following research questions:

1. What is the state of the art of machine learning methods for estimating 3D GRFs from wearable sensors across human movement tasks?

The first research question aims to compile and interpret the methods currently used to estimate the GRF from IMU data. This includes the tasks studied, sensor placements, coordinate systems, normalization approaches, modeling families, and evaluation protocols. It is motivated by the need for a structured understanding of methodological choices in the field and of the gaps in existing literature.

2. How do different machine learning architectures compare in predicting 3D GRFs from IMU data during STW, and which architecture performs best under rigorous evaluation?

The second research question shifts from synthesis to model development and benchmarking. The answer to this question is obtained by implementing and comparing different architectures

from multiple inductive-bias families under a consistent preprocessing and evaluation pipeline to determine which architectures are robust and which inductive biases are best suited to the temporal structure of STW kinetics.

3. What are the generalization characteristics and failure modes of these models for STW kinetics, including subject variability, phase-specific errors, and sensitivity to sensor-derived noise and preprocessing choices?

The third research question examines the reliability and practical considerations for deploying wearable kinetic estimation. In clinical contexts, models must generalize to unseen individuals who may have different anthropometrics, movement strategies, and impairment profiles. They must also remain stable in the presence of sensor noise, minor placement differences, and realistic signal artifacts. This research question, therefore, evaluates model behavior at the subject level and explores error distributions rather than relying only on window-level summary metrics. It also motivates the use of robust evaluation methods and paired statistical tests that reflect the repeated-measures nature of subject-based comparisons.

4. Can the wearable-machine learning framework be extended to estimate GRM and CoP trajectories during STW to enable richer stability-focused kinetic markers?

The final research question expands the problem from GRF estimation to broader kinetic outcomes. GRFs provide the magnitude of the force but do not directly indicate where the force is applied (CoP) or the rotational effects at contact (GRM). Since stability control is often better characterized using CoP and related measures, and because turning and transitional tasks impose substantial rotational demands, estimating CoP and GRM could yield more clinically interpretable outcomes for iTUG. The answer to this question is beneficial in two respects: it enables richer kinetic markers that align with clinical concepts such as weight transfer and stability, and it tests whether machine learning models trained for GRF can be adapted or extended to related kinetic outputs derived from the same ground truth.

1.4 Thesis Structure

The thesis structure and key objectives are outlined in Figure 1-2. This thesis unfolds as a systematic progression of studies, organized into chapters that build upon each other. Chapters 2-4 have been prepared for publication in peer-reviewed journals and may therefore exhibit some recurring content. Within this thesis, each chapter is introduced with a brief preface and, where applicable, retains its original published form.

Chapter 2. Systematic Literature Review. This chapter presents a PRISMA-guided systematic review of machine learning methods for estimating 3D GRFs from wearable IMU sensors across human movement tasks. It synthesizes sensor configurations, preprocessing pipelines, learning targets, model families, and evaluation protocols reported in the literature, and identifies the methodological gaps most relevant to transitional tasks such as STW. This chapter addresses RQ1 and provides the foundation for the architectural choices examined in later chapters.

Chapter 3. Predicting Ground Reaction Force during Sit-to-Walk. This chapter reports an experimental study comparing 13 model architectures across recurrent, convolutional, multiscale encoder–decoder, hybrid, residual-correction, and attention-based families for predicting 3D GRFs from a single sacrum-mounted IMU during STW. Models are evaluated using a consistent pipeline with subject-level metrics and paired statistical testing and are analyzed not only for average error but also for their ability to reproduce salient waveform features and component-wise behavior. This chapter addresses RQ2 and RQ3 and identifies the DS-Conformer as the best-performing architecture for STW GRF prediction.

Chapter 4. Deep Learning-Based Kinetics and Center of Pressure Estimation during Sit-to-Walk. This chapter extends the framework developed in Chapter 3, from GRF prediction to the simultaneous estimation of GRM and CoP trajectories during STW, using a single sacrum-mounted IMU. It defines learning targets and coordinate conventions for CoP and GRM derived from force plate measurements, evaluates models under a leave-one-out cross-validation protocol, and reports component-wise performance for vertical, anteroposterior, and mediolateral outputs. This chapter addresses RQ4 and demonstrates that the wearable–machine

learning framework can yield richer, stability-focused kinetic markers of potential relevance to clinical interpretation.

Chapter 5. Integrated Discussion and Conclusion. This chapter integrates the findings of the preceding chapters and discusses their collective implications for the kinetic instrumentation of the TUG. It revisits the research questions, situates the results within the wider biomechanics and wearable-sensing literature, considers the clinical and methodological significance of the work, acknowledges its limitations, and outlines avenues for future research toward a fully kinetically instrumented TUG test.

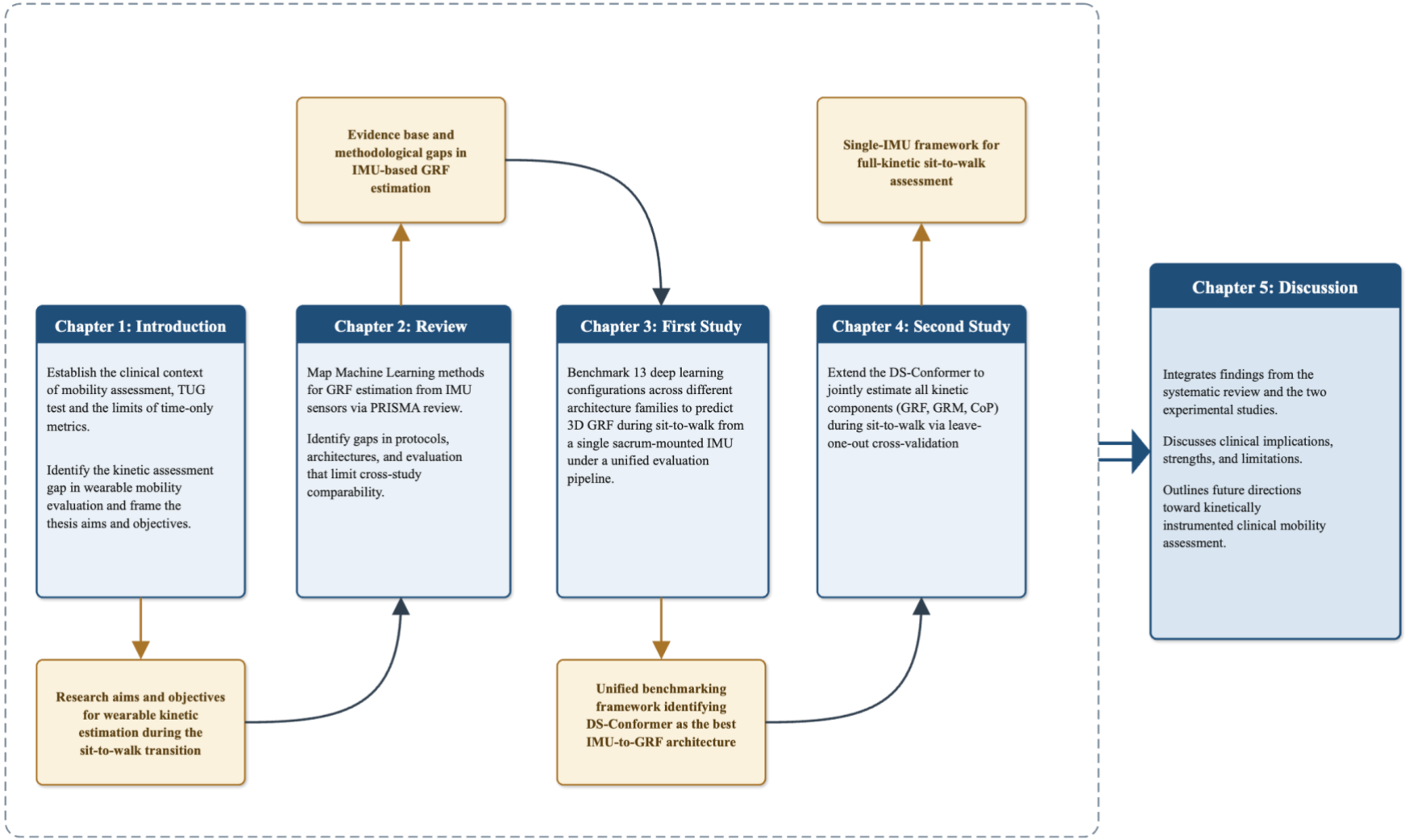


Figure 1-2 Thesis Structure with objectives

Chapter 2. Application of Machine Learning Techniques in the Measurement of Ground Reaction Force (GRF) waveforms using Inertial Measurement Unit (IMU) Sensors - A Systematic Literature Review

2.1 Preface

The content of this chapter is a copy of the article "Application of Machine Learning Techniques in the Measurement of Ground Reaction Force (GRF) Waveforms Using Inertial Measurement Unit (IMU) Sensors - A Systematic Literature Review," submitted for publication. Supplementary materials for this chapter are provided in Appendix A.

Chapter 1 established the broader clinical and biomechanical motivation for this thesis by arguing that functional mobility assessment would benefit from kinetic measures such as ground reaction forces, but these measures require force plates, which are expensive and require fixed installation, thereby limiting their clinical utility. It further identified wearable inertial measurement units and machine learning as a promising route for low-cost kinetic assessment. Building directly on that foundation, the present chapter addresses the first research question of the thesis: "What is the state of the art of machine learning methods for estimating 3D GRFs from wearable sensors across human movement tasks?"

This question is addressed through a systematic literature review undertaken in accordance with PRISMA guidelines. The chapter critically synthesizes prior studies across the entire methodological pipeline, including participant populations, movement tasks, sensor configurations, preprocessing strategies, machine and deep-learning models, validation protocols, and reported performance.

Within the overall thesis, the purpose of this review is not only to summarize existing studies but also to define the current methodological landscape of wearable GRF estimation, identify recurring strengths and limitations in the literature, and clarify the unresolved gaps most

relevant to kinetic instrumentation of functional mobility tasks. These insights then motivate and inform the modeling work presented in the chapters that follow.

Submitted Manuscript Starts here

Application of Machine Learning Techniques in the Measurement of Ground Reaction Force (GRF) waveforms using Inertial Measurement Unit (IMU) Sensors - A Systematic Literature Review

2.2 Abstract

Background: 3D ground reaction forces (GRFs) are important parameters in human motion analysis, yet their measurement is limited to laboratory settings. Over recent years, GRF estimation using machine learning and deep learning from wearable inertial measurement unit (IMU) data has emerged as a promising avenue, enabling kinetic analysis outside laboratory settings. However, no consensus on standardized methods or best practices has emerged yet. This review aims to synthesize the methods used in existing studies and to provide a coherent framework of current practices and recommendations for future work.

Methods: The systematic review was conducted in accordance with PRISMA guidelines (PROSPERO CRD42024578576). Databases including PubMed, Scopus, IEEE Xplore, and ScienceDirect were searched from 2011 to 2025. Studies were included if they used IMUs and machine learning or deep learning algorithms to estimate GRFs and reported quantitative performance metrics. Data was extracted on participant characteristics, IMU specifications and placement, preprocessing techniques, machine learning and deep learning models used, validation protocols, and reported error metrics.

Results: 22 studies met the inclusion criteria, with sample sizes ranging from 5 to 27 (median = 12). IMU number and placement varied widely (n=1 to 12). Most studies used triaxial accelerometers and gyroscopes sampled at 100-1,000 Hz. Classical machine learning approaches with handcrafted features achieved vertical GRF relative RMSE (rRMSE) as low as 5.2%. More recent deep learning models processing raw IMU streams reported vertical GRF rRMSE typically between 4% and 7%. Anteroposterior and mediolateral components remained less accurate, with rRMSE generally in the 8%-12% range. Across studies, models trained and tested on treadmill walking showed lower vertical rRMSE (<6%) than those trained and tested on overground or dynamic tasks (7%-10%). Heterogeneity in validation protocols and demographic samples limited direct cross-study comparisons.

Conclusion: This review synthesizes implementation choices and their impact across machine learning and deep learning pipelines to estimate GRFs from IMU data. By comparing model architectures, feature-engineering strategies, sensor configurations, and validation protocols, we highlight how these decisions directly influence performance, particularly across vertical, anteroposterior, and mediolateral components. The substantial variability in preprocessing, feature extraction, and evaluation frameworks underscores the urgent need for standardized methodologies. Future research should focus on larger, more diverse cohorts, consistent performance metrics, and benchmarking against public datasets to enable fair cross-study comparisons and accelerate clinical translation.

2.3 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 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 & 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, 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).

From a clinical perspective, GRF metrics provide objective insights into 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 (Chockalingam et al., 2016; Williams et al., 2011). 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 (Apatsidis 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 for routine

evaluation, even though GRF analysis could provide richer, objective 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 (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; Huang et al., 2025; Razak et al., 2012), and instrumented shoes (Faivre et al., 2004). However, the presence of sensors mounted inside or under the shoes can affect natural foot biomechanics (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 2-1, highlighting the key differences between traditional feature-engineered machine learning pipelines and modern end-to-end learning deep learning frameworks.

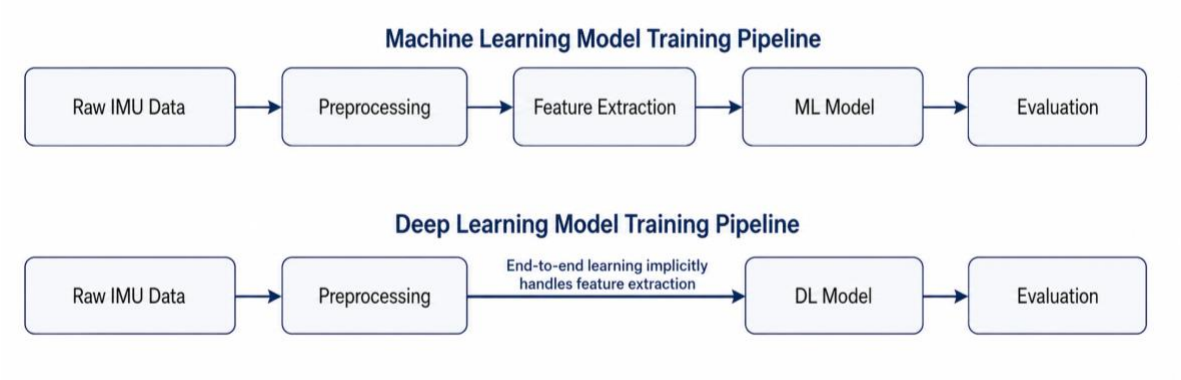


Figure 2-1 Machine learning and deep learning model training pipeline

2.3.1 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.

This review focuses specifically on machine-learning pipeline choices for IMU-based GRF estimation, building on the related reviews (Ancillao et al., 2018; Razak et al., 2012), and 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.

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?

2.4 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).

2.4.1 Search Strategy and Selection Criteria

Electronic databases, including PubMed, Scopus, IEEE Xplore, and ScienceDirect, were searched from 2011 to April 2025. 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 Appendix A, Table S2-1.

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 (PICO) 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 grey 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).

2.4.2 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.

2.4.3 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 Appendix A, Table S2-2

2.4.4 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.

2.5 Results

The initial electronic database search yielded 628 articles, and 15 additional studies were manually identified by examining the reference lists. After removing 99 duplicate articles, 544 studies underwent title and abstract screening, followed by full-text evaluation and quality assessment, resulting in 22 articles included in the systematic review. The PRISMA Diagram is presented in Figure 2-2**.**

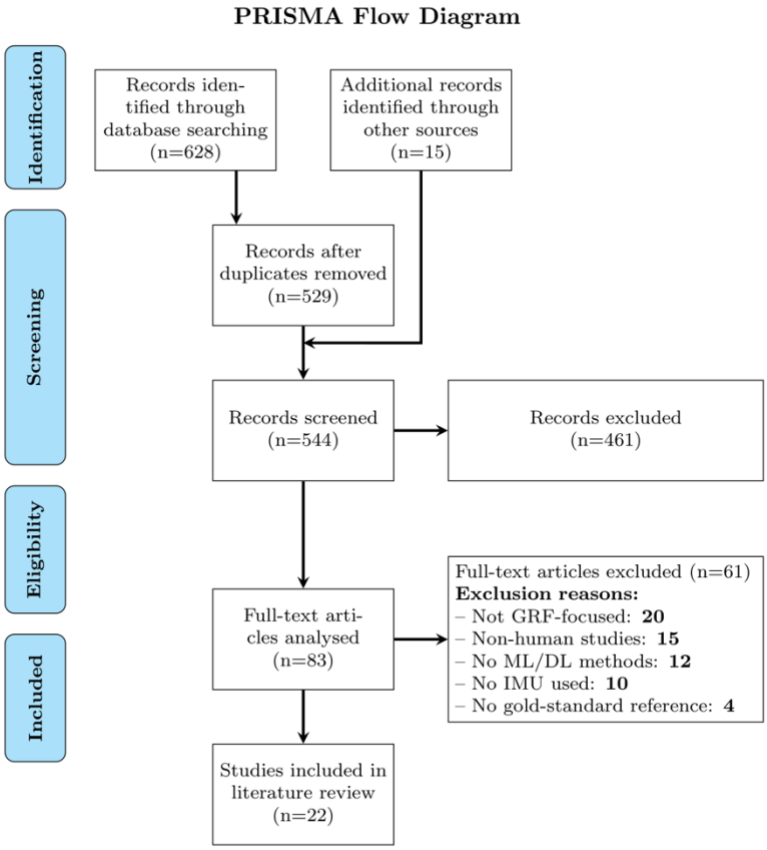


Figure 2-2 PRISMA Flow Diagram

To ensure direct alignment with the review aims, the results are structured around the five research questions listed above. Specifically, Sections 2.5.1 to 2.5.4 report evidence addressing research questions 1 to 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.

2.5.1 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.

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 to 27, with a median of 12. Seven 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 Appendix A, Table S2-3.

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 and 3.1-4.9 m/s, 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 and 0.7-1.6 m/s, respectively). Scheltinga et al. (2023) focused on running at fixed speeds while Wouda et al. (2018) explored multispeed running. Alcantara et al. (2022) introduced variable slopes (±5° and ±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 50m 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 & Hahn, 2023a) assessed gait on level, inclined, and declined surfaces, and their follow-up work Donahue and Hahn (2023b) introduced outdoor running on a 400m track at different speeds.

Detailed description of tasks performed by each study is given in Table 2-2

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, 2023b) used Loadsol insole force sensors (Novel Electronics, St. Paul, MN, USA), sampled at 100Hz, 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 Appendix A, Table S2-4.

4. IMU Sensor Characteristics and Placement

The number of IMU sensors ranged from one to twelve. Common placements included the pelvis, thigh, shank, tibia, and foot. Table 2-1 lists the anatomical distribution of IMU sensor placements across the reviewed studies. Figure 2-3 summarizes sensor placements by study and the frequency with which anatomical locations were used together.

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. None of the studies reported using a magnetometer. 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 6 out of 22 studies used a single IMU, 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 tabulated in Appendix A, Table S2-4.

Table 2-1 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
Jiang et al. (2020); Johnson et al. (2021); Havashinezhadian et al. (2023); Dorschky et al. (2020); Scheltinga et al. (2023); Wouda et al. (2018); Cerfoglio et al. (2021); Sun et al. (2023); Hossain et al. (2023); Leporace et al. (2015); Bach et al. (2022); Shahabpoor et al. (2018); Leporace et al. (2018)	Shank/tibia	13	59.1
Jiang et al. (2020); Havashinezhadian et al. (2023); Dorschky et al. (2020); Scheltinga et al. (2023); Alcantara et al. (2022); Sun et al. (2023); Hossain et al. (2023); Donahue and Hahn (2023b); Donahue and Hahn (2023a); Ngoh et al. (2018); Shahabpoor et al. (2018)	Foot/shoe	11	50.0
Jiang et al. (2020); Johnson et al. (2021); Dorschky et al. (2020); Scheltinga et al. (2023); Cerfoglio et al. (2021); Sun et al. (2023); Hossain et al. (2023); Shahabpoor et al. (2018)	Thigh	8	36.4
Lee and Park (2020); Alcantara et al. (2022); Cerfoglio et al. (2021); Donahue and Hahn (2023b); Donahue and Hahn (2023a); Lim et al. (2019); (Kim et al., 2020)	Sacrum	7	31.8
Johnson et al. (2021); Scheltinga et al. (2023); Wouda et al. (2018); Sun et al. (2023)	Pelvis/waist (ASIS region)	4	18.2
Dorschky et al. (2020); Guo et al. (2017); Shahabpoor et al. (2018)	Lower back (L5 / unspecified lower back)	3	13.6
Guo et al. (2017); Shahabpoor et al. (2018)	Upper trunk (C7)	2	9.1
Scheltinga et al. (2023); Sun et al. (2023)	Trunk/chest (thoracic)	2	9.1
Guo et al. (2017)	Head/forehead	1	4.5
Shahabpoor et al. (2018)	Upper arm	1	4.5
Shahabpoor et al. (2018)	Forearm	1	4.5

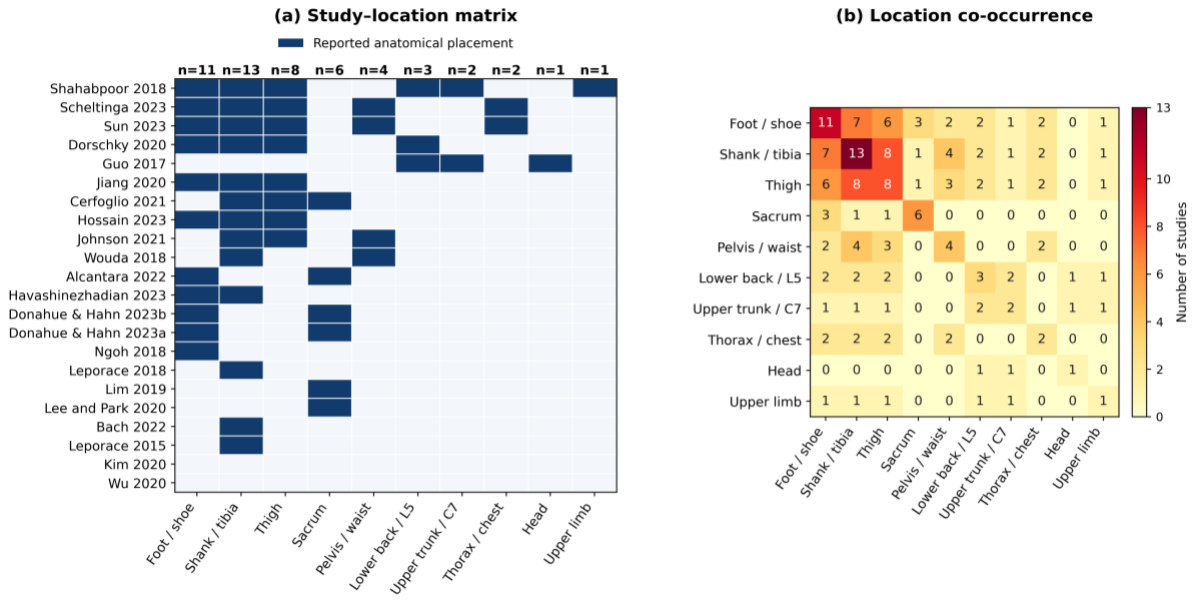


Figure 2-3 Sensor placements across the 22 reviewed studies. (a) Filled cells indicate reported anatomical placements; column totals count studies using each location. (b) The co-occurrence matrix counts studies using each pair of locations; diagonal entries give the location totals. Bilateral or multiple sensors at one location are counted once per study. Wu et al. (2020) used sensors beneath the walking surface, and Kim et al. (2020) used marker-derived sacral/CoM input; these studies are retained as rows but are not counted as anatomical IMU placements. Frequency of use does not indicate predictive accuracy.

2.5.2 RQ2 - Preprocessing and feature extraction

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

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 (Jiang et al., 2020; Ngoh et al., 2018; 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.

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 Hz to 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. 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; Hossain et al., 2023; Jiang et al., 2020; Lee & Park, 2020).

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.

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 from 110ms windows. Two studies (Leporace et al., 2015; Leporace et al., 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 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 & 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.

2.5.3 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 22 included studies is presented in Table 2-2**,** and a detailed description is provided in the subsections below. Notably, Table 2-2 shows that early work emphasized interpretable, low-parameter methods, including NARMAX, feedforward neural networks, multilayer perceptrons (MLPs), and Random Forest. More recent studies favor sequence-modeling approaches, such as LSTM, Echo State Networks, Convolutional Neural Networks (CNN), and hybrid or ensemble frameworks. Implementation details of each study, including layer dimensions and hyperparameter settings are provided in Appendix A, Table S2-6.

Table 2-2 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	Participants Characteristics	IMUs Used	Task Performed	Model Used	Accuracy (ML)	Accuracy (AP)	Accuracy (V)
(Guo et al., 2017)	9 Healthy (6M, 3F) 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)	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)	11 Healthy (6M, 5F) Age: 24.5 ± 1.7 years height: 170.5 ± 10.3 cm	3	Vertical drop jumps	NARX	RMSE: 0.0081 N/kg	RMSE: 0.0350 N/kg	RMSE: 0.0184 N/kg

	weight: 65 ± 11.3 kg						
(Jiang et al., 2020)	9 Healthy Males age: 27 ± 8.0 years height: 176 ± 7cm weight: 72 ± 9 kg	4	Treadmill Walking for 1 min at five different speeds: 0.4 m/s, 0.7 m/s, 1.0 m/s, 1.3 m/s, 1.6 m/s	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%
(Lee & Park, 2020)	17 analyzed participants (20 recruited: 12F, 8M; 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) for 2 minutes at each speed.	ANN	**0.7 m/s Walking Speed:** NRMSE: 17.75 ± 6.41% R^2: 0.54 ± 0.40 **1.0 m/s Walking Speed:** NRMSE: 15.23 ± 5.50% R^2: 0.66 ± 0.25	**0.7 m/s Walking Speed:** NRMSE: 13.52 ± 4.38% R^2: 0.77 ± 0.15 **1.0 m/s Walking Speed:** NRMSE: 8.31 ± 2.89% R^2: 0.91 ± 0.07	**0.7 m/s Walking Speed:** NRMSE: 6.86 ± 2.21% R^2: 0.94 ± 0.04 **1.0 m/s Walking Speed:** NRMSE: 6.17 ± 1.81% R^2: 0.94 ± 0.03 **1.3 m/s Walking Speed:** NRMSE: 6.08 ± 1.88% R^2: 0.94 ± 0.04

					1.3 m/s Walking Speed: NRMSE: 14.19 ± 3.91% R^2: 0.70 ± 0.16 **1.6 m/s Walking Speed:** NRMSE: 15.04 ± 2.78% R^2: 0.63 ± 0.14	**1.3 m/s Walking Speed:** NRMSE: 7.34 ± 2.46% R^2: 0.92 ± 0.05 **1.6 m/s Walking Speed:** NRMSE: 7.67 ± 2.73% R^2: 0.91 ± 0.06.	**1.6 m/s Walking Speed:** NRMSE: 7.68 ± 3.15% R^2: 0.89 ± 0.09.
(Wouda et al., 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 for 3 minutes 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

(Lim et al., 2019)	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-minute 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)	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:** RMSE: 0.017 ± 0.010 BW r: 0.99 **At 9 km/h:** RMSE: 0.015 ± 0.0072 BW r: 0.99 **At 10 km/h:** RMSE: 0.017 ± 0.0091 BW r: 0.99
(Leporace et al., 2015)	17 Healthy (11M, 6F) Age: 27.1 ± 3.4 years	1	Overground Walking - 8m	ANN (MLP)	Single Network with Three Outputs: MAD: 1.4 ±	**Single Network with Three Outputs:**	Single Network with Three Outputs: MAD: 5.2 ± 1.7 %BW

	Weight: 84.3 ± 4.5 kg				0.5 %BW r: 0.801 Three Separate Networks with One Output Each: MAD: 1.35 ± 0.46 %BW r: 0.792	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	r: 0.968 Three Separate Networks with One Output Each: MAD: 5.1 ± 1.64 %BW r: 0.969
(Leporace et al., 2018)	17 Healthy (11M, 6F) Age: 27.1 ± 3.4 years Weight: 84.3 ± 4.5 kg	1	8m 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
(Kim et al., 2020)	7 Healthy Males Age: 25.0 ± 2.9 years Height: 168.8 ± 7.5 cm	1	Treadmill walking at three speed (1.0 to 2.3 m/s)	ANN	N/A	**Method 1 (CoM Kinematics):** NRMSE: 6.48 ±	**Method 1 (CoM Kinematics):** NRMSE: 6.93 ± 4.02% r: 0.98 ± 0.01

	Weight: 72.1 ± 7.7 kg					2.93% r: 0.98 ± 0.01 **Method 2 (SLIP Transform):** NRMSE: 5.28 ± 0.99% r: 0.98 ± 0.01	**Method 2 (SLIP Transform):** NRMSE: 6.30 ± 1.31% r: 0.98 ± 0.01
(Schelting a et al., 2023)	12 Healthy (8M, 4F) 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 ± 0.21 **Hybrid** RMSE = 0.05 ± 0.01 rRMSE = 10.8 ± 2.5% r = 0.58 ± 0.22	**Direct** RMSE =0.08 ± 0.03 rRMSE = 7.8 ± 3.1% r = 0.91 ± 0.09 **Hybrid** RMSE = 0.07 ± 0.03 rRMSE = 7.8 ± 3.2%	**Direct** RMSE = 0.19 ± 0.04 rRMSE = 7.3 ± 1.8% r = 0.97 ± 0.02 **Hybrid** RMSE =0.18 ± 0.04 rRMSE = 6.8 ± 1.7% r = 0.97 ± 0.01

						r = 0.91 ± 0.10	
(Havashinezhadian et al., 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)	21 Healthy (13M, 8F) 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 running Walking speed: 1.24 ± 0.12 m/s Running speed: 2.20 ± 0.14 m/s	ESN	N/A	N/A	NRMSE: 6.8% ± 0.3% R^2: 0.96 ± 0.00
(Donahue & Hahn, 2023a)	13 Healthy (8M, 5F) 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:** RMSE: 0.315 ± 0.102 BW **3.00 m/s:** RMSE: 0.294 ± 0.065

	BW **4.00 m/s:** RMSE: 0.339 ± 0.085 BW **4.50 m/s:** RMSE 0.397 ± 0.180 BW **5.00 m/s:** RMSE: 0.427 ± 0.010 BW **Decline running** **2.25 m/s:** RMSE: 0.253 ± 0.081 BW **3.25 m/s:** RMSE: 0.277 ± 0.065 BW **4.00 m/s:** RMSE: 0.309 ± 0.082 BW **4.50 m/s:** RMSE: 0.576 ± 0.371 BW **5.00 m/s:** RMSE: 0.635 ± 0.000 BW **Incline running** **2.25 m/s:** RMSE: 0.230 ± 0.081 BW

							3.25 m/s: RMSE: 0.277 ± 0.065 BW **4.00 m/s:** RMSE: 0.309 ± 0.082 BW **4.50 m/s:** RMSE: 0.269 ± 0.007 BW
(Donahue & Hahn, 2023b)	15 Healthy (9M, 6F) age: 23.6 years height: 178.3cm weight: 73.5kg	3	Outdoor Running - 400m. 5 trials with increasing velocity (2.33 - 5.36 m/s)	BD-LSTM	N/A	N/A	**Stance phase RMSE:** 0.189-0.288 BW **Whole waveform RMSE:** 0.191-0.309 BW
(Wu et al., 2020)	5 Healthy (3M, 2F) age: 24 ± 2 years height: 168 ± 2 cm weight: 64 ± 11 kg	4	Overground Walking	LSTM	NRMSE: 6.0% ±1.1%.	NRMSE: 3.9% ± 0.5%	NRMSE:4.3% ± 0.8%.
(Alcantara et al., 2022)	19 Healthy (10M, 9F) Age: 29 ± 9 years Height: 173 ± 9 cm Weight: 68.1 ± 9.9 kg	3	Treadmill Running five different slopes: 0°, +5°, -5°, +10°, and -10°. three speeds: 2.5 m/s, 3.33 m/s, and 4.17 m/s for each slope.	LSTM	N/A	N/A	RMSE: 0.16 ± 0.04 BW rRMSE: 6.4 ± 1.5%

(Sun et al., 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 R^2 = 0.85 ± 0.11
(Dorschky et al., 2020)	10 Healthy Males age: 27 ± 2.6 years height: 182 ± 5cm weight: 76.9 ± 8.6 kg	4	Treadmill walking and running at six different speeds Walking Speeds: 0.9-1.0 m/s, 1.2-1.4 m/s, 1.8-2.0 m/s Running Speeds: 3.1-3.3 m/s, 3.9-4.1 m/s, 4.7-4.9 m/s	CNN	N/A	**Only measured data** RMSE: 4.29-4.52% BW r = 0.970-0.976 **With simulated data** RMSE: 4.21-4.95%	**only measured data** RMSE: 11.74-13.24% BW r = 0.979-0.988 **With simulated data** RMSE: 12.16% and 16.07% BW r = 0.974-0.991

						BW r = 0.966-0.980	
(Johnson et al., 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)	20 healthy (12 M, 8F) 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	**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**	**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 ±

	r: 0.847 ± 0.066 **Stair walking** NRMSE: 8.16 ± 1.56 r: 0.885 ± 0.042	NRMSE: 3.99 ± 1.03 r: 0.951 ± 0.040 **Stair walking** NRMSE: 4.64 ± 0.87 r: 0.926 ± 0.030	2.24 r: 0.983 ± 0.024

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. Three IMUs were placed at the fifth lumbar vertebra (L5), seventh cervical vertebra (C7), and forehead; the L5-based estimate produced the errors reported below. The authors first decomposed raw acceleration data into left-dominant and right-dominant components 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:

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

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

$$vGRF(k) = F(asensor(k), asensor(k-1), \dots, asensor(k-L))$$

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 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 using a 12-IMU configuration at C7, L5, and bilaterally on the upper arms, forearms, thighs, shanks, and fourth metatarsals, based on the correlations between segmental accelerations and GRFs. Model 1 utilized Pearson's correlation coefficient, r, 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 three IMUs on the sacrum, right thigh, and right shank. The model reported mean RMSE values of ≈0.0081-0.0350 N/kg, 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.

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 Pearson r 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. Artificial Neural Networks

Artificial Neural Networks (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 and tested on 12 km/h 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.

Feedforward Neural Networks (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 1.7 % of bodyweight (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 multilayer perceptron (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 Pearson's correlation coefficients of r = 0.97 for AP GRF, r = 0.97 for vertical GRF, and r = 0.80 for ML GRF. A further study by Leporace et al. (2018) reported Pearson's correlation coefficients (r) and mean absolute deviations (MAD) for vertical GRF (r = 0.98, MAD = 4.5 %BW), AP GRF (r = 0.97, MAD = 1.8 %BW), and ML GRF (r = 0.80, MAD = 1.4 %BW). These findings demonstrate that even a simple feedforward neural network, when paired with a well-chosen feature set, can capture kinetic parameters using data from just a single IMU.

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:

$$vGRF = (m_b \cdot g) + i = \sum_{i=1}^{3} \left(m_b \cdot WF_i \cdot a_{z,i}\right)$$

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 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 Spring-Loaded Inverted Pendulum model with ANNs to predict GRFs during walking using data from a single IMU mounted on the sacrum. Two methods were evaluated. Method 1 utilized direct Center of Mass (CoM) kinematics, including displacement, velocity, acceleration, and temporal information from IMU sensors placed at the sacrum. 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.

5. Recurrent Neural Networks

Recurrent Neural Networks (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 Echo State Network (ESN) is a type of recurrent neural network 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 ± 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 (~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. Across the 36 combinations of GRF component, gait-cycle intervals and different training approaches evaluated by the authors, the top-of-shoe sensor location produced the lowest MAE in 26 conditions (72.2%) for the healthy group and 15 conditions (41.7%) for the medial knee osteoarthritis (MKOA) group. These percentages describe how often this placement performed best across evaluation conditions, rather than percentages of participants or the magnitude of prediction error.

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 Long Short-Term Memory (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 to 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 ± 0.12 during single-leg drop landings, and 0.85 ± 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 16 ± 4 %BW and rRMSE of 6.4 ± 1.5% normalized to the average range of the measured and predicted waveforms 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.

6. Convolutional Neural Networks

The Convolutional Neural Networks (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× to 16× the original measured data). The simulated data was iteratively evaluated alongside the measured data to assess their impact on the CNN's performance. Inclusion of simulated data improved RMSE by 2% for AP GRF and 6% for vertical GRF.

Johnson et al. (2021) used CNNs to estimate GRFs, training on simulated accelerations and testing on wearable accelerometer recordings from five locations: the pelvis, both thighs, and both shanks. 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.

7. Ensemble Net

Ensemble Nets combine the predictions of multiple individual models. Hossain et al. (2023) introduced an ensemble framework, KineticsFM-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 (GRUs), one and two-dimensional convolutional layers, and fully connected dense layers to capture both temporal and spatial features from the IMU data. Finally, a Fusion Module (FM) integrates outputs from multiple primary models by learning optimal weights. The Double Loss Regression (DLR) strategy employed RMSE and Pearson r 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 Pearson r of 0.931.

2.5.4 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.

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 (Hossain et al., 2023; Lee & Park, 2020; 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 (Johnson et al., 2021; Wu et al., 2020).

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 Appendix A, Table S2-7.

Root Mean Square Error (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.

Relative Root Mean Square Error (rRMSE): rRMSE is the RMSE expressed relative to a reference value, typically as a percentage of the mean or peak value of the observed data. This normalization allows for comparison across datasets or variables with different scales.

Normalized Root Mean Square Error (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, making it easier to compare errors across different datasets or variables.

Mean Absolute Error (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.

Mean Absolute Deviation (MAD): MAD is similar to MAE, representing the average absolute difference between predicted and actual values.

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

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

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 r 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): In prediction evaluation, R^2 compares squared prediction error with the error of a constant mean-value reference. Its maximum is 1, and negative values indicate worse performance than this reference. It is not generally equivalent to squared Pearson correlation.

2.6 Discussion

This systematic review critically examined 22 studies 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.

2.6.1 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 to 27 participants. 7 studies 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 & Hortobagyi, 2000), and they tend to compensate for these deficits by increasing hip extension during the initial vertical GRF peak (DeVita & 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 cerebral palsy, 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 & 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 (Lee & Hidler, 2008; Riley et al., 2007). 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 & 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.

2.6.2 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). Residual synchronization errors can introduce temporal misalignment between predicted and measured GRF, altering correlation and increasing pointwise error. Donahue and Hahn (2023b) used an iterative procedure to correct clock drift between IMUs and force-sensing insoles, aligning initial-contact events between the systems to within approximately ±0.02 s.

Filtering is an important preprocessing consideration because noise reduction must be balanced against preservation of biomechanically relevant signal features. Butterworth low-pass filters are commonly used, although cutoff frequencies vary with the signal and application. For example, Alcantara et al. (2022) used cutoffs of 20 Hz for sacral acceleration and 30 Hz for GRF, whereas Donahue and Hahn (2023b)used 35 Hz for IMU signals and 20 Hz for insole force measurements. These settings represent study-specific choices rather than a generally applicable frequency range. Cutoff selection can affect the resulting biomechanical measurements: Schilaty et al. (2021) demonstrated that changing the cutoff frequency influenced force and kinetic measurements during simulated impacts. For body-mounted IMUs, soft tissue deformation can also cause the sensor to move relative to the underlying skeleton, introducing measurement artefacts. Rong and Kuo (2024) found that the characteristics of these artefacts varied with sensor placement and mass, and that low-pass filtering reduced acceleration overestimation while also attenuating the reference signal. Consequently, cutoff selection should account for the movement task, sensor configuration and signal features required for the intended analysis. Forward–backward filtering avoids phase shifts but requires future samples and therefore cannot be applied unchanged in a strictly causal real-time system.

Studies also differ in their choice of features. Classical machine learning approaches often extract summary statistics (mean, variance, root mean square), temporal features (stance time, stride frequency), or integrate accelerations to obtain velocities and displacements. Physics-informed features, such as estimates of center of mass 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.

2.6.3 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.

In contrast, artificial neural networks (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 (Lee & Park, 2020; Lim et al., 2019) 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 (Donahue & Hahn, 2023a, 2023b; Guo et al., 2017) 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.

2.6.4 Evaluation metrics and validation protocols

Comparisons across studies require consistent definitions of error normalization and evaluation intervals. For example, Alcantara et al. (2022) calculated relative RMSE by dividing RMSE by the average range of the measured and predicted waveforms, whereas Donahue and Hahn (2023b) reported RMSE in bodyweight units. These measures have different denominators and should not be treated as interchangeable percentages. Dividing a waveform's RMSE by its peak, mean or range rescales the error; it does not, by itself, give greater importance to peaks

or movement phases. Bodyweight normalization expresses loading relative to the participant's weight, facilitating comparisons between people of different body masses. However, reporting only normalized values conceals differences in absolute force: a force of 1 BW corresponds to 600 N for a person weighing 600 N and 800 N for a person weighing 800 N. Thus, normalization removes the direct bodyweight-related scale difference but does not eliminate individual differences in relative loading or waveform shape. The normalization formula should therefore be stated explicitly, with absolute forces retained where relevant to interpretation.

The evaluation interval also influences reported error. Donahue and Hahn (2023b) reported both continuous-waveform and stance-phase RMSE, illustrating the distinction between these assessments. Because RMSE averages squared errors across the evaluated samples, adding correctly predicted zero-force samples reduces the overall value without improving prediction during loaded periods. This reduction is conditional on prediction accuracy during those intervals; including unloaded periods does not necessarily reduce RMSE. Studies should therefore specify which movement phases were evaluated. Stance-phase errors are useful for assessing foot–ground loading, while whole-sequence errors remain relevant when the prediction task includes transitions or unloaded periods.

Waveform association should also be distinguished from agreement in force magnitude. Pearson's correlation coefficient measures linear association and is insensitive to constant offsets and positive amplitude scaling; consequently, a high correlation can coexist with substantial differences in predicted force. Correlation should therefore be reported alongside an error measure such as RMSE or MAE (Iosa et al., 2014). Temporal misalignment can also change correlation and increase pointwise error. Acquisition-related timing offsets should be corrected before evaluation using a documented synchronization procedure. If predictions are subsequently shifted to maximize cross-correlation, the resulting measure should be reported separately, because this adjustment removes part of the timing error being evaluated.

Validation should reflect the intended application. For prediction in previously unseen individuals, all trials from a participant must remain within the same partition in each training–test split; otherwise, shared participant characteristics can produce overly optimistic estimates of generalization (Halilaj et al., 2018). Both leave-one-subject-out and participant-grouped k-fold cross-validation can provide this separation. Leave-one-subject-out validation requires one

model-training cycle per participant and can therefore be computationally demanding (Alcantara et al., 2022). A separate, representative test set is also valid when sufficient data are available. Feature selection, data-dependent preprocessing and hyperparameter tuning should use training and validation data without access to the test set, which must remain reserved for final evaluation (Halilaj et al., 2018). Accordingly, studies should report participant-level partitioning, the normalization formula, evaluation intervals, synchronization procedures, and complementary measures of waveform association and prediction error.

2.6.5 Deployment challenges and future directions

Machine learning-based estimation of GRF from IMUs has progressed from proof-of-concept demonstrations to near-clinical viability. High accuracy in vertical GRF under controlled conditions has been achieved, but substantial challenges remain for real-world deployment.

- 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 1-12 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.
- 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 & van den Bogert, 2024).
5. The gait data from 51 healthy participants, collected using motion capture, IMUs, and computer vision (Lavikainen et al., 2024).

2.7 Conclusion

This systematic review identified 22 studies 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 anteroposterior and mediolateral 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.

Submitted Manuscript Ends here

2.8 Chapter Summary

This chapter addressed the first research question of this thesis: "What is the state of the art of machine learning methods for estimating 3D GRFs from wearable sensors across human movement tasks?" A systematic review conducted in accordance with PRISMA guidelines identified 22 studies. Beyond the performance and methodological trends discussed there, three observations are particularly relevant to the direction of this thesis.

First, the field has progressed from interpretable, low-parameter methods such as NARMAX and feedforward neural networks toward deep learning architectures including CNNs, LSTMs, and bidirectional LSTMs. Within this progression, CNNs operating on raw IMU data consistently outperformed purely recurrent models, suggesting that the IMU-to-GRF mapping depends on spatiotemporal rather than purely temporal structure. Attention-based architectures, which replace fixed-length context vectors with dynamic, timestep-specific weighting of encoder states and have demonstrated strong performance in other sequence-to-sequence regression domains, appeared in only a small number of the reviewed studies. Whether attention mechanisms, either standalone or integrated into convolutional and recurrent frameworks, can improve GRF estimation, particularly for the more challenging anteroposterior and mediolateral components, remains an open question.

Second, earlier studies compared selected architectures, including CaffeNet and ResNet-50 (Johnson et al., 2021), motivating broader comparisons across architectural families under consistent preprocessing, validation and evaluation conditions. Because the reviewed studies differ simultaneously in cohorts, sensor configurations, and evaluation protocols, the performance differences highlighted in Section 2.7 cannot be cleanly attributed to the models themselves. It therefore remains unclear which architectural choices are most effective for wearable GRF estimation.

Third, the reviewed studies focused almost exclusively on cyclic, steady-state locomotion. None of the 22 included studies evaluated IMU-based GRF estimation during STW, which involves rapid postural reconfiguration and bilateral force asymmetry that are absent in steady-state gait, and which is the clinically salient transition targeted by this thesis.

Lim et al. (2019) demonstrated single-sensor GRF estimation during walking, while Johnson et al. (2021) compared alternative CNN architectures. Related work on CoP estimation during

walking, including Podobnik et al. (2020), provides additional context for extending wearable kinetic estimation to STW.

These three gaps: the limited exploration of attention-based architectures, the need for broader standardized architectural comparisons, and the lack of STW-specific GRF estimation among the included studies, directly motivate the experimental work presented in Chapter 3.

Chapter 3. Predicting Ground Reaction Force during Sit to Walk: Towards kinetically instrumenting Timed Up and Go Test

3.1 Preface

This chapter presents the experimental study: "Predicting Ground Reaction Force during Sit to Walk: Towards Kinetically Instrumenting Timed Up and Go." The paper has been submitted for publication. Supplementary materials are provided in Appendix B.

Building on the gaps identified in Chapters 1 and 2: a focus of most prior studies on cyclic locomotion tasks, limited use of single-sensor setups (only 6 of 22 studies), and the need for broader, standardized comparisons across machine learning architectures, this chapter evaluates 13 neural network configurations spanning seven architectural families for predicting three-dimensional GRF during sit-to-walk from a single sacrum-mounted IMU. Critically, all models are trained and tested under a single unified pipeline: the same dataset, identical preprocessing, bodyweight normalization, subject-wise cross-validation folds, and a common set of error and waveform-agreement metrics. This controlled design isolates architectural differences as the sole source of performance variation, enabling a consistent comparison across architectural families for STW prediction and allowing an evidence-based determination of which architecture is best suited for STW GRF prediction. Age-group generalization is additionally assessed through a leave-one-age-group-out protocol spanning young, middle-aged, and older adults.

Submitted Manuscript Starts here

Predicting Ground Reaction Force during Sit to Walk: Towards kinetically instrumenting Timed Up and Go Test

3.2 Abstract

Objective: Ground reaction force (GRF) during the sit-to-walk (STW) transition encodes clinically relevant information about lower-limb loading, postural control, and fall risk, yet its measurement currently requires laboratory-grade force plates that are impractical for routine clinical use. This study investigates whether continuous GRF waveforms during the STW can be accurately predicted from a single sacrum-mounted inertial measurement unit (IMU) using Deep Learning and systematically identifies which neural network architectures are best suited to this task.

Methods: Using a public dataset of 65 participants, we evaluated 13 model configurations across seven architectural families, ranging from shallow CNN baselines through recurrent, dilated convolutional, multi-resolution encoder-decoder, hybrid dual-stream fusion, residual-correction, and Conformer-based designs, each motivated by limitations observed in simpler architectures. All models were evaluated using five-fold subject-wise cross-validation. Generalization across age was assessed with a leave-one-age-group-out protocol spanning young (19-35), middle-aged (36-55), and older (56-73) adults.

Results: For vertical GRF, the Downsampled Conformer (DS-Conformer) achieved the best performance (RMSE 4.05 %BW, R^2 0.985), while for anteroposterior GRF, ResCorr (joint) achieved the lowest RMSE (1.68 %BW), though DS-Conformer attained higher waveform agreement (R^2 0.893). For mediolateral GRF, ResCorr (joint) again achieved the lowest RMSE (1.03 %BW), while DS-Conformer yielded the highest R^2 (0.809). No single architecture dominated all three axes. Age-group generalization testing showed that DS-Conformer maintained robust performance across the held-out age groups in this dataset with low inter-fold variability.

Conclusion: Continuous 3D GRF waveforms during STW can be accurately predicted using a single sacrum-mounted IMU, with axis-dependent architectural preferences that reflect the distinct biomechanical characteristics of each force component. Within the present evaluation,

DS-Conformer offers the best balance of accuracy and stability, though axis-specific model selection yields superior clean accuracy.

Significance: This study demonstrates that continuous GRF waveforms during STW transitions can be reconstructed from a single body-worn IMU, establishing a new capability for portable kinetic assessment of a clinically critical movement. The systematic architectural comparison provides evidence-based guidance for model selection in wearable biomechanics, while the age-group generalization and robustness evaluations address key prerequisites for clinical translation.

3.3 Introduction

Sit-to-walk (STW) is a critical activity in our daily lives, with adults performing it around 60 times per day (Dall & Kerr, 2010). This transition from sitting to walking plays a major role in maintaining stability during time-critical activities, such as moving from a chair to answer the phone or the doorbell. Research demonstrates that the ability to execute STW declines with age (Buckley et al., 2009; Perera et al., 2023). Impaired STW performance is associated with increased fall risk and reduced physical confidence, affecting both physical and psychological well-being (Buckley et al., 2009). Investigating the biomechanics of STW is important for assessing balance control during independent living, particularly in older adults (Buckley et al., 2009).

The Timed Up and Go (TUG) test is a widely used clinical assessment of STW performance (Chen & Chou, 2017). In this test, individuals rise from a seated position, walk 3 meters, turn, walk back, and sit down. A clinician times the entire sequence, with longer durations indicating potential mobility impairments. While TUG is simple and reliable, its single timing measure yields only a gross performance metric without insight into movement quality or mechanics (Higashi et al., 2008). To address this limitation, the instrumented TUG (iTUG) was developed, utilizing inertial measurement unit (IMU) sensors to capture detailed kinematic data, including joint angles, velocities, and accelerations (Ortega-Bastidas et al., 2023). However, existing iTUG research has primarily focused on kinematic parameters, which describe motion characteristics without quantifying the forces driving the movement. This oversight is particularly significant because forces and moments across lower limb joints during STW

transitions can reach up to 4.7 times bodyweight (Khemlani et al., 1999). Such high loads pose substantial challenges for people with reduced muscle strength (Hughes et al., 1996). These joint loads directly correlate with ground reaction forces (GRFs), which are the counterforces exerted by the ground during movement. Upward-directed GRFs create external moments about the joints that must be counteracted by internal muscular moments to maintain stability and control. Notably, recent research (Kera et al., 2020) has established a strong correlation between sit-to-stand (a subset of STW) GRF patterns and fall risk, highlighting the importance of kinetic assessments.

Advances in IMU technology have enabled machine learning-based estimation of GRFs across activities such as walking, running, and sports-specific movements, using methods ranging from k-Nearest Neighbors (KNN) regression to deep learning architectures like Long Short-Term Memory (LSTM) networks. Building on single-sensor GRF estimation during walking (Lim et al., 2019) and comparisons of CNN architectures (Johnson et al., 2021), the present study evaluates seven architectural families for continuous three-dimensional GRF estimation during STW. This extension is important because STW poses modeling challenges fundamentally different from those of steady-state locomotion. The transition involves an abrupt onset from a stationary posture, rapid loading transients as bodyweight transfers from the chair to the feet, non-stationary dynamics spanning the postural reorganization into locomotion, and longer-range temporal dependencies that link the rising phase to the initial gait. These characteristics demand models that can simultaneously capture local high-frequency events and global contextual structure across the entire movement sequence. Such requirements are not adequately tested by cyclic gait tasks, where existing IMU-to-GRF methods have been developed and validated.

This study compares models from seven architectural families, selected to test complementary approaches to local feature extraction, temporal memory, multiscale representation, feature fusion, residual correction, and global temporal context. Simple convolutional and recurrent baselines provide reference points for evaluating more complex combinations of these mechanisms. All models use the same dataset, preprocessing, subject-wise evaluation and performance metrics. Leave-one-age-group-out testing additionally examines transfer to young, middle-aged and older adults excluded from training. Two hypotheses guided the architectural and component-wise comparisons. First, architectures combining local temporal feature extraction with longer-range temporal context would provide better waveform

agreement than models relying primarily on a single mechanism. This hypothesis motivated the comparison of convolutional, recurrent, and attention-based architectures, without specifying one model as the anticipated best performer. Second, waveform agreement would be lower for the smaller-amplitude mediolateral GRF component than for the vertical component.

3.4 Materials and Methods

3.4.1 Dataset

We used a publicly available dataset containing lower-body IMU recordings and force-plate measurements during STW extracted from the forward portion of the TUG test (Perera et al., 2024). During the trials, participants began in a seated position on a backless, armless, height-adjustable stool with their hands resting on their laps and then performed TUG trials. The dataset comprises 65 healthy subjects, categorized into three standard age groups: (1) young adults aged 19-35 years; n = 32; mass = 64.12 ± 17.04 kg, (2) middle-aged adults aged 36-55 years; n = 15; mass = 67.94 ± 15.85 kg, and (3) older adults aged 56-73 years; n = 18; mass = 59.38 ± 10.7 kg with a total of 325 trials. All participants were healthy individuals capable of standing and walking comfortably without external aid and exhibited no movement impairments. Data collection was approved by the Monash University Human Research Ethics Committee under project number 32328. All subjects provided written informed consent.

For data acquisition, a wireless Delsys Trigno system (Delsys Inc., Boston, USA) was used that consists of an accelerometer (range ±16g, resolution 10bits) and a gyroscope (range ±2000 °/s, resolution 10bits) with sampling rate of 148Hz. Only the sacrum-mounted IMU sensor data was used in this study. The IMU sensor provided six data channels: three for acceleration along each axis and three for angular velocity along each axis. To capture GRF during STW, three force plates (FP4060-07, Bertec Corp., Ohio, USA) were installed flush within a wooden walkway to provide a flat walking surface. Each force plate measured 600 mm × 400 mm × 75 mm and sampled data at 1,000 Hz. An external hardware trigger was used to time synchronize IMU and force plate data.

3.4.2 Preprocessing

Force plate data was filtered to remove noise and motion artifact using a 4th-order low-pass Butterworth filter. The cutoff frequency was determined using fast Fourier transform (FFT) analysis, by identifying the frequency at which 99% of cumulative signal power was captured (Schilaty et al., 2021).

Figure 3-1 shows cumulative FFT power for each GRF axis for our dataset. 99% of the signal lies below 10 Hz, hence, a cutoff frequency of 10 Hz was applied to the low-pass filter.

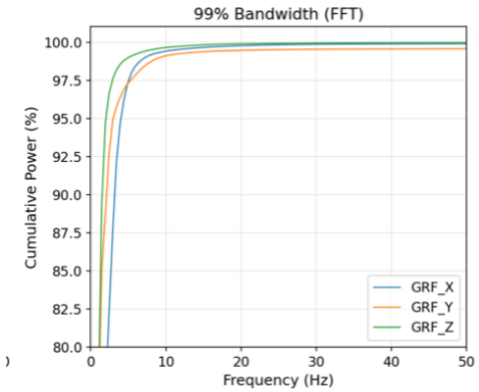


Figure 3-1 Cumulative FFT power for GRF used to determine low-pass filter cutoff frequency. The cumulative percentage of total signal power is plotted against frequency.

Following this, the force plate data was downsampled to 148 Hz to match the IMU sampling rate. GRF amplitudes were normalized by bodyweight (BW) and time-normalized to accelerate neural network convergence. The resulting paired input-output sequences comprised 6-channel IMU time series as input and the GRF waveform as the target output.

3.4.3 Model Architectures

Mapping IMU signals to GRF is a sequence-to-sequence regression task in which a multichannel input time series of accelerometer and gyroscope measurements must be transformed into a synchronized output GRF waveform at each timestep. This mapping is demanding in practice because the model must merge local, high-frequency events such as foot-contact transients and rapid loading-unloading cycles with longer-horizon structure, including phase progression and stride-to-stride consistency (Alcantara et al., 2022; Hochreiter

& Schmidhuber, 1997). Recent studies have demonstrated that deep sequential models can estimate GRF waveforms from IMU data during locomotion (Alcantara et al., 2022; Carter et al., 2024), though they also highlight sensitivity to subject novelty and speed variation, underscoring the need for architectures that learn robust, generalizable representations (Gurchiek et al., 2019; Scheltinga et al., 2023).

Our investigation proceeds as a progressive narrative, beginning with simple baselines and advancing through increasingly sophisticated architectures, each motivated by specific limitations observed in simpler designs. Figure 3-2 illustrates this architectural progression and the design limitations that motivate each transition between model families.

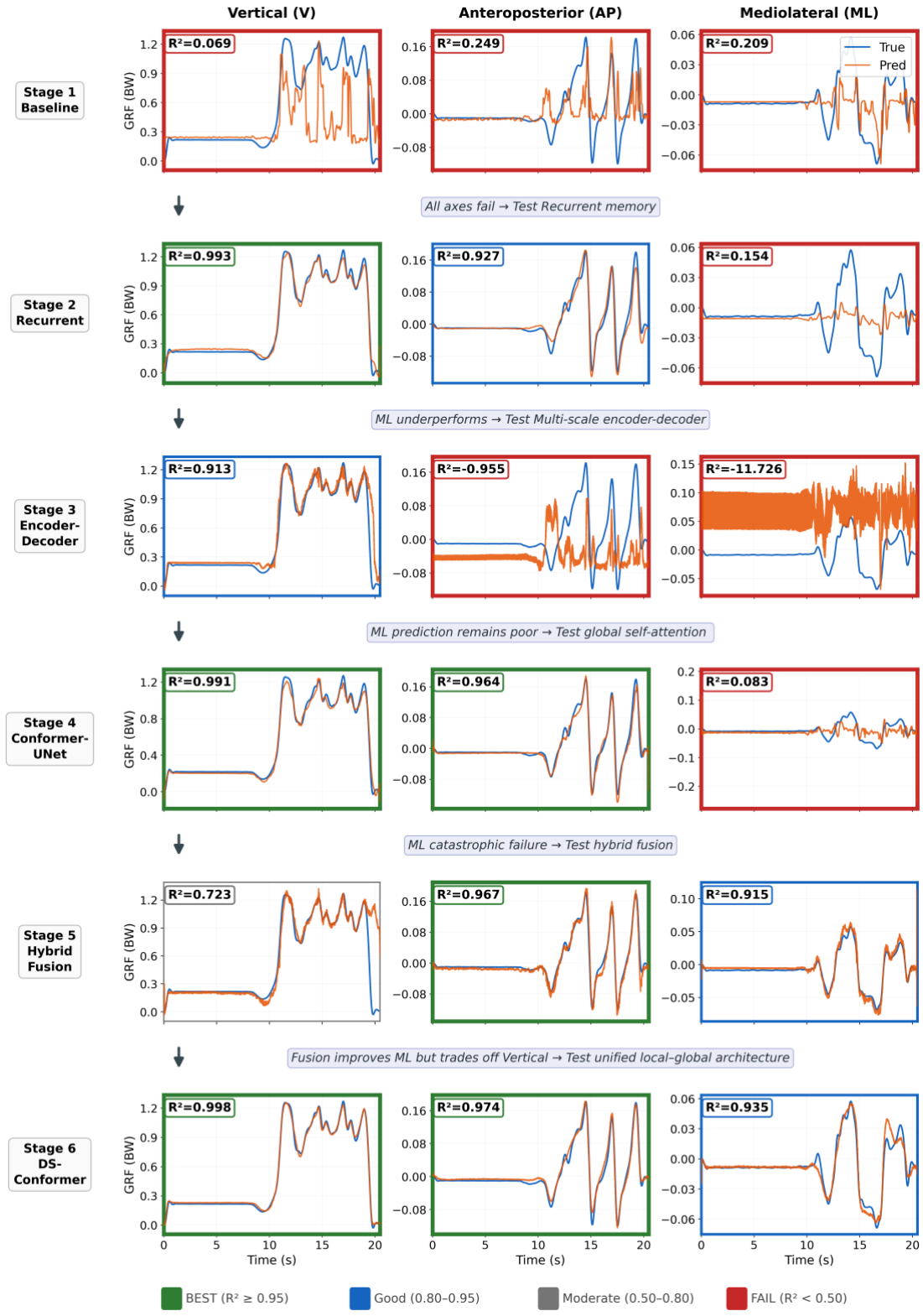


Figure 3-2 Architectural progression for 3D GRF prediction from a single representative fold. Each row represents a stage in the investigation, showing representative predicted (orange) versus true (blue) waveforms for vertical, AP, and ML GRF. Annotations between stages indicate the observed limitation motivating each transition. Border colors denote per-axis R^2 performance: green (≥ 0.95), blue (0.80-0.95), grey (0.50-0.80), and red (< 0.50).

Before presenting individual models, we identify core design principles along which the evaluated architectures differ. These principles directly address the challenges of mapping IMU signals to GRF waveforms during STW.

Temporal memory mechanisms: GRF prediction requires relating events across time, for example, connecting the moment of seat-off to peak loading hundreds of milliseconds later. Recurrent networks (LSTM, GRU) address this through gated memory cells that selectively retain or forget information as the sequence progresses (Cho et al., 2014; Hochreiter & Schmidhuber, 1997).

Local versus global feature extraction: IMU signals contain both localized events, such as impact transients and rapid accelerations, and longer-range patterns, such as phase progression through movement. Convolutional layers in CNNs excel at detecting local patterns through fixed-size filters, while self-attention mechanisms can relate any two timesteps regardless of their temporal distance. The Conformer architecture combines both within a single block (Gulati et al., 2020).

Multi-resolution processing: The STW transition exhibits structure across multiple temporal scales, ranging from millisecond-level impact dynamics to second-scale phase transitions. U-Net-style encoder-decoder networks address this by progressively compressing the signal to capture coarse structure, then expanding it to recover fine detail, using skip connections that preserve high-frequency information (Ronneberger et al., 2015).

Feature selection and fusion: When multiple sensor channels (accelerometer and gyroscope) provide complementary information, the network must learn which channels are most relevant at each timestep. Attention mechanisms and gating units provide learnable weighting schemes for this purpose (Rayatdoost et al., 2020).

For our experiment, we begin with single-mechanism baselines that isolate local feature extraction and temporal memory in turn, then systematically combine and extend these mechanisms, adding multiscale decomposition, attention-based refinement, recurrent-convolutional hybrids, two-stage residual correction, and finally self-attention to determine which inductive biases are most beneficial for IMU-to-GRF prediction during STW transitions.

1. Baseline Architectures

ShallowCNN: To isolate the value of local temporal feature extraction, we included a shallow CNN baseline. Shallow CNNs have been used in physiological signal domains for their ability to learn important features and filter out the noise (Liew et al., 2021).

Bidirectional Gated Recurrent Unit (BiGRU): To complement the CNN's local feature extraction with an explicit test of temporal memory, we evaluated bidirectional GRUs. BiGRU selectively retains or discards information over time, processing the input sequence in both forward and reverse directions so that each timestep is informed by both preceding and subsequent context. This enables the network to capture long-range dependencies that may manifest asymmetrically across the gait cycle (Cho et al., 2014; Gruber & Jockisch, 2020).

Bidirectional Long Short-Term Memory (BiLSTM): The BiLSTM model uses a gating mechanism similar to the BiGRU but includes a dedicated cell state, offering an extra path for information flow (Yu et al., 2019). We tested both bidirectional GRU and LSTM variants to assess whether the choice of recurrent cell type meaningfully affects prediction performance.

2. Temporal Convolutional Networks (TCN)

While the GRU captures long-range dependencies, recurrent networks process inputs sequentially, limiting parallelism and risking memory degradation over very long sequences (Johnston et al., 2025). TCNs offer an alternative mechanism, by preserving convolution's local inductive bias but employing dilated convolutions and residual connections to achieve long effective memory without sequential processing (Bai et al., 2018). For IMU-to-GRF modeling, TCNs are attractive because exponentially increasing dilation factors provide long effective memory without degradation.

3. MultiResUNet Variants

MultiResUNet: TCNs expand temporal reach through dilation while maintaining a single resolution throughout the network, thereby limiting explicit multiscale decomposition. To

directly test the multi-resolution processing principle, we evaluated MultiResUNet, which is a fully convolutional encoder-decoder architecture that extends U-Net (Ronneberger et al., 2015) by incorporating multiresolution blocks to capture features across varying temporal scales (Ibtehaz & Rahman, 2019). Skip connections preserve fine-grained temporal detail through decoding and bypass the encoder's information bottleneck, which proves essential for GRF synthesis.

MRU-AttnGRU: To test whether combining attention-guided feature selection with explicit temporal memory improves prediction, we developed MRU-AttnGRU as a multi-resolution encoder-decoder architecture with attention-gated skip connections and a bidirectional GRU head. The attention gates were included to emphasize task-relevant features while suppressing noisy or irrelevant activations (Faisal et al., 2023), and the BiGRU was used to capture longer-range temporal dependencies after multi-resolution feature extraction. This design therefore isolates the contribution of recurrent temporal modeling within an attention-gated multi-resolution framework.

4. Hybrid Architectures

MultiResUNet provides strong multiscale representations but lacks a dedicated temporal memory mechanism. The bottleneck block in MultiResUNet may suppress long-distance dependencies. Hybrid models address this by combining dual-stream, multi-resolution encoding with a bidirectional GRU head and a principled modality fusion strategy (Cho et al., 2014). We evaluated two fusion strategies: Hybrid(Concat), where stream features are concatenated, and the GRU learns implicit channel weighting, and Hybrid(Gated), where a Gated Multimodal Unit (Arevalo et al., 2020) computes per-timestep weights to explicitly learn which modality is most informative at each point in the movement cycle (Kim et al., 2019).

5. Residual-Correction Pipelines

Single-stage pipelines sometimes fail when reproducing both sharp transients and smooth low-frequency trends. Based on stacked generalization (Wolpert, 1992) and residual learning (He

et al., 2016) we decomposed prediction into two stages: a MultiResUNet generates an initial waveform estimate, and a TCN correction network refines the residual errors using dilated convolutions (Bai et al., 2018). The jointly trained variant, ResCorr(joint), trains both stages end-to-end, while ResCorr(frozen) locks the MultiResUNet weights after initial training and optimizes only the TCN correction network, testing whether the correction stage benefits from adapting the primary model's representations or can operate independently on fixed features.

6. Downsampled Transformer

Architectures above implement temporal dependency modeling via recurrence and/or convolution. A Transformer encoder instead uses self-attention, which can model dependencies between any two timesteps regardless of their distance. However, standard attention scales quadratically with sequence length, which is problematic for high-sample-rate IMU data (Vaswani et al., 2017). Downsampling and hierarchical time-series Transformer designs mitigate this by reducing the effective sequence length while maintaining informative representations (Liu et al., 2021; Nie et al., 2022).

7. Conformer-Based Architectures

A known limitation of pure Transformers is the absence of local inductive bias. Self-attention treats all timestep pairs equally, which can hinder the learning of localized features, such as impact transients, without large training sets. The Conformer architecture (Gulati et al., 2020) was designed to resolve this by combining self-attention for content-based global interactions with depthwise convolution for local feature extraction within a single block.

ConformerUNet. Integrates Conformer blocks at the bottleneck of a U-shaped encoder-decoder, testing whether combining local-global modeling with skip-connection-based multi-resolution reconstruction outperforms either mechanism alone (Burchi & Vielzeuf, 2021).

Downsampled Conformer (DS-Conformer). The DS-Conformer shares the downsample-process-upsample structure of the DS-Transformer but replaces Transformer blocks with

Conformer blocks, reintroducing local inductive bias without an explicit decoder or skip connections (Burchi & Vielzeuf, 2021).

3.4.4 Loss Functions

All models were trained using a two-phase optimization strategy designed to leverage complementary strengths of different loss functions. In the first phase, each model was trained from random initialization using mean absolute error (MAE) as the primary objective function, optimized with the Adam optimizer. MAE was selected for the initial training phase because it is more robust to outliers than mean squared error (MSE). GRF waveforms during STW contain sharp loading transients and peak force events that can dominate an MSE-based objective. L1 loss provides more uniform gradient contributions across all timesteps, preventing the optimizer from disproportionately fitting high-amplitude regions at the expense of lower-amplitude phases, such as the seated baseline or early weight transfer.

In the second phase, the best checkpoint from Phase 1 was fine-tuned using the Pearson correlation coefficient (r) loss function. The rationale for the two-phase approach is that MAE provides a stable training landscape for learning coarse signal structure, after which the correlation-aware loss refines waveform fidelity. Because Pearson r is invariant to scale and offset, amplitude drift during fine-tuning can be a concern. We mitigated this risk in practice by three factors: initialization from an amplitude-calibrated MAE checkpoint, a reduced learning rate applied over limited epochs, and continued monitoring of validation MAE to confirm that magnitude accuracy was preserved.

3.4.5 Validation Protocol

Model performance was evaluated using five-fold subject-wise cross-validation, where all trials from a given participant appeared exclusively in either the training or test set within each fold to prevent within-subject data leakage. Within each training partition, 20% of subjects were held out as a validation set for early stopping and checkpoint selection.

All models were trained with the Adam optimizer. Each fold was trained from a fresh random initialization. Final metrics were computed on held-out test subjects and aggregated across folds. Hyperparameters for each model were tuned using Optuna (Akiba et al., 2019) with 40 trials per model, selected once on a development split from the training data, and then fixed across all cross-validation folds. The held-out test subjects in each fold were used only for final evaluation and were not used during hyperparameter selection. Parameter counts and GPU inference costs are reported in Supplementary Material Section B6.

1. Evaluation Metrics

The following evaluation metrics were reported:

Root mean square error (RMSE) was computed between predicted and measured GRF waveforms to quantify the average magnitude of prediction errors, with results reported as a percentage of bodyweight (%BW) to facilitate cross-study comparison.

Normalized RMSE (NRMSE) was computed by dividing RMSE by the range of the measured GRF waveform, yielding a scale-independent accuracy measure expressed as a percentage.

Mean absolute error (MAE) provides a complementary error measure that is less sensitive to outlier timesteps than RMSE.

The coefficient of determination (R^2) compares squared prediction error with the error obtained by predicting the mean measured value at every evaluated timestep:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where (y_i) and $(\hat{y}_i)$ are the measured and predicted values, and $(\bar{y})$ is the mean measured value over the evaluated samples. For a nonconstant measured signal, R^2 has a maximum of 1 and no finite lower bound. A negative value indicates greater squared error than the constant mean-value reference. Predictive R^2 is not generally equal to squared Pearson correlation, because it is sensitive to amplitude and offset errors as well as waveform association.

For the five-fold evaluation, R^2 was calculated separately for each subject after concatenating the samples from that subject's evaluated trials. For the leave-one-subject-out evaluation, R^2 was calculated for each trial and then averaged within each subject. Results were summarized as the mean and sample standard deviation across subjects. These aggregation procedures are not mathematically equivalent and should be considered when comparing R^2 between the two evaluation protocols.

Pearson's correlation coefficient (r) measures the linear association between predicted and measured waveforms, capturing how well the model reproduces overall waveform morphology independent of amplitude scaling.

Lin's concordance correlation coefficient (CCC) was reported because it simultaneously penalizes deviations in both correlation and bias, unlike Pearson's r, which is insensitive to constant offsets or scaling errors. CCC is defined as the product of the Pearson correlation and a bias correction factor, and ranges from −1 to +1, where +1 indicates perfect agreement.

Together, these metrics provide a comprehensive evaluation of the model's performance.

3.4.6 Age-Group Generalization Experiment

The five-fold cross-validation protocol described in Section 3.4.5 evaluates generalization to unseen subjects drawn from the same population distribution. However, younger and older adults exhibit systematically different STW kinematics and kinetics (Buckley et al., 2009; van der Kruk et al., 2022). To test whether the model can generalize across these age-related differences, an age-group holdout experiment was conducted in which entire age cohorts were withheld from training and used exclusively for testing. The 65 participants were split into three age groups: young adults (19-35 years; n = 32), middle-aged adults (36-55 years; n = 15), and older adults (56-73 years; n = 18). Age groups are unequally sized, and the number of training subjects varies across splits, so differences in held-out performance may partly reflect class imbalance rather than demographic mismatch alone.

Three leave-one-age-group-out splits were evaluated: (1) train on young + middle-aged, test on older adults; (2) train on middle-aged + older adults, test on young adults; and (3) train on young + older adults, test on middle-aged adults. In each split, all trials from the held-out age group were reserved entirely for testing, ensuring zero overlap of age demographics between

training and evaluation. Within each training partition, a subject-wise validation set comprising 20% of the training subjects was carved out to enable early stopping and checkpoint selection, in line with the protocol in Section 3.4.5. All training hyperparameters, loss functions, and the two-phase optimization strategy were kept identical to those used in the cross-validation experiments to ensure direct comparability.

3.5 Results

This section presents the prediction performance of the 13 model configurations across all three GRF axes. Section 3.5.1 reports five-fold cross-validated waveform accuracy for vertical, AP, and ML GRF components. Section 3.5.2 presents Leave-One-Out Cross-Validation Results, and Section 3.5.3 examines age-group generalization to assess whether models trained on one demographic can accurately predict GRF waveforms for held-out age groups.

3.5.1 5-Fold Cross-Validated Waveform Prediction

Each of the 13 architectures was trained and evaluated using five-fold subject-wise cross-validation on 65 participants. Performance is presented separately for each GRF axis.

1. Vertical GRF Prediction

Table 3-1 presents the cross-validated waveform metrics for the 13 model configurations that predict vertical GRF during STW. Values are mean ± SD across 65 subjects aggregated over all folds. Models are ranked by RMSE (%BW) (lower is better). Bold indicates the best value per column. R^2, Pearson r, and CCC are higher-is-better metrics.

Table 3-1 Five-fold subject-wise cross-validation vertical GRF results for 13 model configurations

Model	RMSE (%BW)	NRMSE %	MAE (%BW)	R^2	Pearson r	CCC
DS-Conformer	**4.05 ± 2.81**	**2.90 ± 1.98**	**2.80 ± 1.75**	**0.985 ± 0.024**	**0.994 ± 0.009**	**0.992 ± 0.012**
ConformerUNet	4.44 ± 3.84	3.15 ± 2.49	3.01 ± 2.16	0.978 ± 0.062	0.991 ± 0.026	0.989 ± 0.030
TCN	5.59 ± 3.34	4.00 ± 2.17	3.78 ± 1.62	0.975 ± 0.039	0.989 ± 0.017	0.987 ± 0.018
MultiResUNet	7.00 ± 6.41	5.04 ± 4.59	4.34 ± 3.19	0.944 ± 0.152	0.977 ± 0.058	0.974 ± 0.068
BiGRU	7.16 ± 5.07	5.10 ± 3.21	4.75 ± 2.75	0.955 ± 0.090	0.980 ± 0.042	0.977 ± 0.045
Hybrid-Concat	7.53 ± 5.07	5.46 ± 3.75	4.81 ± 2.63	0.950 ± 0.081	0.980 ± 0.034	0.975 ± 0.040
Hybrid-Gated	7.96 ± 10.40	5.67 ± 7.08	4.13 ± 2.70	0.901 ± 0.457	0.974 ± 0.070	0.969 ± 0.085
DS-Transformer	9.45 ± 4.36	6.77 ± 2.90	5.24 ± 2.09	0.934 ± 0.071	0.969 ± 0.034	0.967 ± 0.035
ResCorr (frozen)	11.86 ± 9.63	8.59 ± 6.90	7.14 ± 6.34	0.863 ± 0.249	0.944 ± 0.096	0.930 ± 0.125
MRU-AttnGRU	12.39 ± 11.73	8.91 ± 8.45	9.52 ± 9.95	0.823 ± 0.286	0.961 ± 0.054	0.937 ± 0.091
ResCorr (joint)	16.26 ± 18.86	11.70 ± 13.58	12.75 ± 16.43	0.627 ± 0.653	0.925 ± 0.122	0.893 ± 0.174
BiLSTM	27.15 ± 11.43	19.78 ± 8.64	18.24 ± 10.91	0.485 ± 0.516	0.758 ± 0.222	0.705 ± 0.248
ShallowCNN	29.92 ± 9.85	21.81 ± 7.51	18.89 ± 9.87	0.407 ± 0.463	0.700 ± 0.190	0.668 ± 0.217

The DS-Conformer achieved the best overall performance with RMSE 4.05 %BW, R^2 0.985. The Conformer-based architectures dominated the vertical axis, with ConformerUNet ranking second (RMSE 4.44 %BW). Among recurrent architectures, BiGRU achieved respectable performance (RMSE 7.16 %BW), whereas ShallowCNN performed poorly (RMSE 29.92 %BW), confirming that minimal local filtering is insufficient to capture the complex temporal structure of vertical GRF during STW.

2. AP GRF Prediction

Table 3-2 presents results for AP GRF prediction. The ResCorr (joint) architecture achieved the lowest mean RMSE (1.68 %BW). DS-Conformer ranked second (RMSE 1.75 %BW), followed by BiGRU (RMSE 2.41 %BW). However, DS-Conformer attained higher correlation metrics across all three agreement measures (R^2 = 0.893 vs. 0.886; Pearson r = 0.949 vs. 0.946; CCC = 0.942 vs. 0.938), suggesting superior waveform shape preservation.

Table 3-2 Five-fold subject-wise cross-validation AP GRF results for 13 model configurations.

Model	RMSE (%BW)	NRMSE (%)	MAE (%BW)	R^2	Pearson r	CCC
ResCorr (joint)	**1.68 ± 1.34**	**3.93 ± 2.64**	**1.12 ± 0.74**	0.886 ± 0.177	0.946 ± 0.087	0.938 ± 0.096
DS-Conformer	1.75 ± 1.07	4.16 ± 2.09	1.19 ± 0.70	**0.893 ± 0.126**	**0.949 ± 0.068**	**0.942 ± 0.079**
BiGRU	2.41 ± 0.99	5.80 ± 1.84	1.65 ± 0.62	0.821 ± 0.136	0.916 ± 0.065	0.902 ± 0.077
DS-Transformer	2.64 ± 1.40	6.28 ± 2.66	1.62 ± 0.90	0.771 ± 0.234	0.882 ± 0.133	0.871 ± 0.141
Hybrid-Gated	2.65 ± 2.16	6.33 ± 5.11	1.77 ± 1.38	0.698 ± 0.403	0.784 ± 0.334	0.763 ± 0.362
Hybrid-Concat	2.66 ± 1.78	6.41 ± 4.09	1.72 ± 1.09	0.720 ± 0.301	0.841 ± 0.188	0.810 ± 0.218
TCN	2.70 ± 4.99	5.99 ± 7.40	1.53 ± 0.74	0.418 ± 3.610	0.925 ± 0.116	0.915 ± 0.128
ConformerUNet	2.92 ± 1.91	7.09 ± 4.47	1.94 ± 1.13	0.663 ± 0.416	0.760 ± 0.363	0.749 ± 0.369
ResCorr (frozen)	3.10 ± 1.52	7.45 ± 2.41	2.00 ± 0.53	0.699 ± 0.381	0.864 ± 0.096	0.846 ± 0.108
MultiResUNet	3.95 ± 2.69	9.67 ± 6.73	2.85 ± 2.11	0.374 ± 0.656	0.625 ± 0.409	0.587 ± 0.431
BiLSTM	4.53 ± 0.74	11.04 ± 1.43	2.96 ± 0.46	0.424 ± 0.144	0.679 ± 0.120	0.583 ± 0.136
ShallowCNN	5.00 ± 0.77	12.17 ± 1.35	3.14 ± 0.49	0.304 ± 0.146	0.572 ± 0.112	0.497 ± 0.109
MRU-AttnGRU	7.02 ± 1.48	17.14 ± 3.47	5.00 ± 1.58	-0.409 ± 0.613	0.080 ± 0.188	0.028 ± 0.063

3. ML GRF Prediction

Table 3-3 presents results for ML GRF prediction. ML forces are the smallest-amplitude GRF component during STW, making this the most challenging prediction axis. ResCorr (joint) achieved the best performance (RMSE 1.03 %BW). DS-Conformer again yielded higher correlation metrics (R^2 = 0.809 vs. 0.545; Pearson r = 0.915 vs. 0.909; CCC = 0.902 vs. 0.897). ResCorr (joint) had a mean R^2 of 0.545 (SD 2.309), with negative values in 3 of 65 subjects and a minimum of −17.633. DS-Conformer had a mean R^2 of 0.809 (SD 0.169), with no negative subject-level values. All UNet-based models underperformed, indicating that conformer-based encoder-decoder structures may overfit to larger-magnitude signals and fail on the smaller ML component.

Table 3-3 Five-fold subject-wise cross-validation ML GRF results for 13 model configurations.

Model	RMSE (%BW)	NRMSE (%)	MAE (%BW)	R^2	Pearson r	CCC
ResCorr (joint)	**1.03 ± 1.50**	**5.97 ± 6.92**	**0.61 ± 0.47**	0.545 ± 2.309	0.909 ± 0.179	0.897 ± 0.190
DS-Conformer	1.09 ± 0.39	6.75 ± 2.14	0.71 ± 0.23	**0.809 ± 0.169**	**0.915 ± 0.084**	**0.902 ± 0.102**
DS-Transformer	1.59 ± 0.61	9.93 ± 3.59	0.95 ± 0.33	0.591 ± 0.322	0.772 ± 0.204	0.747 ± 0.214
TCN	1.81 ± 4.98	10.20 ± 21.98	0.83 ± 0.43	-2.723 ± 28.267	0.884 ± 0.127	0.866 ± 0.141
Hybrid-Gated	2.15 ± 1.19	13.52 ± 7.66	1.51 ± 0.94	0.109 ± 0.861	0.591 ± 0.362	0.568 ± 0.375
ConformerUNet	2.31 ± 0.74	14.64 ± 4.74	1.60 ± 0.55	0.181 ± 0.379	0.460 ± 0.265	0.343 ± 0.294
Hybrid-Concat	2.34 ± 0.80	14.58 ± 4.81	1.55 ± 0.56	0.197 ± 0.335	0.419 ± 0.284	0.249 ± 0.336
ResCorr (frozen)	2.40 ± 0.49	15.08 ± 2.81	1.60 ± 0.35	0.187 ± 0.279	0.504 ± 0.202	0.341 ± 0.255
BiLSTM	2.53 ± 0.38	15.90 ± 2.31	1.71 ± 0.36	0.108 ± 0.283	0.492 ± 0.161	0.291 ± 0.131
ShallowCNN	2.55 ± 0.35	16.03 ± 1.93	1.67 ± 0.29	0.106 ± 0.164	0.411 ± 0.131	0.258 ± 0.091
BiGRU	2.63 ± 0.38	16.48 ± 2.01	1.77 ± 0.30	0.066 ± 0.075	0.400 ± 0.138	0.131 ± 0.043
MultiResUNet	4.33 ± 2.75	27.15 ± 17.52	3.55 ± 2.58	-2.662 ± 4.464	0.213 ± 0.376	0.196 ± 0.372
MRU-AttnGRU	4.91 ± 1.05	31.00 ± 7.36	4.16 ± 1.12	-2.529 ± 1.718	0.023 ± 0.123	0.009 ± 0.054

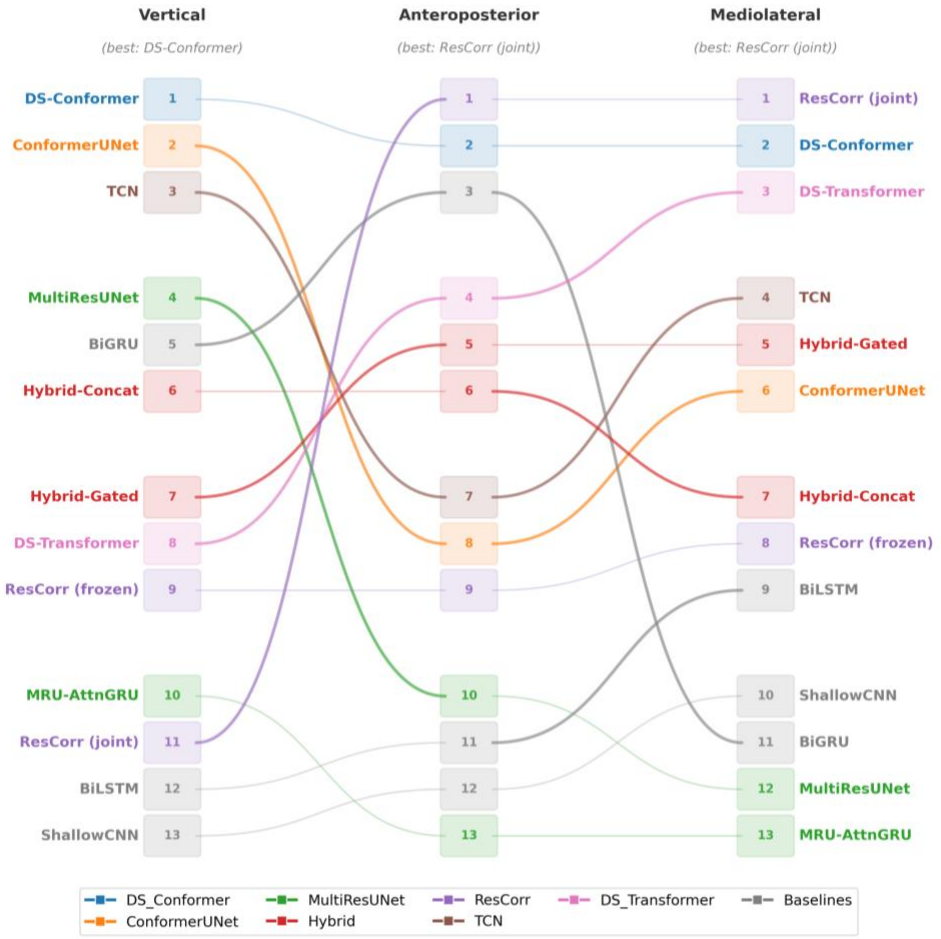


Figure 3-3 Model Rank flow across GRF axes based on RMSE. Rank lines cross extensively between axes, with DS-Conformer and ResCorr (joint) being the only models remaining in the top three on at least two axes.

3.5.2 Leave-One-Out Cross-Validation Results

To provide a more rigorous test of inter-subject generalization, leave-one-out cross-validation (LOOCV) was performed using the top two performing models identified in Section 3.5.1. In each of the 65 folds, all trials from a single subject were held out for testing, while the remaining 64 subjects were used for training, ensuring complete subject independence and eliminating fold-level averaging effects that might mask individual prediction failures.

Table 3-4 Leave-one-out cross-validation results for the top two performing models for each GRF axis.

Component	Model	RMSE (%BW)	NRMSE (%)	MAE (%BW)	R^2	Pearson r	CCC
GRF (V)	DS-Conformer	6.80 ± 5.14	4.92 ± 3.66	4.66 ± 2.86	0.950 ± 0.086	0.980 ± 0.035	0.9737 ± 0.0502
	ConformerUNet	6.79 ± 5.33	4.90 ± 3.75	4.72 ± 2.88	0.951 ± 0.094	0.981 ± 0.040	0.9753 ± 0.0472
GRF (AP)	DS-Conformer	2.40 ± 0.83	5.86 ± 1.79	1.77 ± 0.57	0.809 ± 0.146	0.918 ± 0.079	0.8999 ± 0.0873
	ResCorr (joint)	2.66 ± 1.24	6.44 ± 2.39	1.82 ± 0.55	0.743 ± 0.361	0.895 ± 0.096	0.8760 ± 0.1181
GRF (ML)	DS-Conformer	1.35 ± 0.40	8.50 ± 2.42	0.90 ± 0.26	0.697 ± 0.209	0.860 ± 0.157	0.8380 ± 0.1609
	ResCorr (joint)	1.52 ± 1.15	9.39 ± 5.67	0.95 ± 0.37	0.333 ± 2.164	0.848 ± 0.157	0.8144 ± 0.1859

Table 3-4 shows that under the stricter leave-one-out cross-validation protocol, both leading models retained very high accuracy for vertical GRF prediction. However, DS-Conformer provided the more consistent subject-independent performance across all three axes and outperformed the comparator models for both AP and ML GRF.

3.5.3 Age-Group Generalization

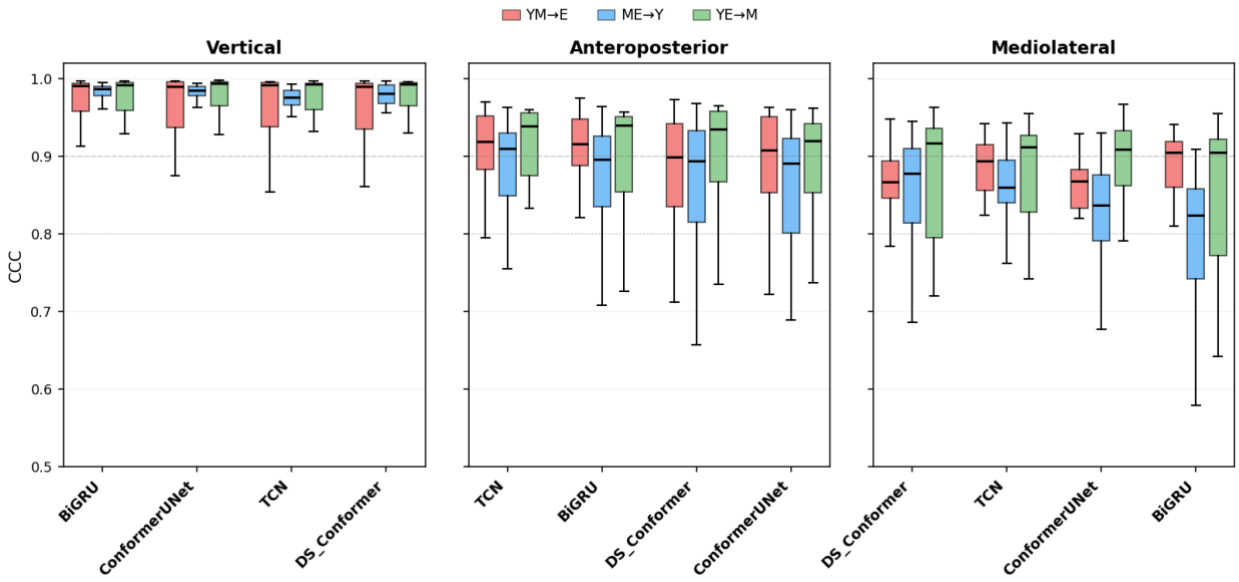


Figure 3-4 Per-subject concordance correlation coefficient (CCC) distributions for the top four models on each GRF axis under the age-group holdout protocol. Box plots show median line, interquartile range (box), and 1.5×IQR whiskers across subjects.

Figure 3-4 shows a clear and consistent axis hierarchy, V > AP > ML. Vertical GRF is near-ceiling for all four models with tight boxplots clustered around ~0.97-0.99 CCC, indicating strong cross-age robustness with minimal between-subject spread. In contrast, AP performance drops (medians 0.88-0.93) and the distributions widen, reflecting reduced stability with age shift. ML is the most challenging, with much lower medians (0.84-0.91) and the largest variability. Overall, DS-Conformer maintains consistent holdout distributions across all directions.

3.6 Discussion

This study demonstrates that continuous 3D GRF waveforms during STW transitions can be accurately predicted from a single sacrum-mounted IMU. The systematic comparison of 13

model configurations across seven architectural families revealed axis-dependent performance patterns, with DS-Conformer emerging as the most consistently high-performing and robust architecture. In this section, we interpret these findings in relation to the biomechanical characteristics of each GRF axis and compare performance with prior IMU-to-GRF work on other movement tasks.

3.6.1 Axis-Dependent Architectural Preferences

The extensive rank crossings visible in Figure 3-3 confirm that no single architecture dominates all three force components. For instance, BiGRU ranks 5th on vertical but drops to 11th on ML, while TCN rises from 3rd on vertical to 4th on ML despite falling to 7th on AP. These rank shifts reflect fundamental biomechanical differences between axes that impose distinct requirements on model inductive biases, interpretable through the design principles introduced in section 3.4.3. DS-Conformer emerges as the most consistently high-performing architecture overall.

Vertical GRF is dominated by the bodyweight support pattern: a seated baseline near zero, a rapid loading transient as bodyweight transfers to the feet, peak force reaching approximately 1.2 BW, and a transition into the oscillatory pattern of gait. This large-amplitude waveform combines clear global temporal structure with sharp local transients at seat-off and initial contact. DS-Conformer excelled here (RMSE = 4.05 %BW, $R^2 = 0.985$) because its Conformer blocks combine self-attention for the global loading-unloading envelope with depthwise convolution for local impact transients. ConformerUNet ranked second ($R^2 = 0.978$), confirming that coupling Conformer modeling with U-Net skip connections further preserves high-frequency detail. The large signal amplitude provides a strong learning signal that rewards long-range temporal modeling, while ShallowCNN's poor performance ($R^2 = 0.407$) confirms that minimal local filtering is insufficient for this axis.

AP forces reflect braking and propulsion dynamics during the transition into forward locomotion and are substantially smaller in magnitude. A notable finding is the divergence between error and fidelity metrics: ResCorr (joint) achieved the lowest RMSE (1.68 %BW) while DS-Conformer attained a higher R^2 (0.893 vs. 0.886). This divergence arises because the two-stage residual correction pipeline is explicitly designed to minimize pointwise residual

errors through its secondary TCN refinement, which can reduce absolute errors at individual timesteps. However, this local error minimization does not guarantee preservation of global waveform morphology. DS-Conformer, by contrast, processes the entire sequence through integrated attention-convolution blocks that inherently maintain temporal coherence, yielding superior shape agreement.

ML forces are the smallest-amplitude GRF component, reflecting lateral weight transfer and balance control. ResCorr (joint) achieved the lowest RMSE (1.03 %BW), but its subject-level R^2 values were more variable (mean 0.545, SD 2.309) than those of DS-Conformer (RMSE 1.09 %BW; mean R^2 0.809, SD 0.169). ResCorr produced negative R^2 values in 3 of 65 subjects, indicating greater squared error than the constant mean-value reference for those subjects. Low absolute error can coexist with poor R^2 when the measured signal has limited variation, because R^2 evaluates error relative to that variation. MultiResUNet and MRU-AttnGRU also had negative mean R^2 values (−2.662 and −2.529), indicating poor average performance relative to their subject-specific mean references. These scores identify prediction limitations but do not establish which architectural component caused them.

3.6.2 Architectural Component Analysis

Beyond axis-level rankings, the 13-architecture comparison isolates each design principle's contribution to STW GRF synthesis. Local convolutional filtering alone is insufficient: ShallowCNN collapses on the vertical GRF (R^2 = 0.407) and fails on both horizontal axes, confirming that a short receptive field cannot capture the multi-second loading envelope spanning seat-off through stance. Adding recurrent memory to BiGRU and BiLSTM recovers vertical accuracy but does not generalize across axes, with BiGRU dropping to R^2 = 0.066 on ML. Replacing sequential recurrence with dilated convolution, as in TCN, yields comparable vertical GRF performance but offers no advantage in AP or ML, indicating that long effective memory alone is not the missing ingredient. Both recurrent and dilated-convolutional models expand the temporal receptive field, but neither injects the inductive biases needed for small-amplitude, subject-variable components (Lim & Zohren, 2021). Multi-resolution decomposition, as implemented in MultiResUNet and its attention-gated variant (MRU-AttnGRU) fails on ML (R^2 = −2.662 and −2.529, respectively), although multi-scale filtering

was introduced to enrich feature diversity over the standard U-Net (Ibtehaz & Rahman, 2019). The aggressive strided downsampling inherent to the encoder removes the low-amplitude information needed for the smallest GRF component. ResCorr (joint) achieved the lowest RMSE on AP and ML GRF, but its ML R^2 varied substantially across subjects (SD 2.309), with negative values in 3 of 65 subjects. Global attention on its own also falls short: DS-Transformer improves over recurrent and convolutional baselines on vertical GRF but lacks the local inductive bias needed to capture the sharp transients at seat-off and initial contact. Replacing Transformer blocks with Conformer blocks, which integrate depthwise convolution alongside self-attention within a single block (Gulati et al., 2020), restores the local bias without losing global context. DS-Conformer and ConformerUNet are the only architectures that remain in the top three on at least two axes, and the separation between them is informative. ConformerUNet matches DS-Conformer on vertical GRF, where U-Net skip connections help preserve high-frequency detail, but loses ground on AP under LOOCV (Section 3.5.2), suggesting that skip connections offer diminishing returns on smaller-amplitude axes and can reintroduce variance under strict subject independence. Taken together, these patterns support a minimal architecture that combines self-attention with depthwise convolution inside a single block and applies downsampling conservatively enough to preserve low-amplitude signals. DS-Conformer is the instance of this principle that best satisfies all three axes.

3.6.3 Comparison with Prior IMU-to-GRF Studies

Table 3-5 Comparison of IMU-to-GRF prediction accuracy across studies, movement tasks, and sensor configurations. Results for the present study (Ours) are reported under leave-one-out cross-validation to ensure subject-independent evaluation consistent with other studies. A dash (-) indicates no metric was reported for that axis. "r" is Pearson Correlation Coefficient.

Study	Subjects	Task	IMUs	Model	Vertical GRF Accuracy	AP GRF Accuracy	ML GRF Accuracy
(Alcantara et al., 2022)	19	Running	1	Bi-LSTM	RMSE: 16% BW	-	-
(Schelting a et al., 2023)	12	Running	3	Hybrid ANN ensemble	RMSE: 18% BW, r: 0.97	RMSE: 7% BW, r: 0.91	RMSE: 5% BW, r: 0.58
(Dorschky et al., 2020)	10	Walking + running	7	CNN	RMSE: 3% BW	RMSE: 10% BW	-
Ours	65	Sit-to-Walk	1	DS-Conformer	RMSE 6.80% BW, r = 0.980	RMSE 2.40% BW, r = 0.918	RMSE 1.35% BW, r = 0.860

Table 3-5 provides context by comparing our results with related locomotion-based studies. These cross-study comparisons should be interpreted cautiously, as they are not like-for-like due to differences in task, population, number and placement of IMUs, dataset size, and evaluation protocol. Therefore, Table 3-5 is intended to contextualize, rather than directly benchmark the model performance. Within this broader context, our DS-Conformer, using a single sacrum IMU, achieved strong accuracy across all axes, suggesting competitive performance despite using only one sensor and the added complexity of a non-cyclic STW task.

3.6.4 Generalizability

For clinical translation, models must generalize across demographic groups with systematically different movement characteristics.

As shown in Figure 3-4 for DS-Conformer, the worst-performing split varies by axis and reflects an interplay between training set size and biomechanical specificity. YM→E denotes a model trained on Young + Middle-aged and tested on Older adults. On the vertical axis, ME→Y shows a marginally greater spread because its training partition (middle-aged + older, n = 33) is the smallest, a notable drop from the full 5-fold setting (CCC 0.992), where all 52 training subjects are available per fold. The same training-size limitation also explains why ME→Y is the weakest split on the AP axis, where reduced training diversity appears to limit the model's ability to learn the full range of braking and propulsion strategies, thereby degrading the 5-fold CCC to 0.942. On the ML axis, however, the pattern shifts: YM→E becomes the weakest split despite having a larger training set (n = 47). Here, training set size is no longer the bottleneck. Instead, the challenge is biomechanical extrapolation. The model trained on young and middle-aged adults has never seen the compensatory lateral balance strategies of older adults (n = 18 test subjects), and because ML forces are the most age-sensitive component (Schrager et al., 2008), this demographic gap outweighs the training-size advantage. This contrasts with the 5-fold CCC of 0.902, in which elderly subjects are always included in the training set, allowing the model to learn age-specific lateral weight-shifting patterns directly.

LOOCV across all 65 subjects provided an additional test of inter-subject generalizability within the same dataset, preventing the fold-averaging that can hide individual prediction failures. Under full subject independence, DS-Conformer maintained strong vertical GRF prediction (R^2 = 0.950, RMSE = 6.80 ± 5.14 %BW). On the AP axis, DS-Conformer retained the highest waveform fidelity (R^2 = 0.809, r = 0.918), whereas ResCorr (joint) showed more severe degradation (R^2 = 0.743). ML performance degraded the most (R^2 = 0.697), consistent with this axis being the most subject-specific and biomechanically variable component. Notably, ConformerUNet matched DS-Conformer on vertical GRF under LOOCV (R^2 = 0.951 vs 0.950; RMSE 6.79 vs 6.80 %BW), whereas DS-Conformer retained a clear lead on AP, suggesting that U-Net skip connections provide comparable vertical prediction capacity but offer less benefit for the smaller-amplitude AP waveform under strict generalization conditions.

3.6.5 Limitations

Several limitations should be considered. First, all models were evaluated using a single dataset of healthy participants performing TUG tests, so the subject-wise results assess prediction for new participants under the same acquisition conditions. Generalization to clinical populations with gait impairments, neurological conditions, or musculoskeletal disorders remains untested. Transfer to other sensors, laboratories or recording setups was also not evaluated. Second, the dataset comprises 65 subjects with 325 trials, which is relatively small for deep learning. The fold instability observed in several architectures may partially reflect overfitting to this sample size. This is corroborated by the LOOCV vertical GRF RMSE standard deviation of ±5.14 %BW, which quantifies the per-subject prediction variability that fold-averaged results obscure. Third, the experimental setup used a backless, armless stool, which does not reflect the real-world diversity of STW initiation conditions. Chair height, armrest use, and seat compliance all influence the kinetic profile during the rising phase and may alter the IMU signal characteristics on which the models were trained. Fourth, all trials were performed at a self-selected comfortable pace on level ground, limiting speed and strategy variation. Clinical populations often exhibit altered movement strategies, asymmetric loading, or pathological timing patterns that are outside the training distribution, with varying gait speed.

3.7 Conclusion

This study presents a comprehensive architectural comparison of IMU-based methods for predicting 3D ground reaction forces during sit-to-walk transitions. The systematic evaluation of 13 model configurations across seven architectural families revealed that no single architecture dominates all three GRF axes, establishing axis-dependent design principles for wearable biomechanics. The DS-Conformer achieved the best vertical GRF prediction (R^2 = 0.985, RMSE = 4.05 %BW). For anteroposterior forces, DS-Conformer achieved the highest R^2 (0.893) while ResCorr (joint) achieved the lowest RMSE (1.68 %BW). For mediolateral forces, DS-Conformer again yielded the highest R^2 (0.809) while ResCorr (joint) achieved the lowest RMSE (1.03 %BW). These axis-specific preferences reflect fundamental biomechanical differences: vertical GRF benefits from global temporal context (self-attention), anteroposterior GRF from two-stage residual correction for amplitude accuracy or combined

attention-convolution for waveform fidelity, and mediolateral GRF from architectures that avoid aggressive temporal downsampling. Age-group generalization testing confirmed robust vertical GRF prediction across held-out age groups in this dataset, with DS-Conformer demonstrating the most consistent cross-demographic performance across all axes. Leave-one-out cross-validation over all 65 subjects further confirmed inter-subject generalizability within the same dataset, with DS-Conformer achieving $R^2 = 0.950$ on vertical, $R^2 = 0.809$ on AP, and $R^2 = 0.697$ on ML direction. These findings provide evidence-based guidance for model selection in wearable kinetic assessment systems and establish the feasibility of portable GRF estimation during a clinically critical movement task.

Submitted Manuscript Ends here

3.8 Chapter Summary

This chapter evaluated two hypotheses: (1) architectures combining local temporal feature extraction with longer-range temporal context would provide better waveform agreement than models relying primarily on a single mechanism, and (2) waveform agreement would be lower for mediolateral than vertical GRF. This chapter addressed the second and third research questions of this thesis: "How do different machine learning architectures compare in predicting 3D GRFs from IMU data during STW, and which architecture performs best under rigorous evaluation?" and "What are the generalization characteristics and failure modes of these models?" It presented IMU-based prediction of continuous GRF waveforms during sit-to-walk transitions, using a single sacrum-mounted IMU and a publicly available dataset of 65 participants.

Building on the synthesis gap identified in Chapter 2, 13 neural network configurations across different architectural families were evaluated within a single pipeline to ensure that observed differences could be attributed to the architectures themselves. No architecture proved uniformly superior across the three GRF axes. The DS-Conformer was the most consistently strong and robust performer overall, while ResCorr (joint) achieved the lowest pointwise error on the horizontal axes at the cost of greater inter-subject variability. Age-group holdouts and simulated sensor degradation tests (Appendix B, Table S3-11 to Table S3-13) assessed robustness within the available dataset, with the mediolateral axis remaining the most sensitive to demographic variation; performance under independently recorded deployment conditions remains untested.

These findings demonstrate that accurate, portable GRF estimation during STW is feasible from a single wearable sensor, establishing a foundation for kinetic instrumentation of the TUG test. However, GRFs characterize only the magnitude and direction of the foot-ground interaction. A complete kinetic description of postural control during STW also requires ground reaction moments, which capture rotational loading about the support surface, and center-of-pressure trajectories, which indicate where the resultant force is applied and how support is controlled throughout the transition.

The results were consistent with the first hypothesis for DS-Conformer, which achieved the highest mean R^2 across all three GRF components in five-fold cross-validation. However, architectures combining processing mechanisms did not consistently outperform simpler models, indicating that the advantage depended on the specific architecture. The comparisons therefore support the performance of DS-Conformer without establishing the independent contribution of its architectural components. The second hypothesis was supported by lower waveform agreement for mediolateral than vertical GRF; for DS-Conformer, mean R^2 was 0.809 and 0.985, respectively.

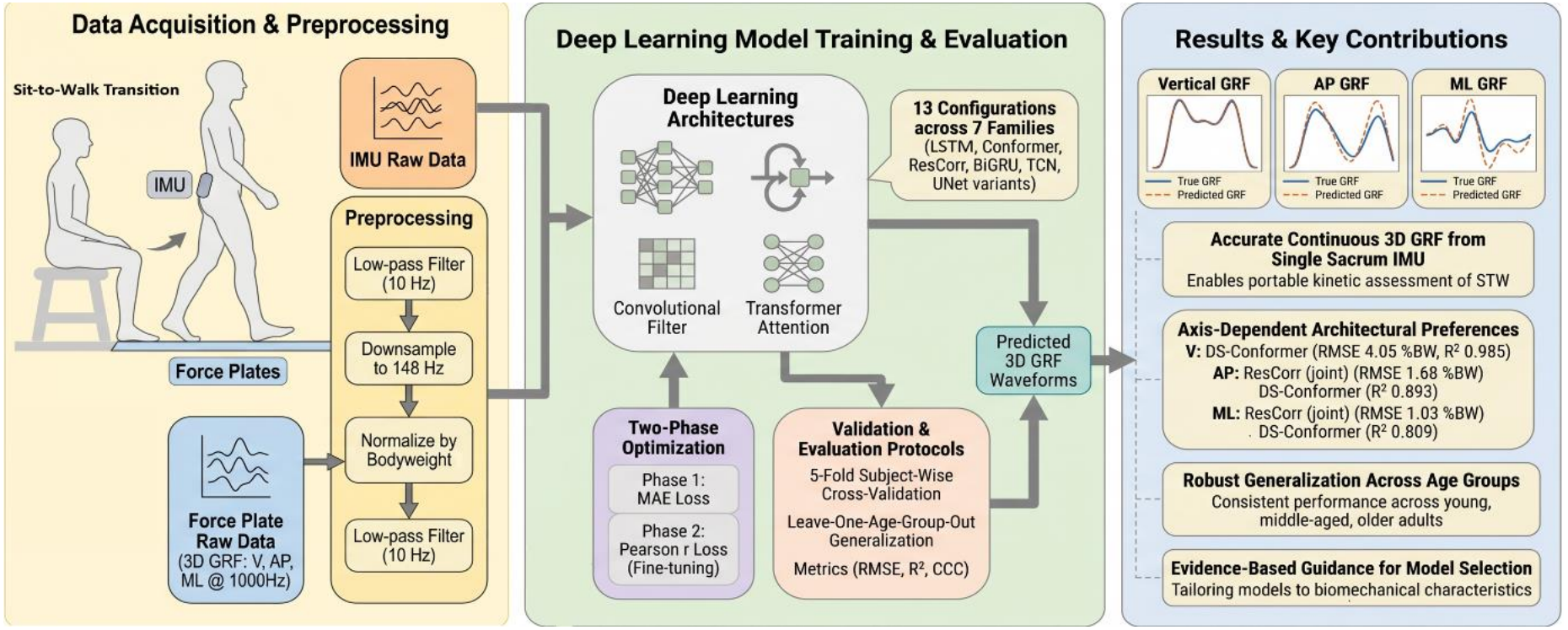


Figure 3-5 Graphical Summary of Chapter 3.

Chapter 4. Deep Learning-Based Kinetics and Center of Pressure Estimation During Sit-to-Walk Transitions from a Single Sacrum-Mounted IMU

4.1 Preface

This chapter presents the extension study: "Deep Learning-Based Kinetics and Center of Pressure Estimation During Sit-to-Walk Transitions from a Single Sacrum-Mounted IMU." The paper has been submitted for publication.

This chapter addresses the fourth research question: "Can the wearable-machine learning framework (identified in Chapter 3) be extended to estimate GRM and CoP trajectories during STW to enable richer stability-focused kinetic markers?" It does so by applying an extended DS-Conformer architecture to predict all kinetic measures, including three-dimensional GRF, three-dimensional GRM, and two-dimensional CoP trajectories during sit-to-walk using a single sacrum-mounted IMU. This extension builds on the architectural findings of Chapter 3 to address the need for wearable assessment of loading and postural control outlined in Chapter 1.

Building on Chapter 3, this extension examined two research hypotheses: (1) the DS-Conformer framework could estimate GRM and CoP waveforms in addition to GRF during STW from a single sacrum-mounted IMU, and (2) AP CoP would have lower range-normalized error and higher waveform correlation than ML CoP.

In addition to expanding the scope of outputs, this chapter refines the Phase 2 training objective. Chapter 3 fine-tuned all models using a pure Pearson correlation loss, which is scale-invariant and can permit amplitude drift. Here, Phase 2 instead employs a composite loss defined as $(1 - \mathrm{r}) \times \mathrm{MAE}$, which jointly optimizes waveform shape and amplitude accuracy. This modification yielded further improvements in GRF prediction relative to the Chapter 3 LOOCV baseline.

Submitted Manuscript Starts here

Deep Learning-Based Kinetics and Center of Pressure Estimation During Sit-to-Walk Transitions from a Single Sacrum-Mounted IMU

4.2 Abstract

The sit-to-walk (STW) transition is a common transitional activity of daily living, yet quantifying its biomechanics outside the laboratory remains challenging because ground reaction forces (GRF), moments (GRM), and center of pressure (CoP) require fixed force platforms. This study presents a deep learning approach for estimating continuous three-dimensional GRF, GRM, and two-dimensional CoP trajectories during STW from a single sacrum-mounted inertial measurement unit. A DS-Conformer architecture was developed that integrates strided convolutional downsampling with conformer blocks, combining multi-head self-attention with depthwise convolution, and a bidirectional gated recurrent refinement to capture the rapid, localized force transients at seat-off and the broader temporal dependencies spanning the full sitting-to-walking sequence. Evaluation on 65 healthy adults (aged 19-73 years) using leave-one-out cross-validation yielded strong kinetic prediction, with NRMSE ranging from 3.78% to 7.69% and Pearson's correlation coefficients (r) from 0.893 to 0.982 across GRF and GRM components. For CoP, the anteroposterior direction achieved an NRMSE of 4.46% and r = 0.990, while the mediolateral direction achieved an NRMSE of 11.48% and r = 0.787, reflecting the lower signal amplitude and greater inter-subject variability of lateral weight transfer. These results support single-sensor STW kinetic estimation in held-out participants from this dataset, motivating independent validation before clinical or community use.

4.3 Introduction

The sit-to-walk (STW) transition is a critical weight-bearing activity of daily living, requiring simultaneous coordination of rising from a seated position and initiating forward gait (Kerr et al., 2004; Perera et al., 2023). Unlike the more commonly studied sit-to-stand movement, STW requires maintaining continuous forward momentum through seat-off and into the first walking

step, rather than terminating in a static standing posture (Jones et al., 2021; Magnan et al., 1996), thereby placing greater demands on postural control. Quantifying the external forces and moments acting on the body during this transition is essential for understanding its biomechanical demands.

The three-dimensional ground reaction force (GRF) characterizes the magnitude and direction of loading at the foot-ground interface, while the three-dimensional ground reaction moment (GRM) captures the rotational loading about the support surface. Together, these six kinetic components determine the center of pressure (CoP), defined as the point of application of the resultant GRF vector on the support surface. During STW, the CoP trajectory provides insight into the coordination of weight transfer, forward progression, and lateral stability (Bestaven et al., 2013). Deviations in CoP behavior during STW have been associated with aging, fall history, and neurological impairment (Buckley et al., 2009; Chen & Chou, 2013; Jones et al., 2021; Mezzarobba et al., 2018), making CoP a clinically meaningful marker of postural control.

Measuring these kinetic variables requires laboratory force plates, which are expensive and require fixed installation, limiting their use in clinical settings (Jones et al., 2021). Wearable Inertial measurement units (IMUs) provide a low-cost alternative to force plates (Ancillao et al., 2018; Lim et al., 2019). Prior studies have demonstrated that GRF, GRM, and CoP can be estimated from IMUs during level walking using machine learning (Lim et al., 2019; Podobnik et al., 2020; Wu et al., 2020). The present study extends this walking-focused work to estimation of GRF, GRM and CoP during STW.

Predicting CoP during STW poses a distinct modeling challenge because the signal encompasses multiple temporally heterogeneous phases within a window of only a few seconds. These include a quiet sitting baseline, rapid anterior weight shift at seat-off, and asymmetric mediolateral loading during the transition to single-limb support (Frykberg et al., 2009; Kerr et al., 2004; Miller et al., 2024). Each deep learning family can capture part of this pattern, but each also has limitations. CNNs with fixed-size kernels are effective at capturing local temporal patterns within each phase, but their receptive fields are bounded by the kernel size and network depth (Lim & Zohren, 2021), thereby limiting their ability to model dependencies across the full STW sequence. Recurrent models such as LSTMs explicitly capture temporal dependencies, but they rely on passing information through a limited hidden state. As a result, fine details from earlier phases may be compressed or overwritten, making

long-range dependencies harder to model (Cho et al., 2014). Self-attention avoids this bottleneck by allowing all timesteps to interact directly (Vaswani et al., 2017). However, a plain Transformer is not optimal for STW CoP prediction because the signal is dominated by short, localized, and ordered events such as seat-off and weight transfer, and Transformers lack inherent local temporal inductive bias for modeling short-duration sequential dependencies. Effectively capturing STW CoP, therefore, requires an architecture that combines fine-grained local feature extraction with global cross-phase modeling.

We propose a Downsampled-Conformer architecture that combines three integrated components: (1) convolutional blocks to extract local temporal features within each STW phase, (2) downsampled multi-head self-attention to model broader dependencies across the sequence, and (3) bidirectional recurrent layers to refine the predicted waveform by integrating contextual information from both earlier and later portions of the movement.

The main contributions of this research are:

1. We present a deep learning model for estimating three-dimensional GRF, GRM, and two-dimensional CoP trajectories during STW from a single wearable IMU sensor, extending CoP prediction beyond steady-state gait to the more challenging STW transition.
2. We develop and validate the DS-Conformer architecture, specifically designed for the temporally heterogeneous dynamics of STW-based CoP prediction.
3. We conduct a rigorous evaluation using leave-one-out cross-validation across 65 subjects, providing a stringent inter-subject generalization assessment.

4.4 Materials and Methods

4.4.1 Dataset

We used a publicly available dataset (Perera et al., 2024) comprising force plate and IMU data of 65 healthy community-dwelling adults (32 males, 33 females; age 39.4 ± 17.5 years) performing STW trials. Participants were categorized into young (19-35 years, n = 32), middle-aged (36-55 years, n = 15), and older adults (56-73 years, n = 18). All participants could stand and walk without external aid and had no declared physical, musculoskeletal, or neurological impairments. The study was ethically approved by the Monash University Human Research

Ethics Committee (project 32328), and written informed consent was obtained from all subjects.

During data collection, participants sat on a backless, armless, height-adjustable stool with their hands resting on their laps, both feet placed on the first force plate, and knees flexed to 90°. On cue, they performed the standard Timed Up and Go (TUG) test protocol, which involved standing up, walking 3 m, turning, returning, and sitting down again as quickly and safely as possible. Each participant completed five repetitions. STW segments were extracted from the initial forward phase of each trial, spanning from quiet sitting to the point at which the participant left the third force plate during gait initiation.

A schematic of the experimental setup is shown in Figure 4-1. For data acquisition, three serially aligned BERTEC force plates were used together with a wireless Delsys Trigno system comprising six Trigno Avanti sensors and two Trigno Duo sensors, all time-synchronized using a Delsys Trigger Module. Only the sacrum-mounted IMU was used as model input in this study. This sensor consisted of an accelerometer (range ±16g, resolution 10bits) and a gyroscope (range ±2000 °/s, resolution 10bits) with a sampling rate of 148Hz and provided six channels: three-axis linear acceleration and three-axis angular velocity.

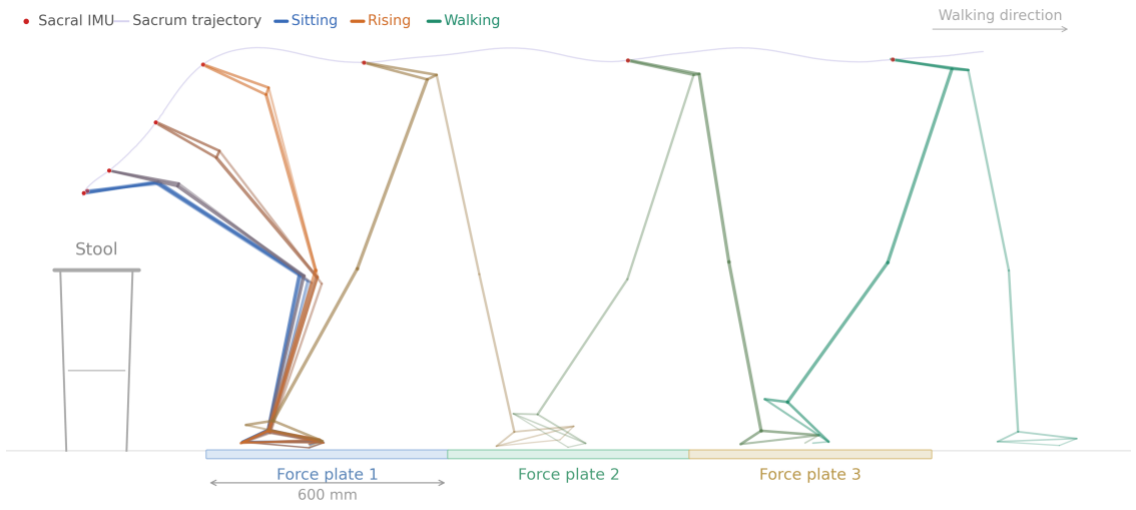


Figure 4-1 Sagittal-view representation of the experimental setup. Skeleton color indicates movement phase. Right leg segments are shown as bold lines. Red dots denote the sacrum-mounted IMU location at each pose. Three serially aligned BERTEC force plates (600 × 400 mm each) are shown at floor level. The participant began seated on a backless, height-adjustable stool with both feet on Force plate 1, then performed the STW transition progressing across all three force plates.

4.4.2 Data Processing

Force plate signals were filtered to reduce noise and motion artifacts using a fourth-order low-pass Butterworth filter. The cutoff frequency was determined using fast Fourier transform (FFT) analysis, by identifying the frequency at which 99% of cumulative signal power was captured (Schilaty et al., 2021). Based on this analysis, a cutoff frequency of 10 Hz was applied. Figure 4-2 shows cumulative FFT power for each GRF and GRM axis.

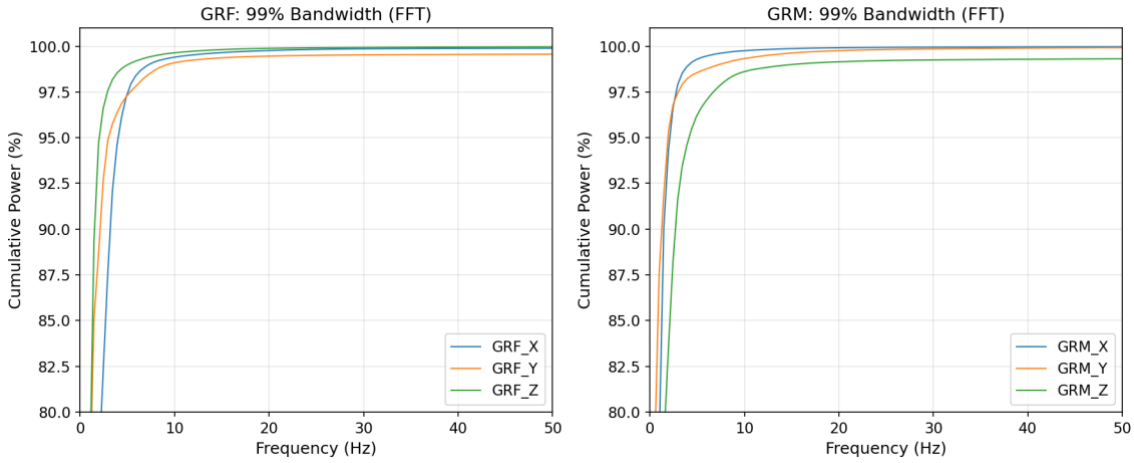


Figure 4-2 Cumulative FFT power for GRF and GRM channels used to determine low-pass filter cutoff frequency. For each axis, the cumulative percentage of total signal power is plotted against frequency.

The filtered force plate data was then downsampled to 148 Hz to match the IMU sampling frequency. GRF and GRM amplitudes were normalized to bodyweight, and the signals were time-normalized to facilitate neural network training. The final paired input-output sequences consisted of IMU time-series data as input and the corresponding 3D GRF, GRM, and 2D CoP trajectory as the target outputs.

4.4.3 Center of Pressure Computation

CoP was computed from the filtered force plate data. For each timestep, the Anteroposterior (AP) and Mediolateral (ML) CoP coordinates were derived from the measured GRFs and GRMs using the standard formulations:

$$CoP_{AP} = -\frac{M_y}{F_z} \quad , \quad CoP_{ML} = \frac{M_x}{F_z}$$

where F_z is the vertical GRF and M_x and M_y are the GRMs about the ML and AP axes, respectively. During the STW transition, participants progressed across three serially aligned force plates. At each timestep, the resultant GRF and GRM were computed by summing the contributions from all active force plates, and CoP was then derived from these resultant values. Each force plate's local CoP was translated to the global lab frame using the plate's known position. When multiple plates were active, the overall CoP was the vertical-force-weighted average of their translated positions. The computed CoP_{AP} and CoP_{ML} trajectories were used directly as model targets.

During the initial quiet-sitting phase of each trial, the stool supported the participant's body weight while the feet rested on the first force plate, resulting in a near-zero vertical force on the plate. Under these conditions, the CoP is physically undefined as F_z approaches zero. As the participant initiates the STW transition, the flexion-momentum phase begins, in which the trunk leans forward to generate horizontal momentum, producing the first measurable change in vertical GRF (Kerr et al., 2004). The ground-truth CoP trajectories were therefore defined from the onset of the flexion-momentum phase through to the end of the trial.

4.4.4 Model Architecture

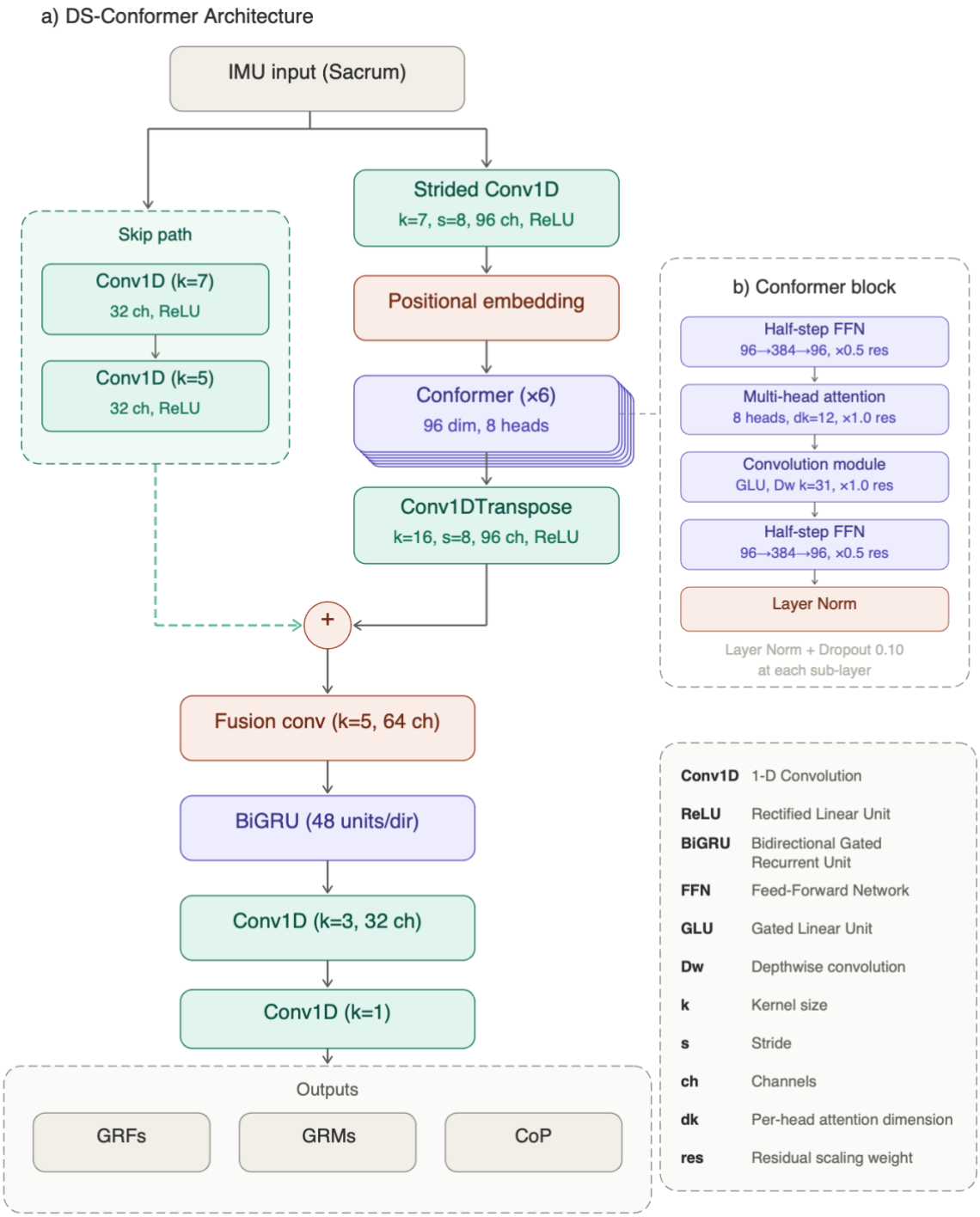


Figure 4-3 DS-Conformer neural network architecture. (a) Overall model architecture comprising a skip-connection branch, a strided downsampling encoder, a bottleneck, and a transposed-convolution upsampling decoder. (b) Detail of a single Conformer block: feedforward half-step, multi-head self-attention, depth-wise convolution, and feedforward half-step, each wrapped in residual connections with layer normalization.

The input signal to the model is processed through two parallel paths (Figure 4-3a). A skip-connection branch applies two one-dimensional convolutional layers to extract fine-grained temporal features at full resolution. This branch provides a route for local features to bypass temporal compression. Two successive convolutional layers with nonlinear activations expand the temporal neighborhood represented by each output and allow further transformation of the extracted features. The use of two layers was a design choice; its advantage over a single layer was not independently tested. In the main encoding path, a strided convolution compresses the temporal dimension by a factor of 8. A learned positional embedding (Gehring et al., 2017) is added to the input. This is a trainable weight matrix that maps each downsampled timestep to a unique vector. It encodes temporal position information, which is essential because the self-attention mechanism in the Conformer blocks is permutation-invariant. Without positional information, the model would be unable to distinguish the temporal order of its inputs. The compressed sequence passes through stacked conformer blocks. Each block (Figure 4-3b) follows a four-module structure: a half-step feedforward network (FFN), multi-head self-attention (MHSA), a convolution module, and a second half-step FFN. Every sub-layer applies layer normalization (Ba et al., 2016) before its transformation and a residual connection (He et al., 2016) afterward. The half-step FFNs scale their outputs by 0.5 before residual addition. The convolution module chains a pointwise expansion, a gated linear unit (Dauphin et al., 2017), a depthwise separable convolution, batch normalization (Ioffe & Szegedy, 2015), ReLU activation, and a pointwise projection. Dropout is applied after each sub-layer, and a final layer normalization closes each block.

Within each block, the feedforward modules provide nonlinear feature transformations at each timestep, while multi-head self-attention relates information across different parts of the movement sequence. The convolution module supplies local temporal processing: pointwise expansion increases the feature representation, the gated linear unit regulates feature transmission, depthwise convolution processes neighboring timesteps within each channel, and pointwise projection combines the resulting channel information. The half-step feedforward scaling follows the Conformer design (Gulati et al., 2020). Residual connections provide direct paths for information and gradients, normalization controls activation scales, and dropout regularizes learning. These complementary functions motivate the combination of local and longer-range temporal processing.

After the Conformer stack, a transposed convolution learns to expand the compressed features back to the original temporal resolution. These features are concatenated with the full-resolution skip features, making both representations available for reconstruction. A fusion convolution learns to combine them while reducing the channel dimension. The bidirectional gated recurrent unit then integrates information from preceding and subsequent timesteps to provide temporal context for each prediction. Of the final two convolutional layers, the first applies nonlinear local refinement, while the second linearly projects the features to a scalar kinetic value at each timestep.

4.4.5 Training Strategy

A dual-stage training strategy was employed to optimize both waveform accuracy and temporal correlation of the predicted CoP trajectories. In Stage 1, the model was trained to minimize mean absolute error (MAE) using the Adam optimizer up to 1000 epochs with early stopping (patience = 50 epochs) based on validation loss. In Stage 2, the best Stage 1 weights were loaded and fine-tuned using a custom correlation-coefficient loss, computed as $(1 - r) \times \text{MAE}$, where r is the per-sample Pearson correlation along the time axis, with the same epoch and early stopping settings. This two-stage approach first establishes an accurate amplitude prediction and then refines the waveform's temporal shape.

Hyperparameters were tuned once using Optuna (Akiba et al., 2019) with 40 trials per model on a development split drawn from the training data, then held fixed across all cross-validation folds.

4.4.6 Model Evaluation

A leave-one-out cross-validation (LOOCV) scheme was employed to evaluate model performance, ensuring complete subject independence between training and test sets. In each of the 65 folds, data from one subject served as the test set, while data from the remaining 64 subjects were used for training. Within the training set, data from 20% of subjects were held

out as a validation set for early stopping, using a group-based split to prevent cross-subject data leakage.

Model performance was assessed using metrics listed below, computed per subject and then averaged across all 65 subjects. The overall performance is reported as the mean ± standard deviation across all 65 subjects.

Root Mean Square Error (RMSE): Measures the average magnitude of prediction errors in the original units.

Normalized RMSE (NRMSE): The RMSE normalized by the range of the corresponding ground truth signal, expressed as a percentage. This allows comparison across components with different scales.

Mean Absolute Error (MAE): Provides the average absolute deviation between predicted and ground truth signals.

The coefficient of determination (R^2) compares squared prediction error with the error obtained by predicting the mean measured value at every evaluated timestep:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where (y_i) and $(\hat{y}_i)$ are the measured and predicted values, and $(\bar{y})$ is the mean measured value over the evaluated samples. It was computed for each trial and averaged within each subject, before calculating the mean and standard deviation across subjects. Negative values were retained because they indicate greater squared error than the corresponding constant mean-value reference.

Pearson Correlation Coefficient (r): Measures the linear association between predicted and ground truth waveforms, reflecting temporal shape agreement.

Lin's concordance correlation coefficient (CCC) was reported because it penalizes deviations in both correlation and bias at the same time, unlike Pearson's r, which is insensitive to constant offsets or scaling errors. The CCC is defined as the product of the Pearson correlation and a bias correction factor, ranging from −1 to +1, with +1 indicating perfect agreement. Together, these metrics offer a thorough assessment of model's performance.

4.5 Results

Figure 4-4 to Figure 4-6 provide a qualitative overview of DS-Conformer performance under leave-one-out cross-validation. Figure 4-4 summarizes the agreement between the mean predicted and ground-truth waveforms across all trials, Figure 4-5 shows a representative single-trial example, and Figure 4-6 illustrates the predicted CoP trajectory relative to the ground truth.

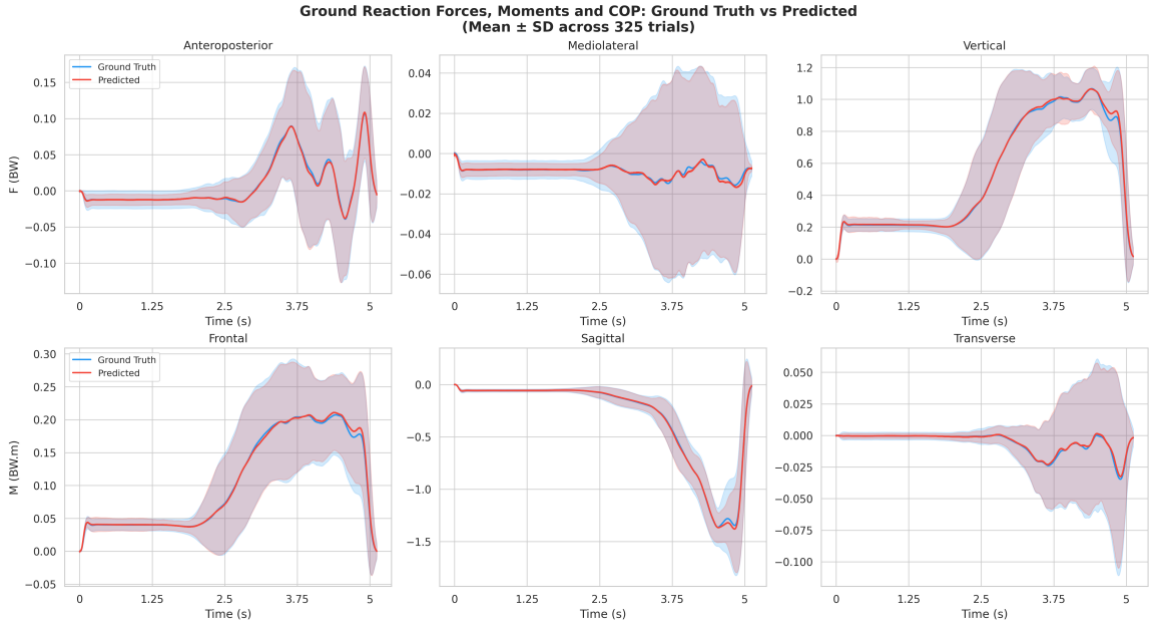


Figure 4-4 Mean predicted (red) and ground truth (blue) waveforms across all 325 trials (65 subjects × 5 windows) from leave-one-out cross-validation using the DS-Conformer model. Solid lines represent mean, shaded regions represent ± 1 SD

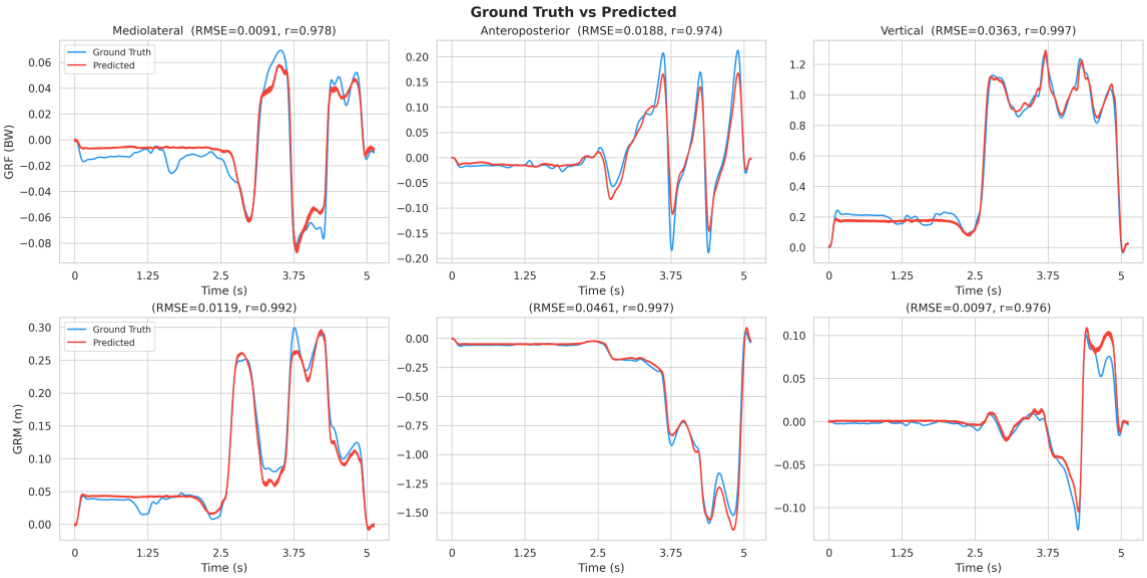


Figure 4-5 Representative single-trial prediction, showing ground truth (blue) and predicted (red) waveforms. RMSE and Pearson r are reported for each component.

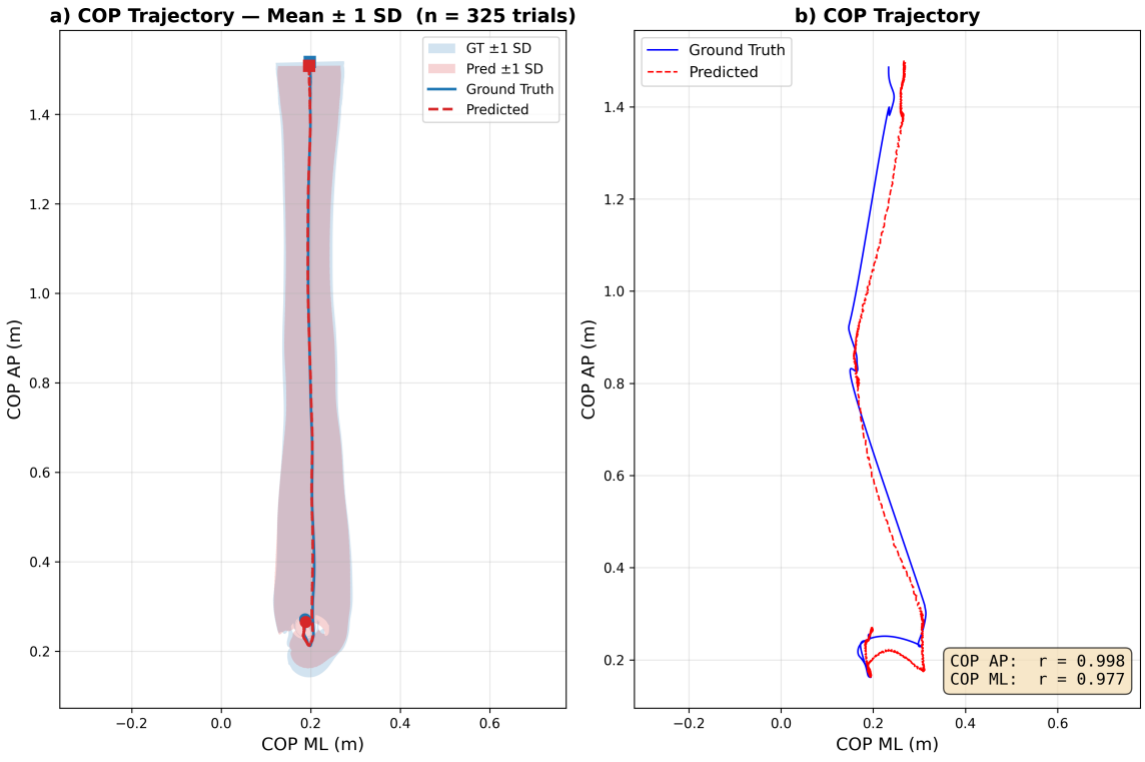


Figure 4-6 CoP prediction from IMU data using the DS-Conformer model. (a) Mean CoP trajectory ± 1 SD across all 325 trials, showing close agreement between ground truth (blue) and predicted (red) paths with largely overlapping variability bands. (b) Representative single-trial CoP trajectory demonstrating high prediction accuracy for both AP (r = 0.998) and ML (r = 0.977) components. CoP coordinates are expressed in meters.

Note: The broad ML SD in Figure 4-6 *(a) arises because participants initiated stepping with either the left or right leg; these opposing lateral excursions cancel in the mean (giving a near-straight trajectory) but are preserved in the SD envelope. The near-perfect overlap of the predicted and ground-truth SD bands demonstrates that the model captures not only the average CoP path but also the full range of stepping strategies across the cohort.*

Overall, the model captured the main waveform shapes and temporal patterns across the kinetic outputs, with the closest visual match seen in the larger sagittal-plane signals. Differences were more noticeable in the lower-amplitude mediolateral components, especially for CoP.

Table 4-1 presents the LOOCV prediction performance across all eight kinetic output components: three GRF channels, three GRM channels, and two CoP coordinates.

Table 4-1 Leave-One-Out Cross-Validation Results for GRF, GRM, and CoP Prediction (N = 65)

Component	RMSE (%BW)	MAE (%BW)	NRMSE (%)	R^2	Pearson r	CCC
GRF (V)	6.55 ± 4.76	4.46 ± 2.36	4.75 ± 3.45	0.955 ± 0.074	0.982 ± 0.033	0.977 ± 0.038
GRF (AP)	2.24 ± 0.76	1.63 ± 0.51	5.48 ± 1.65	0.832 ± 0.132	0.927 ± 0.074	0.913 ± 0.079
GRF (ML)	1.22 ± 0.34	0.83 ± 0.21	7.69 ± 2.05	0.742 ± 0.200	0.893 ± 0.097	0.872 ± 0.105
GRM (Sagittal)	6.92 ± 5.55	3.92 ± 2.37	3.78 ± 3.10	0.949 ± 0.112	0.979 ± 0.047	0.975 ± 0.052
GRM (Frontal)	2.39 ± 1.17	1.67 ± 0.73	6.63 ± 3.28	0.907 ± 0.119	0.966 ± 0.048	0.956 ± 0.054
GRM (Transverse)	1.24 ± 0.33	0.73 ± 0.16	5.74 ± 1.83	0.757 ± 0.195	0.906 ± 0.089	0.883 ± 0.097
CoP (AP)	0.071 ± 0.033	0.050 ± 0.022	4.46 ± 1.97	0.963 ± 0.054	0.990 ± 0.015	0.981 ± 0.035
CoP (ML)	0.032 ± 0.008	0.022 ± 0.006	11.48 ± 3.69	0.468 ± 0.259	0.787 ± 0.134	0.731 ± 0.139

Note. Values represent mean ± SD across 65 subjects from leave-one-out cross-validation. RMSE and MAE are reported in percent bodyweight (%BW) for GRF, in meters (m) for CoP, and in %BW·m for GRM components. NRMSE = RMSE normalized by the ground-truth signal range. R^2 = coefficient of determination; r = Pearson correlation coefficient; CCC = Lin's concordance correlation coefficient. Cell shading indicates performance. Thresholds for R^2, Pearson r, and CCC: > 0.90: excellent (green), 0.75–0.90: good (yellow), 0.50–0.75: moderate (orange), and < 0.50: poor (red). NRMSE thresholds: < 5%, 5–10%, 10–20%, and > 20% respectively. RMSE and MAE are unshaded (raw units).

Prediction accuracy followed a consistent directional pattern across all three kinetic groups: sagittal-plane components were estimated more accurately than lateral or transverse ones. Among the GRF components, vertical force achieved the highest accuracy (NRMSE = 4.75%, r = 0.982), followed by AP (NRMSE = 5.48%) and ML (NRMSE = 7.69%). The same order appeared in the GRM results, where the sagittal moment achieved the lowest NRMSE (3.78%, r = 0.979). This directional disparity was most pronounced for CoP. The AP direction was estimated with high accuracy (NRMSE = 4.46%, r = 0.990), while the ML direction showed substantially higher error (NRMSE = 11.48%, r = 0.787). The biomechanical basis for this pattern is examined in the Discussion.

4.5.1 Performance Across Age Groups

To examine whether prediction accuracy differed across age groups, a two-stage non-parametric testing procedure was employed. Non-parametric tests were chosen because the NRMSE distributions did not satisfy the normality assumption (Kolmogorov-Smirnov test). First, the Kruskal-Wallis H test was applied as an omnibus test for each output component across the three age groups. Pairwise, two-sided Mann–Whitney U tests were calculated for all three age-group pairs, with Bonferroni adjustment applied within each outcome (adjusted $p = \min[3p, 1]$). The Kruskal–Wallis test assessed the overall difference among the three groups; pairwise results were interpreted as post-hoc evidence only following a significant omnibus result ($p < 0.05$).

Table 4-2 Normalized RMSE (%) by Age Group for All Output Components (N = 65).

Component	Young[1]	Mid-aged[2]	Older[3]	*H*	*p*	$P_{1,2}$	$P_{1,3}$	$P_{2,3}$
GRF (V)	4.29 (2.68)	6.12 (5.15)	5.17 (3.51)	1.84	0.398	0.405	1.000	1.000
GRF (AP)	5.74 (1.82)	6.91 (2.32)	6.12 (1.35)	4.57	0.102	0.145	0.517	1.000
GRF (ML)	8.81 (2.60)	8.76 (2.92)	9.03 (2.18)	0.80	0.670	1.000	1.000	1.000
GRM (Sagittal)	3.14 (2.03)	4.65 (4.57)	4.85 (3.63)	5.11	0.078	0.747	0.069	1.000
GRM (Frontal)	6.22 (1.78)	8.67 (5.58)	7.66 (3.36)	3.05	0.218	0.338	0.736	1.000
GRM (Transverse)	6.70 (2.46)	6.24 (2.12)	7.17 (1.73)	2.50	0.287	1.000	1.000	0.258
CoP (AP)	4.43 (2.31)	4.26 (1.72)	5.77 (1.96)	**10.76**	**0.005**	1.000	**0.014**	**0.010**
CoP (ML)	12.41 (3.83)	14.42 (3.97)	17.77 (3.72)	**15.88**	**<.001**	0.370	**<.001**	0.054

Note: Values are NRMSE (%) reported as mean (SD). H = Kruskal-Wallis test statistic (omnibus, df = 2), p = omnibus p-value. Pi,j = pairwise Mann-Whitney U tests with Bonferroni correction. Young[1] n = 32 (age 19-35), Mid-aged[2] n = 15 (age 36-55), Older[3] n = 18 (age 56-73). Bold indicates p < .05.

For GRF and GRM components, no significant age-group differences in NRMSE were observed (Kruskal-Wallis, all $p > 0.05$). By contrast, significant effects of age group were found for both CoP components. CoP (ML) showed the strongest effect ($H = 15.88$, $p < 0.001$); post-hoc tests indicated that the older group had significantly higher NRMSE than the young group ($p < 0.001$), while the difference between the middle-aged and older groups did not reach significance. CoP (AP) also showed a significant omnibus effect ($H = 10.76$, $p = 0.005$), with the older group performing significantly worse than both the young group ($p = 0.014$) and the middle-aged group ($p = 0.010$). Figure 4-7 presents the leave-one-age-group-out analysis: CoP (AP), GRM (Sagittal), and GRF (V) maintained consistently high prediction accuracy ($r > 0.95$) regardless of which age group was held out, whereas CoP (ML) showed the greatest performance variation across hold-out conditions.

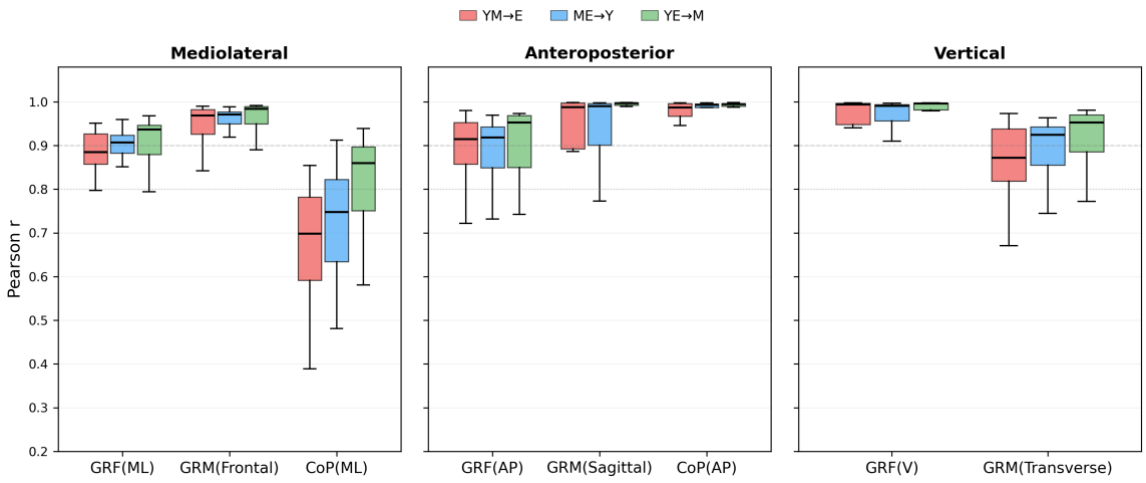


Figure 4-7 Age-group generalization performance (Pearson r) of the DS-Conformer model across all eight kinetic output components, organized by biomechanical axis. Each panel displays per-subject distributions for GRF, GRM, and CoP predictions under three leave-one-age-group-out holdout scenarios: train on Young+Mid, test on Elder (YM→E, n=18); train on Mid+Elder, test on Young (ME→Y, n=32); and train on Young+Elder, test on Mid (YE→M, n=15). Dashed and dotted horizontal lines indicate r = 0.9 and r = 0.8, respectively. The vertical axis has no associated CoP component. CoP(ML) exhibits the lowest prediction accuracy across all splits, while CoP(AP), GRM(Sagittal), and GRF(V) consistently achieve r > 0.95.

4.5.2 Comparison with Existing Methods

Table 4-3 compares the reported prediction errors for GRF, GRM and CoP in the present study and previous studies. Values retain each study's normalization and evaluation conditions; differences in task, population, sensor configuration, validation protocol and GRM reference point mean that the table provides descriptive context rather than a direct ranking of methods.

Table 4-3 Comparison of reported GRF, GRM and CoP prediction errors across studies

Component	Present study	Wu et al. (2020)	Lee and Park (2020)	Karatsidis et al. (2016)	Liu et al. (2024)
Task	STW	Walking	Walking	Normal walking	Stair ascent and descent
Participants (N)	65	5	17	11	172
Input	1 IMU	1–4 IMUs; foot and waist combinations	1 IMU (sacrum)	17 IMUs	Optical motion capture; complete kinematics
Ground reaction force (GRF)					
Vertical	4.75 ± 3.45	—	6.70 ± 2.41	5.30 ± 3.10	2.76 ± 1.30
Anteroposterior	5.48 ± 1.65	—	9.21 ± 4.06	9.40 ± 2.50	7.57 ± 3.43
Mediolateral	7.69 ± 2.05	—	15.55 ± 5.03	13.10 ± 2.80	5.10 ± 2.94
Ground reaction moment (GRM)					
Sagittal	3.78 ± 3.10	—	—	12.40 ± 3.40	—
Frontal	6.63 ± 3.28	—	—	29.60 ± 9.30	—
Transverse	5.74 ± 1.83	—	—	18.20 ± 4.70	—
Centre of pressure (CoP)					
Anteroposterior	4.46 ± 1.97	5.88 ± 0.96	8.22 ± 4.96	21.30 ± 5.30	5.52 ± 2.30
Mediolateral	11.48 ± 3.69	25.33 ± 2.35	19.54 ± 9.59	32.40 ± 10.90	13.98 ± 12.98

These cross-study comparisons should be approached with caution, as they are not directly comparable due to variations in task, population, and the number and placement of IMUs.

4.6 Discussion

Taken together, the leave-one-out results confirm that a single trunk-mounted IMU contains sufficient information to reconstruct all eight kinetic components of the STW transition with good accuracy. The following paragraphs examine why prediction quality varies across

components and what this reveals about both the biomechanics of STW and the architectural choices of the DS-Conformer.

Table 4-1 revealed a directional accuracy hierarchy. The model performed best along the sagittal plane and worst in the mediolateral direction. Forward progression generates a large-amplitude signal at the sacrum, whereas lateral loading produces lower-amplitude forces with greater movement-strategy variation across individuals (Figure 4-4). The same directional asymmetry has been reported in other studies, as listed in Table 4-3. The mediolateral CoP component is inherently low-amplitude, and at this scale, the inertial sensor's noise floor is no longer negligible. The signal-to-noise ratio deteriorates to the point where meaningful kinematic correlates of lateral loading become difficult to recover.

The age-group analysis (Table 4-2, Figure 4-7) revealed that CoP prediction is the most sensitive output to age-related biomechanical differences, while GRF and GRM accuracy remained stable across groups. The significantly higher ML CoP error in the older group aligns with documented increases in lateral stabilization variability among older adults during STW (Miller et al., 2024; van der Kruk et al., 2021), an effect likely compounded by training-set imbalance (Young n = 32 vs. Older n = 18). That CoP (AP) also degraded with age, despite its overall high accuracy, indicates the model does not fully capture age-related changes in forward progression strategy.

Several aspects distinguish our work from prior CoP prediction studies (Table 4-3). Unlike the steady-state movements examined previously, STW involves rapid transitions between mechanically distinct postures within a few seconds, thereby increasing the difficulty of modeling. Moreover, only a single IMU is used as input, in contrast to the multi-sensor setups of Karatsidis et al. (2016) (seventeen IMUs), Wu et al. (2020) (four IMUs), and Liu et al. (2024) (full motion capture). This single-sensor approach maximizes clinical feasibility through reduced cost and setup time.

The observed accuracy reflects the complete model and cannot be attributed to individual architectural stages. Conceptually, self-attention relates distant timesteps, convolution supplies local context, the skip branch retains full-resolution features, and the BiGRU integrates temporal context in both directions (Cho et al., 2014; Gulati et al., 2020). These functions motivated the design, but the present results do not show which component caused the strong

sagittal-plane performance or the weaker ML CoP prediction. Controlled ablations would be needed to test those contributions.

4.6.1 Limitations and Future Work

This study has several limitations. Evaluation was limited to a single dataset comprising 65 subjects with 325 trials, which is relatively large compared to many previous CoP prediction studies that employed fewer than 20 subjects, and the wide age range (19-73 years) allows age-group comparisons within this cohort. These comparisons assess prediction across participants recorded under the same conditions; transfer to other sensors, laboratories or acquisition setups remains untested. However, compared to the 172-subject dataset of Liu et al. (2024) for stair walking, the present sample is smaller, and increasing its size could further improve accuracy. Moreover, the dataset includes only healthy adults; including individuals with pathological gait patterns is necessary to validate its applicability beyond this population. Addressing the ML CoP limitation will likely require a combination of strategies: incorporating additional sensor locations closer to the lower limbs to capture lateral dynamics more directly, employing subject-specific adaptation or fine-tuning to account for individual stabilization strategies, and augmenting training data with synthetic variations in lateral weight-shift patterns.

Parameter counts and inference costs for the DS-Conformer are reported in Supplementary Material Section C2. However, performance on resource-limited hardware remains untested, so suitability for real-time wearable deployment remains unproven. The final design combines several elements, including skip-connection fusion, conformer blocks, and bidirectional recurrent refinement, but no formal ablation study was conducted. As a result, the individual impact of each component on overall performance cannot be determined, and it remains possible that a simpler design could achieve similar accuracy at lower computational cost.

4.7 Conclusion

Leave-one-out evaluation across 65 adults (aged 19-73) confirmed that the DS-Conformer can estimate GRF, GRM, and CoP during STW from a single sacrum-mounted IMU, achieving NRMSE values of 3.78-11.48% across all eight kinetic components. Sagittal-plane outputs were predicted with consistently high fidelity, while mediolateral CoP remains the primary

challenge, particularly for older adults. These findings support the development of portable STW assessment tools, although performance outside the study's acquisition setting remains to be established.

Submitted Manuscript Ends here

4.8 Chapter Summary

This chapter evaluated two hypotheses: (1) the DS-Conformer framework could estimate GRM and CoP waveforms in addition to GRF from a single sacrum-mounted IMU, and (2) AP CoP would have lower range-normalized error and higher waveform correlation than ML CoP. This chapter addressed the fourth research question of this thesis: "Can the wearable-machine learning framework identified in Chapter 3 be extended to estimate GRM and CoP trajectories during STW to enable richer stability-focused kinetic markers?" Building on the model-selection findings of Chapter 3, the DS-Conformer framework was extended from GRF estimation to the simultaneous prediction of GRF, GRM, and CoP from a single sacrum-mounted IMU under subject-independent evaluation. In doing so, this chapter broadened the thesis from force estimation alone to a more complete wearable description of STW kinetics and postural control, while indicating that lateral-balance-related outputs remain the main challenge for further development. The extension demonstrates the feasibility of estimating a broader set of STW kinetic measures from one wearable sensor. The lower ML CoP accuracy, especially in older adults, identifies lateral balance estimation as a priority for improvement before clinical use.

The first hypothesis was supported by the estimation of GRM and CoP waveforms under leave-one-out cross-validation, although accuracy varied across components and was lowest for ML CoP. The second hypothesis was supported by lower NRMSE and higher Pearson's r for AP CoP (4.46%, 0.990) than ML CoP (11.48%, 0.787). Both findings support the hypotheses within the healthy cohort studied; clinical validity and transfer to other acquisition settings remain untested.

Supplementary Material Section C3 complements these findings with an exploratory analysis of GRF estimation during sit-to-stand in an external dataset. AP and vertical GRF predictions showed good waveform agreement, whereas ML estimation remained weak. These findings motivate further external evaluation but do not establish generalization to measured IMUs or external validity for GRM and CoP.

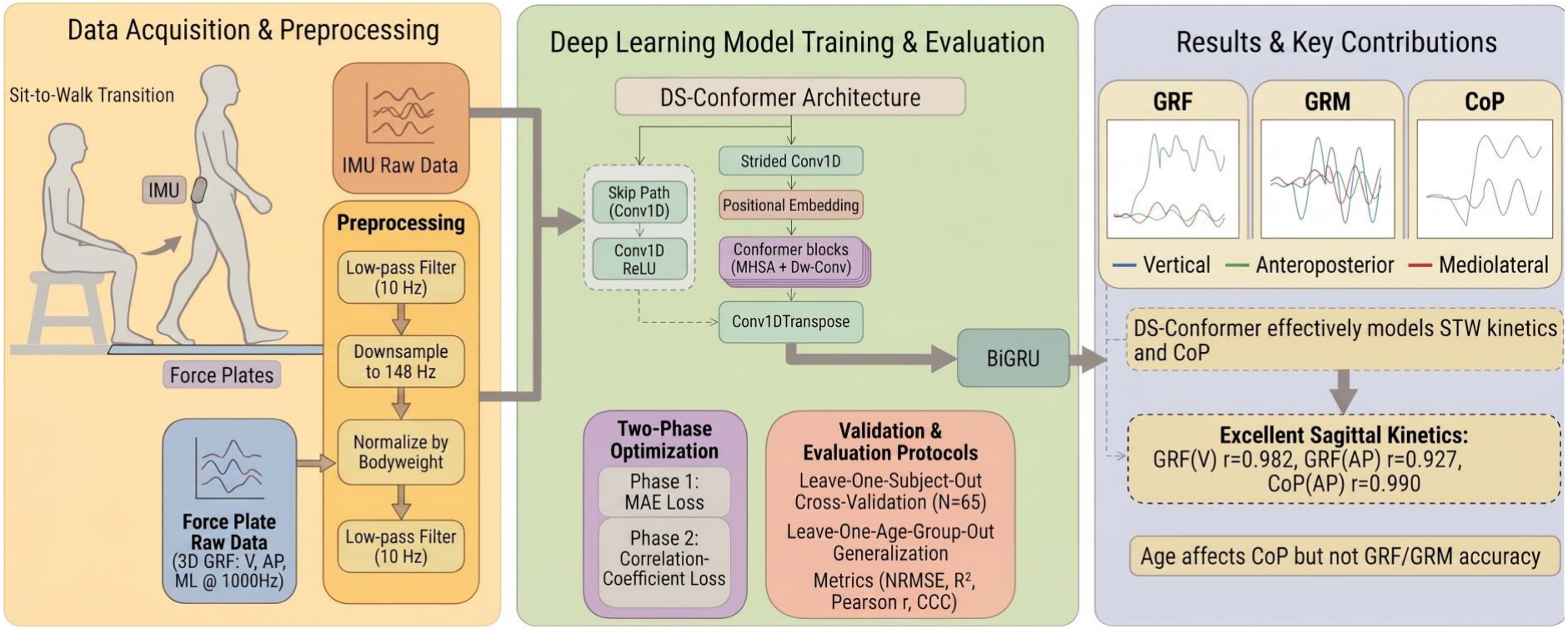


Figure 4-8 Graphical Summary of Chapter 4

Chapter 5. Discussion and Conclusion

5.1 Prologue

This research aimed to develop a framework for estimating kinetic variables during the sit-to-walk (STW) transition of the Timed Up and Go (TUG) test using a single wearable inertial measurement unit (IMU), without laboratory force plates. To achieve this, a systematic review of existing machine learning methods for ground reaction force (GRF) estimation from IMU sensors was conducted, followed by two experimental studies: the first benchmarking thirteen model architectures for continuous 3D GRF prediction during STW, and the second extending the best-performing architecture to simultaneously estimate GRF, ground reaction moments (GRM), and center of pressure (CoP) trajectories. This chapter provides an overview of the main findings and discusses their implications for clinical practice and future research, as well as the strengths and weaknesses of this work.

5.2 Overview of Findings

This thesis approached the estimation of continuous kinetic information from the STW transition of the TUG using a clinically feasible sensor setup through three coordinated stages. The systematic review (Chapter 2) mapped the field of IMU-to-GRF estimation and established the methodological standards that the experimental work would adopt. Chapter 3 then applied those standards in a controlled multi-architecture benchmark on a single sacrum-mounted IMU, and Chapter 4 carried the selected architecture forward to simultaneous estimation of all eight kinetic channels under a stricter validation protocol.

Across the experimental chapters, the strongest conclusion is not about which model performed best but about which kinetic outputs are recoverable from a single midline sensor at this body location, and which are not. The results showed a clear directional hierarchy across Chapters 3 and 4. Vertical GRF was the strongest force output, sagittal GRM was the best-estimated kinetic output overall, and AP CoP was also well estimated. By contrast, ML outputs were the weakest. This suggests that the dominant sagittal dynamics of STW are more strongly encoded in a sacrum-mounted IMU than fine lateral balance-control signals. This interpretation is

biomechanically plausible as during STW, the sacrum moves through a pronounced forward and upward trajectory as the body rises and transitions into gait (Kerr et al., 2004; Magnan et al., 1996). These dynamics are tightly coupled to vertical support and AP progression, both of which are high-amplitude, temporally structured features of the movement. Lateral control, by contrast, depends on smaller-amplitude shifts in load and CoP as the body moves from bilateral to unilateral support (Bestaven et al., 2013). A single midline sensor mounted at the sacrum is less directly informative about left-right events. The results do not prove that sensor location is the only reason for this pattern, as the model's inductive bias and the target signal's properties also play a role. However, the consistent sagittal-plane advantage observed across GRF, GRM, and CoP in various model architectures implies that this limitation is not accidental. Instead, it reflects a core connection between trunk-level inertial signals and sagittal-plane ground kinetics. The practical implication is that improving lateral kinetic estimation will require either additional sensor locations, for example, on the shanks or feet, where inter-limb weight shifting is more directly observable, or complementary sensing modalities, such as pressure-sensing insoles.

Age-group generalization in Chapter 4 further complements these findings. There were no significant age-group differences in prediction accuracy for GRF and GRM, but significant age effects for ML CoP, with the older group showing substantially higher error. Buckley et al. (2009) showed that older adults adopt different sagittal-plane strategies during STW, generating less momentum before rising and delaying gait initiation, yet the model's sagittal-plane accuracy was maintained across age groups. The inability of ML CoP to generalize across age groups may be attributed either to the baseline prediction accuracy on this channel, which was already the lowest among all outputs, or to the biomechanical signal varying more with age in the ML direction than in the sagittal plane. Miller et al. (2024) and Perera et al. (2023) documented a significant increase in ML CoP excursion (median 164 mm in older adults compared to 84 mm in young adults) and greater variability in lateral stabilization among older adults during STW assessments measured with force plates. Consequently, a model primarily trained on younger individuals would likely exhibit reduced generalizability to an older population characterized by inherently more variable lateral control strategies.

Another insight concerns how model performance is assessed. ResCorr (joint) achieved the lowest mean pointwise RMSE for AP and ML GRF in the five-fold comparison. However, for ML GRF, its R^2 values were substantially more variable across subjects, with negative values

in 3 of 65 subjects. DS-Conformer achieved slightly higher RMSE but higher mean R^2, Pearson's r and CCC, together with a smaller subject-level R^2 standard deviation. These findings support considering prediction error, waveform agreement and consistency across subjects when selecting a model.

DS-Conformer was the most reliable architecture for this dataset and task, rather than being the universally best architecture for all IMU-to-kinetics problems. In a clinical context, that difference matters: a model that occasionally performs very well but fails unpredictably in some participants is less useful than one that is slightly more conservative yet consistent (Enderling & Wolkenhauer, 2020). Hence, DS-Conformer was selected for the Chapter 4 extension because of this reliability criterion, and the LOOCV results confirmed that this reliability carried over to a stricter evaluation protocol. The reliability of DS-Conformer stems from how its architecture represents time. During the STW transition, the network must anchor predictions to brief events such as seat-off, initial contact, and lateral weight shifts, while also tracking the overall progression of the transition. Architectures that address only one of these scales may perform well for some participants but generalize less consistently across others. Combining depthwise convolution with self-attention inside a single block allows DS-Conformer to preserve both representations rather than forcing the model to rely predominantly on one scale (Gulati et al., 2020). This may explain its lower inter-subject variance in this thesis: the model was not forced to choose between local event-level information and global transition-level information. Instead, DS-Conformer retained both sources of temporal context, allowing predictions to remain stable when different kinetic outputs depended more strongly on one scale than the other.

The parallel use of convolution and attention also separates DS-Conformer from the downsampled pure-attention variant evaluated alongside it. Self-attention alone has no built-in preference for temporal locality. It must learn that neighboring samples are often more related than distant ones, and this weak locality bias is a known limitation of attention-only architectures (d'Ascoli et al., 2022; Gulati et al., 2020). This makes pure-attention models more prone to smoothing or misrepresenting short transients that encode loading and balance events. Depthwise convolution directly supplies this locality bias, allowing attention to focus on relating more distant moments within the transition, such as the relationship between early seat-off dynamics and subsequent force development. It is therefore reasonable to expect that a

pairing of complementary inductive biases will yield more consistent performance across participants than attention or convolution alone, as observed in this thesis.

In summary, the primary aim of this thesis was met across a systematic review and two experimental studies, providing a framework for estimating continuous 3D GRF, GRM, and CoP trajectories during the STW phase of the TUG test from a single sacrum-mounted IMU. The key contributions and main findings of each chapter are summarized in Table 5-1.

Table 5-1 Summary of research questions and main findings by chapter.

Chapter	Research Question	Main Findings
2	What is the state of the art of machine learning methods for estimating 3D GRFs from wearable sensors across human movement tasks?	A PRISMA-compliant systematic review of 22 studies identified model families including NARMAX, feedforward neural networks, CNNs, LSTMs, and bidirectional LSTMs. Vertical GRF was generally more accurately predicted than AP and ML components, but performance depended strongly on task design, model used, and validation choices. The reviewed studies included walking and running tasks as well as drop jumps, drop landings, ramp walking, and stair climbing. No included study addressed STW. Only six studies used a single IMU, with protocols and reference systems varying widely, and outdoor studies often relied on vertical-only insoles, which limited direct comparison. The review thus identified both a gap in STW application and the need for standardized evaluation using BW normalization, clearly defined temporal windows, subject-independent validation, and shared public benchmarks.
3	How do different machine learning architectures compare in predicting 3D GRFs from IMU data during STW, and which architecture performs best?	Thirteen model configurations spanning seven architectural families were evaluated using a single sacrum-mounted IMU on 65 participants through a unified pipeline with subject-aware five-fold cross-validation. No architecture outperformed others across all three axes. DS-Conformer achieved the best overall performance under five-fold cross-validation: vertical GRF $R^2 = 0.985$, RMSE = 4.05 %BW; AP $R^2 = 0.893$; ML $R^2 = 0.809$. Under LOOCV, DS-Conformer achieved $R^2 = 0.950$ on vertical GRF, $R^2 = 0.809$ on AP, and $R^2 = 0.697$ on ML. ResCorr (joint) had the lowest AP and ML RMSE but showed more instability and less waveform fidelity, so the choice of model depended on whether pointwise error or waveform shape was more important.
3	What are the generalization characteristics and failure modes of these models for STW kinetics?	Age-group generalization within the study dataset using leave-one-age-group-out cross-validation confirmed robust vertical GRF prediction across young, middle-aged, and older adult groups, while ML was most sensitive to demographic changes. Leave-one-out cross-validation across all 65 subjects confirmed strong subject-independent performance for DS-Conformer (vertical $R^2 = 0.950$; AP $R^2 = 0.809$; ML $R^2 = 0.697$). Robustness analysis under added noise, axis misalignment, and signal dropout showed DS-Conformer as the least-affected architecture overall.
4	Can the framework be extended to estimate GRM and CoP trajectories during STW?	The DS-Conformer was expanded to simultaneously estimate GRF, GRM, and CoP from a single sacrum-mounted IMU using LOOCV with 65 participants. CoP AP: r = 0.990, NRMSE = 4.46%; CoP ML: r = 0.787, NRMSE = 11.48%. GRF and GRM components achieved NRMSE of 3.78%-7.69% and correlations of r = 0.893-0.982. Age-group effects were significant for CoP but not for GRF or GRM. Performance was comparable to or better than previous multi-sensor CoP studies during simpler walking tasks.

The following sections discuss the implications of these findings for clinical practice and future research, as well as their strengths and weaknesses.

5.3 What This Thesis Adds

At a conceptual level, this thesis frames STW within the TUG test as a kinetic measurement problem rather than a purely time- or kinematic-based problem. Across the empirical chapters, the most accurate predictions were those most directly related to support generation, forward movement, and sagittal control (vertical GRF, AP CoP, and sagittal-plane GRM). This suggests that clinically relevant information about how a person rises and initiates gait can be recovered from wearable data in a way that total TUG time alone cannot provide. That matters because it changes what can reasonably be asked of an instrumented functional test. Instead of treating the TUG only as a faster-or-slower task, it becomes possible to ask whether a person generated sufficient support, transferred weight efficiently, or adopted a more cautious movement strategy during the transition.

Methodologically, the work demonstrates that a single sacrum-mounted IMU, when paired with a rigorously evaluated deep learning pipeline, can estimate continuous GRF, GRM, and CoP waveforms during STW under subject-independent evaluation. The systematic review (Chapter 2) established why prior studies were difficult to compare. Chapter 3 then isolated architecture as the main variable by holding preprocessing, normalization, validation, and metrics constant across model configurations. Chapter 4 extended the selected architecture to multi-output prediction. The contribution is therefore a defensible pathway for model selection and extension, not simply a favorable result for DS-Conformer. The benchmark also revealed two transferable findings: performance is axis-dependent, and mean accuracy and deployment reliability can diverge, cautioning against single-metric model selection.

In applied terms, the thesis provides a laboratory-based proof of concept for wearable STW kinetic estimation. Prior to this work, laboratory force plates were required to directly characterize STW kinetics. This work shows that all kinetic channels can be estimated from a single wearable sacrum-mounted sensor during the transition, with subject-independent

validation in 65 adults aged 19–73 years. It means that a capability previously limited to laboratory instrumentation could potentially be approximated using a single wearable sensor, pending independent validation. However, how far that approximation extends and which indicators it is reliable for depend on the analysis presented in the following sections.

5.4 Implications for clinical practice

In current clinical practice, mobility during the TUG is assessed using total completion time and, when iTUG systems are used, kinematic and spatiotemporal features. Neither approach directly measures the forces and moments that govern balance control during the transition. This limitation is important because the TUG instruction of "stand up and walk" (Podsiadlo & Richardson, 1991) can be biomechanically ambiguous. One individual might perform this as a single fluent action, rising from the chair and continuing directly into gait, while another might perform it as a sequence: stand up fully, stabilize, then initiate walking. van der Kruk et al. (2021) and Perera et al. (2023) showed that age-related decline can make it difficult to merge rising and walking into a fluent STW transition, and pausing between the sitting and walking may represent a compensatory response to reduced neuromuscular strength. Buckley et al. (2009) further showed that older adults generated less momentum before rising and delayed gait initiation compared with younger adults. Taken together, these findings suggest that the presence, duration, and kinetic signature of STW may instead represent a clinically informative feature that total TUG time alone fails to capture.

Following independent-dataset and clinical validation, the methods developed here could support estimation of STW kinetics during routine TUG assessment from a single wearable sensor. Vertical GRF waveforms can inform clinicians about lower limb loading capacity and the rate of weight transfer during rising, AP GRF waveforms can differentiate participants who generate adequate forward momentum from those adopting a cautious, momentum-limited strategy associated with elevated fall risk, CoP AP trajectories provide a direct representation of how the base of support is controlled during the transition, enabling identification of slow or posteriorly biased CoP migration that is characteristic of impaired balance. GRMs provide additional information on rotational demands at ground contact and may reveal mediolateral

trunk-loading asymmetries in individuals with unilateral impairments, such as post-stroke hemiplegia or knee osteoarthritis.

Beyond waveform interpretation, the estimated kinetic signals may also improve the temporal resolution of the TUG test. Previous iTUG studies have shown that sub-task durations provide clinically relevant information that total completion time obscures (Higashi et al., 2008; Hoskovcová et al., 2015; Ortega-Bastidas et al., 2023; Vervoort et al., 2016; Zarzeczny et al., 2017). This is important because the TUG is not a single movement, but a sequence of sub-tasks, including standing, walking, turning, and sitting, and impairments may appear in one sub-task even when the total completion time is unchanged. For the STW component, the sub-phases are segmented based on the kinetic events (Chandler et al., 2021; Kerr et al., 2004). Estimating kinetics, therefore, also improves the granularity of total TUG time by reporting time spent within individual STW sub-phases

However, measurement without actionable interpretation is insufficient. A clinician measures blood pressure because a threshold exists, an intervention exists (medication, lifestyle change), a monitoring cadence exists, and a response criterion exists. Equivalent clinical decision pathways do not yet exist for STW kinetics. Before this framework can become clinically useful rather than merely scientifically interesting, several steps are required. First, normative kinetic reference ranges during STW must be established so that clinicians can interpret an individual patient's GRF, GRM, and CoP profiles relative to expected values for their age, sex, functional level, and clinical condition. These norms are currently limited because existing STW force plate studies have generally used limited sample sizes and populations. The wearable framework developed in this thesis could help address this limitation by enabling kinetic data collection at scale outside the laboratory, but such reference ranges must be developed before individual clinical interpretation is possible. Second, clinically actionable thresholds need to be defined through prospective studies linking kinetic markers to outcomes such as fall incidence, functional decline, and rehabilitation response. Third, the predicted kinetic outputs must be translated into formats that support rapid clinical reasoning. Summary scores, indicators for values outside expected ranges, and overlay plots comparing a patient's waveform with their previous assessment may be more clinically useful than raw waveform displays. This aligns with Higashi et al. (2008), who argued that the challenge in clinical movement assessment is not simply sensor availability but the generation of objective data that clinicians can interpret and act upon. Finally, the framework must complement existing

workflows rather than add complexity. The single-sensor approach is clinically attractive because it requires minimal setup, avoids inter-device calibration, and can be embedded within existing TUG protocols. However, the analytical layer linking predicted waveforms to clinical decisions remains to be developed.

Further research is therefore needed before these methods are viable in clinical practice. Validation across independent acquisition settings and in clinically representative populations, including older adults with documented fall risk or individuals with a clinical condition is required. Following such validation, the framework could provide clinicians with mechanistic kinetic information during routine TUG assessments that completion time alone does not capture. This information could support a more targeted interpretation of mobility performance and a more specific rehabilitation prescription. For example, it could help identify whether a patient needs to improve loading rate, forward propulsion, or lateral stability, while also enabling more objective monitoring of the response to intervention over time. The proposed framework also provides a foundation for developing a portable kinetic iTUG tool that could be deployed in primary care settings to improve the sensitivity of fall-risk screening and rehabilitation monitoring.

5.5 Strengths of this thesis

Comprehensive preliminary work was conducted to map the existing literature on machine learning-based GRF estimation from wearable sensors, applying a PRISMA-compliant methodology and a structured eight-criterion quality assessment. This review provided a clear foundation for the design of the experimental studies and identified the two gaps: the lack of STW-specific studies among those reviewed and the need for broader, standardized architectural comparisons, which the subsequent chapters addressed directly. The first experimental study evaluated thirteen model configurations under a single unified pipeline with consistent preprocessing, normalization, cross-validation, and evaluation metrics, ensuring that observed performance differences were attributable to architectural differences rather than methodological variation. This approach addresses the review's recommendation for standardized evaluation and extends earlier model comparisons to STW. The dataset used across both experimental studies spans 65 participants aged 19-73 years, comprising young

adults (19–35 years, n = 32), middle-aged adults (36–55 years, n = 15), and older adults (56–73 years, n = 18). This sample size is roughly five times larger than the median sample size (n = 12, range 5–27) across the 22 studies reviewed in Chapter 2, and even the smallest subgroup (middle-aged adults) has 15 participants, which also exceeds that median. This distribution enabled age-group generalization analyses and robustness testing that are not possible with the small, homogeneous cohorts typical of this field. The extension study in Chapter 4 estimates GRF, GRM, and CoP from a single wearable IMU during STW, and its results, which match or exceed prior multi-sensor CoP studies during simpler tasks, establish the feasibility of comprehensive wearable kinetic assessment during a clinically critical transitional movement.

5.6 Limitations of this thesis

Clinical gait laboratories have traditionally focused on steady walking, and its kinetics are well understood. But everyday function involves much more than walking. Transitions such as sit-to-walk, turning, and stand-to-sit are a routine part of daily activity, yet their kinetic details remain poorly understood. Force plates are effective at capturing repeatable foot strikes during steady gait but are harder to use for movements in which the feet are rising or repositioning. An accessible method for measuring these movements could facilitate the study of their control mechanisms and may eventually provide endpoints useful in clinical settings. The framework presented in this thesis is a step in that direction. However, certain limitations restrict its current usefulness. The most consequential is that all participants were healthy adults performing STW in a controlled laboratory setting. No participant had a neurological impairment, musculoskeletal condition, documented fall history, or used a mobility aid. Clinical target populations exhibit kinetic patterns that differ markedly from healthy adults, and it is not known whether the models generalize reliably to these groups. Prospective validation in clinically representative cohorts is therefore necessary before any component of this framework can be used to inform clinical decisions.

Both experimental studies used a single sacrum-mounted IMU because the aim was to develop a simple system that could be deployed with minimal setup. However, this design choice limits the kinematic information available to the model. The substantially lower ML CoP accuracy (NRMSE = 11.48%) compared with AP CoP (NRMSE = 4.46%) directly reflects this

constraint. Mediolateral balance control during STW involves fine-grained inter-limb weight shifting that a single midline sensor cannot adequately encode. The dataset, while broader in age range than most comparable studies, comprises 65 participants, which is relatively small for deep learning, and the fold-to-fold instability observed in some architectures may partially reflect overfitting to this sample size. Furthermore, both experimental studies drew on the same dataset, collected using one sensor system (Delsys Trigno by Delsys Inc., Boston, USA) and laboratory protocol. The participant and age-group holdouts therefore assess transfer between people under these acquisition conditions, while performance with other sensors, laboratories or setups remains unknown.

Real-time deployment on resource-constrained wearable hardware also remains unaddressed. The DS-Conformer uses multi-head self-attention over full-trial sequences, with computational requirements that may limit direct implementation on current low-power embedded processors. The Chapter 4 model combined skip-connection fusion, conformer blocks, and bidirectional recurrent refinement. Parameter counts and GPU inference measurements are reported in Supplementary Material Sections B6 and C2, while the feasibility of edge deployment remains uncertain.

The study is additionally limited to standardized stool-based STW protocols in laboratory conditions. Real-world STW frequently occurs from chairs with armrests, at varied seat heights, and on different surface materials, all of which can alter kinetic profiles. Importantly, this thesis instruments only the STW subtask, which is the first transition of the TUG, rather than the whole test. Therefore, its current clinical utility is limited. The work should be interpreted as a proof of concept showing that kinetic variables, particularly force components such as AP GRF, can be estimated reasonably well from a wearable sensor during the transition from sitting to walking. It does not yet constitute a complete instrumented kinetic TUG. The remaining subtasks, including steady-state walking, walk-to-turn, turning, turn-to-walk, and walk-to-sit, have distinct kinetic characteristics and clinical relevance. A clinically complete kinetic iTUG system would therefore require instrumentation and validation across the full TUG sequence.

5.7 Implications for future work

Transforming this framework into a practical clinical tool will require more than additional data or architectural tuning. Given this thesis's limitations and recent directions in wearable-kinetics research, the main priorities are methodological refinement to address bottlenecks identified here, population extension that treats clinical cohorts as biomechanically distinct rather than simply more variable, and integration into an uncertainty-aware, instrumented TUG.

The weakest output identified in this thesis, ML CoP, appears to reflect a broader field-wide limitation rather than an isolated model failure. Mengarelli et al. (2019), Podobnik et al. (2020) and Wu et al. (2020) reported similar degradation in the ML plane using both sacrum-only and multi-IMU setups, suggesting that the missing element is not more inertial sensors but direct foot-contact information. A practical next step would therefore be to combine the sacrum IMU with a minimal plantar-pressure channel, as suggested by Fukushi et al. (2019), and Moon et al. (2022), to restore the information that a midline sensor cannot capture while preserving clinical simplicity.

The immediate translational barrier is that all participants in this thesis were healthy adults. Future work, therefore, needs to test the models in clinical cohorts, where movement strategies may differ from those seen in healthy participants. Prior studies suggest that older adults and people with balance difficulties may use different STW strategies (Buckley et al., 2009; Perera et al., 2023; van der Kruk et al., 2021). These subtle changes are clinically important and should not be treated simply as prediction errors or variability. Future validation should therefore examine whether the models can preserve these compensatory patterns, rather than only reporting average accuracy across all participants.

Uncertainty quantification would also be important for clinical use because a waveform estimate alone does not indicate to the clinician how much confidence to place in it. For example, if a participant uses an unusual strategy, moves outside the range represented in the training data, or produces noisy sensor signals, the model may still generate a GRF, GRM, or CoP waveform that looks plausible, but without uncertainty estimates, there is no indication that this prediction may be unreliable. Adding uncertainty quantification would allow the system to flag low-confidence trials, identify outputs that should not be used for clinical

interpretation, and distinguish reliable kinetic patterns from estimates that require caution or repeat measurement.

The final priority is to extend kinetic instrumentation across the entire TUG sequence. The present work focused on the first transition of the TUG, the sit-to-walk phase, whereas most existing wearable and gait-analysis studies have focused primarily on straight walking. Future research should therefore move beyond STW and examine the other transitional phases of the TUG, including walk-to-turn, turning, turn-to-walk, and walk-to-sit. This extension would require reliable phase segmentation. Nguyen et al. (2017) showed that inertial sensor signals can be used to identify and separate TUG subtasks. Building on this segmentation approach would allow each phase of the TUG to be detected first, then modeled separately to capture its relevant kinetic outputs. This is necessary before the approach can be considered a complete kinetically instrumented TUG, rather than merely a proof of concept for a single transition.

5.8 Conclusion

In this thesis, the feasibility of wearable kinetic assessment during the sit-to-walk transition of the Timed Up and Go test has been established using a single sacrum-mounted IMU and deep learning. First, a PRISMA-compliant systematic review of 22 studies mapped the state of the art in machine learning-based GRF estimation from wearable sensors, identifying an STW-specific gap among the included studies and a need for broader architectural comparisons under a consistent evaluation pipeline. To address these gaps, thirteen model configurations across seven architectural families were benchmarked on a 65-participant STW dataset, establishing DS-Conformer as the most consistently high-performing and robust architecture and identifying axis-dependent architectural principles that reflect the biomechanical structure of STW. Finally, the DS-Conformer was extended to simultaneously estimate GRF, GRM, and CoP from a single sensor under LOOCV, achieving good accuracy across all eight kinetic components and matching or exceeding prior multi-sensor studies in simpler movement contexts. Taken together, these studies demonstrated not only that wearable kinetic estimation during STW is feasible, but also that clinically meaningful application relies on standardized evaluation, clear treatment of generalization and robustness, and model selection criteria that balance pointwise accuracy with faithful waveform reconstruction.

Within the laboratory dataset studied, the proposed framework estimated STW kinetic information that is not captured by TUG completion time alone. Its performance outside these acquisition conditions remains to be established. The ML CoP component remains the primary source of accuracy issues due to the limited data from a single IMU sensor. As the models are refined through clinical validation, expanded sensor configurations, and larger, more diverse training datasets, the error levels and generalization properties required for routine clinical use remain to be established and tested. Overall, the present research provides a scientific and methodological foundation for developing a portable, clinically validated kinetic assessment tool for the TUG test, with the eventual aim of enabling mechanistic mobility assessment outside the laboratory and supporting fall-risk screening, rehabilitation monitoring, and outcome evaluation in routine clinical settings.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July 25, 2019). Optuna: A Next-generation Hyperparameter Optimization Framework.*KDD '19*

Alcantara, R. S., Edwards, W. B., Millet, G. Y., & Grabowski, A. M. (2022). Predicting continuous ground reaction forces from accelerometers during uphill and downhill running: a recurrent neural network solution. *PeerJ*, *10*, e12752. https://doi.org/10.7717/peerj.12752

Ancillao, A., Tedesco, S., Barton, J., & O'Flynn, B. (2018). Indirect Measurement of Ground Reaction Forces and Moments by Means of Wearable Inertial Sensors: A Systematic Review. *Sensors (Basel)*, *18*(8), 2564. https://doi.org/10.3390/s18082564

Apatsidis, D. P., Solomonidis, S. E., Spence, W. D., & Kelleher, K. J. (2009). The Effect of Impaired Plantar Sensation on Gait in People with Multiple Sclerosis. *International Journal of MS Care*, *11*(1), 25–31. https://doi.org/10.7224/1537-2073-11.1.25

Arevalo, J., Solorio, T., Montes-y-Gómez, M., & González, F. A. (2020). Gated multimodal networks. *Neural Computing and Applications*, *32*(14), 10209–10228. https://doi.org/10.1007/s00521-019-04559-1

Avrutov, V., Bouraou, N., Davydenko, S., Hehelskyi, O., Lakoza, S., Matvienko, O., & Pazdrii, O. (2022). Wavelet De-Noising of Mems Inertial Measurement Unit for Autonomous Latitude Determination. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4042780

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. https://doi.org/10.48550/arXiv.1607.06450

Bach, M. M., Dominici, N., & 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. https://doi.org/10.3389/fspor.2022.1037438

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. https://doi.org/10.48550/arXiv.1409.0473

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *CoRR*, *abs/1803.01271*.

Barela, A. M., de Freitas, P. B., Celestino, M. L., Camargo, M. R., & Barela, J. A. (2014). Ground reaction forces during level ground walking with body weight unloading. *Braz J Phys Ther*, *18*(6), 572–579. https://doi.org/10.1590/bjpt-rbf.2014.0058

Barry, E., Galvin, R., Keogh, C., Horgan, F., & Fahey, T. (2014). Is the Timed Up and Go test a useful predictor of risk of falls in community dwelling older adults: a systematic review and meta-analysis. *BMC Geriatr*, *14*, 14. https://doi.org/10.1186/1471-2318-14-14

Bestaven, E., Petit, J., Robert, B., & Dehail, P. (2013). Center of pressure path during Sit-to-walk tasks in young and elderly humans. *Annals of Physical and Rehabilitation Medicine*, *56*(9-10), 644–651. https://doi.org/10.1016/j.rehab.2013.09.004

Boyer, K. A., Johnson, R. T., Banks, J. J., Jewell, C., & Hafer, J. F. (2017). Systematic review and meta-analysis of gait mechanics in young and older adults. *Exp Gerontol*, *95*, 63–70. https://doi.org/10.1016/j.exger.2017.05.005

Bradley, C., & Harrison, J. E. (2011). Trends in hospitalised injuries due to falls by older people, Australia 1999–2007. *Injury Prevention*, *16*(Suppl 1), A197.192–A197. https://doi.org/10.1136/ip.2010.029215.702

Buckley, T., Pitsikoulis, C., Barthelemy, E., & Hass, C. J. (2009). Age impairs sit-to-walk motor performance. *J Biomech*, *42*(14), 2318–2322. https://doi.org/10.1016/j.jbiomech.2009.06.023

Buckley, T. A., Pitsikoulis, C., & Hass, C. J. (2008). Dynamic postural stability during sit-to-walk transitions in Parkinson disease patients. *Movement Disorders: Official Journal of the Movement Disorder Society*, *23*(9), 1274–1280. https://doi.org/10.1002/mds.22079

Burchi, M., & Vielzeuf, V. (2021). Efficient Conformer: Progressive Downsampling and Grouped Attention for Automatic Speech Recognition. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),

Camargo, J., Ramanathan, A., Flanagan, W., & 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. https://doi.org/10.1016/j.jbiomech.2021.110320

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

Cerfoglio, S., Galli, M., Tarabini, M., Bertozzi, F., Sforza, C., & 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 (Basel)*, *21*(22), 7709. https://doi.org/10.3390/s21227709

Chandler, E. A., Stone, T., Pomeroy, V. M., Clark, A. B., Kerr, A., Rowe, P., Ugbolue, U. C., Smith, J., & Hancock, N. J. (2021). Investigating the Relationships Between Three Important Functional Tasks Early After Stroke: Movement Characteristics of Sit-To-Stand, Sit-To-Walk, and Walking. *Front Neurol*, *12*, 660383. https://doi.org/10.3389/fneur.2021.660383

Chen, G., Patten, C., Kothari, D. H., & Zajac, F. E. (2005). Gait differences between individuals with post-stroke hemiparesis and non-disabled controls at matched speeds. *Gait Posture*, *22*(1), 51–56. https://doi.org/10.1016/j.gaitpost.2004.06.009

Chen, T., & Chou, L.-S. (2013). Altered center of mass control during sit-to-walk in elderly adults with and without history of falling. *Gait & Posture*, *38*(4), 696–701. https://doi.org/10.1016/j.gaitpost.2013.03.007

Chen, T., & Chou, L. S. (2017). Effects of Muscle Strength and Balance Control on Sit-to-Walk and Turn Durations in the Timed Up and Go Test. *Arch Phys Med Rehabil*, *98*(12), 2471–2476. https://doi.org/10.1016/j.apmr.2017.04.003

Childs, J. D., Sparto, P. J., Fitzgerald, G. K., Bizzini, M., & Irrgang, J. J. (2004). Alterations in lower extremity movement and muscle activation patterns in individuals with knee osteoarthritis. *Clin Biomech (Bristol)*, *19*(1), 44–49. https://doi.org/10.1016/j.clinbiomech.2003.08.007

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2015). How much data is needed to train a medical image deep learning system to achieve necessary high accuracy. *arXiv: Learning*.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, 2014–10). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. EMNLP 2014,

Chockalingam, N., Healy, A., & Needham, R. (2016). Interpreting Ground Reaction Forces in Gait. In B. Müller, S. I. Wolf, G.-P. Brueggemann, Z. Deng, A. McIntosh, F. Miller, & W. S. Selbie (Eds.), *Handbook of Human Motion* (pp. 1–15). Springer International Publishing. https://doi.org/10.1007/978-3-319-30808-1_31-1

Costello, K. E., Felson, D. T., Neogi, T., Segal, N., Lewis, C. E., Gross, K. D., Nevitt, M., Lewis, C. L., & Kumar, D. (2019). Ground reaction force patterns in knees with and without pain and radiographic osteoarthritis: descriptive analyses from a large cohort study. *Osteoarthritis and Cartilage*, *27*, S142–S143. https://doi.org/10.1016/j.joca.2019.02.208

Cummings, S. R., & Melton, L. J. (2002). Epidemiology and outcomes of osteoporotic fractures. *Lancet*, *359*(9319), 1761–1767. https://doi.org/10.1016/S0140-6736(02)08657-9

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., & Sagun, L. (2022). ConViT: improving vision transformers with soft convolutional inductive biases*. *Journal of Statistical Mechanics: Theory and Experiment*, *2022*(11). https://doi.org/10.1088/1742-5468/ac9830

Dall, P. M., & Kerr, A. (2010). Frequency of the sit to stand task: An observational study of free-living adults. *Appl Ergon*, *41*(1), 58–61. https://doi.org/10.1016/j.apergo.2009.04.005

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017, August 6,). Language modeling with gated convolutional networks.*ICML'17*

Davis, B. L., Perry, J. E., Neth, D. C., & Waters, K. C. (1998). A Device for Simultaneous Measurement of Pressure and Shear Force Distribution on the Plantar Surface of the

Foot. *Journal of Applied Biomechanics*, *14*(1), 93–104. https://doi.org/10.1123/jab.14.1.93

Della Croce, U., Leardini, A., Chiari, L., & Cappozzo, A. (2005). Human movement analysis using stereophotogrammetry. Part 4: assessment of anatomical landmark misplacement and its effects on joint kinematics. *Gait Posture*, *21*(2), 226–237. https://doi.org/10.1016/j.gaitpost.2004.05.003

DeVita, P., & Hortobagyi, T. (2000). Age causes a redistribution of joint torques and powers during gait. *J Appl Physiol (1985)*, *88*(5), 1804–1811. https://doi.org/10.1152/jappl.2000.88.5.1804

Dimitrov, H., Bull, A. M. J., & Farina, D. (2023). High-density EMG, IMU, kinetic, and kinematic open-source data for comprehensive locomotion activities. *Sci Data*, *10*(1), 789. https://doi.org/10.1038/s41597-023-02679-x

Dingwell, J. B., & Cusumano, J. P. (2000). Nonlinear time series analysis of normal and pathological human walking. *Chaos*, *10*(4), 848–863. https://doi.org/10.1063/1.1324008

Donahue, S. R., & 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*(1), 2339. https://doi.org/10.1038/s41598-023-29314-4

Donahue, S. R., & Hahn, M. E. (2023b). Estimation of ground reaction force waveforms during fixed pace running outside the laboratory. *Front Sports Act Living*, *5*, 974186. https://doi.org/10.3389/fspor.2023.974186

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

Enderling, H., & Wolkenhauer, O. (2020). Are all models wrong? *Comput Syst Oncol*, *1*(1). https://doi.org/10.1002/cso2.1008

Faisal, M. A. A., Mahmud, S., Chowdhury, M. E. H., Khandakar, A., Ahmed, M. U., Alqahtani, A., & Alhatou, M. (2023). Robust and novel attention guided MultiResUnet model for 3D ground reaction force and moment prediction from foot kinematics. *Neural Computing and Applications*, *36*(3), 1105–1121. https://doi.org/10.1007/s00521-023-09081-z

Faivre, A., Dahan, M., Parratte, B., & Monnier, G. (2004). Instrumented shoes for pathological gait assessment. *Mechanics Research Communications*, *31*(5), 627–632. https://doi.org/10.1016/j.mechrescom.2003.10.008

Freiberger, E., Sieber, C. C., & Kob, R. (2020). Mobility in Older Community-Dwelling Persons: A Narrative Review. *Front Physiol*, *11*, 881. https://doi.org/10.3389/fphys.2020.00881

Frykberg, G. E., Aberg, A. C., Halvorsen, K., Borg, J., & Hirschfeld, H. (2009). Temporal coordination of the sit-to-walk task in subjects with stroke and in controls. *Archives of Physical Medicine and Rehabilitation*, *90*(6), 1009–1017. https://doi.org/10.1016/j.apmr.2008.12.023

Fukushi, K., Sekiguchi, Y., Honda, K., Yaguchi, H., & Izumi, S.-I. (2019, 2019–07). Three-dimensional GRF and CoP Estimation during Stair and Slope Ascent/Descent with Wearable IMUs and Foot Pressure Sensors. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),

Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017, August 6,). Convolutional sequence to sequence learning.*ICML'17*

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: continual prediction with LSTM. *Neural Comput*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Giandolini, M., Poupard, T., Gimenez, P., Horvais, N., Millet, G. Y., Morin, J. B., & Samozino, P. (2014). A simple field method to identify foot strike pattern during running. *J Biomech*, *47*(7), 1588–1593. https://doi.org/10.1016/j.jbiomech.2014.03.002

Giardini, M., Turcato, A. M., Arcolin, I., Corna, S., & Godi, M. (2023). Vertical Ground Reaction Forces in Parkinson's Disease: A Speed-Matched Comparative Analysis with Healthy Subjects. *Sensors (Basel)*, *24*(1), 179. https://doi.org/10.3390/s24010179

Goethel, M. F., Becker, K. M., Parolini, F. C. S., Ervilha, U. F., & Vilas-Boas, J. P. (2025). Development of a Tool for Comprehensive Balance Assessment Based on Artificial Intelligence and Anomaly Detection. *Life*, *15*(4). https://doi.org/10.3390/life15040632

Grouvel, G., Carcreff, L., Moissenet, F., & Armand, S. (2023). A dataset of asymptomatic human gait and movements obtained from markers, IMUs, insoles and force plates. *Sci Data*, *10*(1), 180. https://doi.org/10.1038/s41597-023-02077-3

Gruber, N., & Jockisch, A. (2020). Are GRU Cells More Specific and LSTM Cells More Sensitive in Motive Classification of Text? *Front Artif Intell*, *3*, 40. https://doi.org/10.3389/frai.2020.00040

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020, 2020–10–25). *Conformer: Convolution-augmented Transformer for Speech Recognition* Interspeech 2020, https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.pdf

Guo, Y., Storm, F., Zhao, Y., Billings, S. A., Pavic, A., Mazza, C., & Guo, L. Z. (2017). A New Proxy Measurement Algorithm with Application to the Estimation of Vertical Ground Reaction Forces Using Wearable Sensors. *Sensors (Basel)*, *17*(10), 2181. https://doi.org/10.3390/s17102181

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

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

Hartholt, K. A., van Beeck, E. F., Polinder, S., van der Velde, N., van Lieshout, E. M., Panneman, M. J., van der Cammen, T. J., & Patka, P. (2011). Societal consequences of falls in the older population: injuries, healthcare costs, and long-term reduced quality of life. *J Trauma*, *71*(3), 748–753. https://doi.org/10.1097/TA.0b013e3181f6f5e5

Havashinezhadian, S., Chiasson-Poirier, L., Sylvestre, J., & 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*(4), 3120. https://doi.org/10.3390/ijerph20043120

He, K., Zhang, X., Ren, S., & Sun, J. (2016, 2016–06). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),

Higashi, Y., Yamakoshi, K., Fujimoto, T., Sekine, M., & Tamura, T. (2008). Quantitative evaluation of movement using the timed up-and-go test. *IEEE Engineering in Medicine and Biology Magazine*, *27*(4), 38–46. https://doi.org/10.1109/memb.2008.919494

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoskovcová, M., Dušek, P., Sieger, T., Brožová, H., Zárubová, K., Bezdíček, O., Šprdlík, O., Jech, R., Štochl, J., Roth, J., & Růžička, E. (2015). Predicting Falls in Parkinson Disease: What Is the Value of Instrumented Testing in OFF Medication State? *PLoS One*, *10*(10), e0139849. https://doi.org/10.1371/journal.pone.0139849

Hossain, M. S. B., Guo, Z., & 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*(6), 2829–2840. https://doi.org/10.1109/JBHI.2023.3262164

Hu, Z., Zhang, J., & Ge, Y. (2021). Handling Vanishing Gradient Problem Using Artificial Derivative. *IEEE Access*, *9*, 22371–22377. https://doi.org/10.1109/access.2021.3054915

Huang, P., Mostovov, A., Cohen, R., Cadilhac, C., & Pionnier, R. (2025). Comparison of feetme insoles with a motion capture system coupled to force plates for assessing gait and posture. *Sci Rep*, *15*(1), 13476. https://doi.org/10.1038/s41598-025-96878-8

Hughes, M. A., Myers, B. S., & Schenkman, M. L. (1996). The role of strength in rising from a chair in the functionally impaired elderly. *J Biomech*, *29*(12), 1509–1513. https://doi.org/10.1016/S0021-9290(96)80001-7

Ibtehaz, N., & Rahman, M. S. (2019). MultiResUNet : Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. https://doi.org/10.48550/arXiv.1902.04049

Ioffe, S., & Szegedy, C. (2015, July 6, 2015). Batch normalization: accelerating deep network training by reducing internal covariate shift.*ICML'15*

Iosa, M., Cereatti, A., Merlo, A., Campanini, I., Paolucci, S., & Cappozzo, A. (2014). Assessment of waveform similarity in clinical gait data: the linear fit method. *Biomed Res Int*, *2014*, 214156. https://doi.org/10.1155/2014/214156

Janssen, I., Heymsfield, S. B., & Ross, R. (2002). Low relative skeletal muscle mass (sarcopenia) in older persons is associated with functional impairment and physical disability. *J Am Geriatr Soc*, *50*(5), 889–896. https://doi.org/10.1046/j.1532-5415.2002.50216.x

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014, November 3, 2014). Caffe: Convolutional Architecture for Fast Feature Embedding.*MM '14*

Jiang, X., Napier, C., Hannigan, B., Eng, J. J., & Menon, C. (2020). Estimating Vertical Ground Reaction Force during Walking Using a Single Inertial Sensor. *Sensors (Basel)*, *20*(15), 4345. https://doi.org/10.3390/s20154345

Johnson, W. R., Mian, A., Robinson, M. A., Verheul, J., Lloyd, D. G., & Alderson, J. A. (2021). Multidimensional Ground Reaction Forces and Moments From Wearable Sensor Accelerations via Deep Learning. *IEEE Trans Biomed Eng*, *68*(1), 289–297. https://doi.org/10.1109/TBME.2020.3006158

Johnston, L., Patel, V., Cui, Y., & Balaprakash, P. (2025). Revisiting the problem of learning long-term dependencies in recurrent neural networks. *Neural Netw*, *183*, 106887. https://doi.org/10.1016/j.neunet.2024.106887

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jones, G. D., Jones, G. L., James, D. C., Thacker, M., & Green, D. A. (2021). Identifying consistent biomechanical parameters across rising-to-walk subtasks to inform rehabilitation in practice: A systematic literature review. *Gait & Posture*, *83*, 67–82. https://doi.org/10.1016/j.gaitpost.2020.10.001

Jonsdottir, J., Lencioni, T., Gervasoni, E., Crippa, A., Anastasi, D., Carpinella, I., Rovaris, M., Cattaneo, D., & 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. https://doi.org/10.3389/fneur.2020.00668

Karatsidis, A., Bellusci, G., Schepers, H. M., de Zee, M., Andersen, M. S., & Veltink, P. H. (2016). Estimation of Ground Reaction Forces and Moments During Gait Using Only Inertial Motion Capture. *Sensors (Basel)*, *17*(1), 75. https://doi.org/10.3390/s17010075

Kellermeyer, L., Harnke, B., & Knight, S. (2018). Covidence and Rayyan. *Journal of the Medical Library Association*, *106*(4). https://doi.org/10.5195/jmla.2018.513

Kera, T., Kawai, H., Takahashi, J., Hirano, H., Watanabe, Y., Fujiwara, Y., Ihara, K., Kim, H., & Obuchi, S. (2020). Association between ground reaction force in sit-to-stand motion and falls in community-dwelling older Japanese individuals. *Arch Gerontol Geriatr*, *91*, 104221. https://doi.org/10.1016/j.archger.2020.104221

Kerns, J. A., Zwart, A. S., Perez, P. S., Gurchiek, R. D., & McBride, J. M. (2023). Effect of IMU location on estimation of vertical ground reaction force during jumping. *Front Bioeng Biotechnol*, *11*, 1112866. https://doi.org/10.3389/fbioe.2023.1112866

Kerr, A., Durward, B., & Kerr, K. M. (2004). Defining phases for the sit-to-walk movement. *Clinical Biomechanics*, *19*(4), 385–390. https://doi.org/10.1016/j.clinbiomech.2003.12.012

Khemlani, M. M., Carr, J. H., & Crosbie, W. J. (1999). Muscle synergies and joint linkages in sit-to-stand under two initial foot positions. *Clin Biomech (Bristol)*, *14*(4), 236–246. https://doi.org/10.1016/s0268-0033(98)00072-2

Kim, B., Lim, H., & Park, S. (2020). Spring-loaded inverted pendulum modeling improves neural network estimation of ground reaction forces. *J Biomech*, *113*, 110069. https://doi.org/10.1016/j.jbiomech.2020.110069

Kim, D., Jin, Y., Cho, H., Jones, T., Zhou, Y. M., Fadaie, A., Popov, D., Swaminathan, K., & Walsh, C. J. (2025). Learning-based 3D human kinematics estimation using behavioral constraints from activity classification. *Nat Commun*, *16*(1), 3454. https://doi.org/10.1038/s41467-025-58624-6

Kim, J., Koh, J., Kim, Y., Choi, J., Hwang, Y., & Choi, J. W. (2019, 2019). Robust Deep Multi-modal Learning Based on Gated Information Fusion Network.

Kloeckner, J., Visscher, R. M. S., Taylor, W. R., Viehweger, E., & De Pieri, E. (2023). Prediction of ground reaction forces and moments during walking in children with cerebral palsy. *Front Hum Neurosci*, *17*, 1127613. https://doi.org/10.3389/fnhum.2023.1127613

Kmet, L. M., Cook, L. S., & 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., & Buurke, J. H. (2024). Estimation of Kinetics Using IMUs to Monitor and Aid in Clinical Decision-Making during ACL Rehabilitation: A Systematic Review. *Sensors (Basel)*, *24*(7), 2163. https://doi.org/10.3390/s24072163

Lavikainen, J., Vartiainen, P., Stenroth, L., Karjalainen, P. A., Korhonen, R. K., Liukkonen, M. K., & Mononen, M. E. (2024). Gait data from 51 healthy participants with motion capture, inertial measurement units, and computer vision. *Data Brief*, *56*, 110841. https://doi.org/10.1016/j.dib.2024.110841

Leardini, A., Chiari, L., Della Croce, U., & Cappozzo, A. (2005). Human movement analysis using stereophotogrammetry. Part 3. Soft tissue artifact assessment and compensation. *Gait Posture*, *21*(2), 212–225. https://doi.org/10.1016/j.gaitpost.2004.05.002

Lee, M., & Park, S. (2020). Estimation of Three-Dimensional Lower Limb Kinetics Data during Walking Using Machine Learning from a Single IMU Attached to the Sacrum. *Sensors (Basel)*, *20*(21), 6277. https://doi.org/10.3390/s20216277

Lee, S. J., & Hidler, J. (2008). Biomechanics of overground vs. treadmill walking in healthy individuals. *J Appl Physiol (1985)*, *104*(3), 747–755. https://doi.org/10.1152/japplphysiol.01380.2006

Lektip, C., Rungruangbaiyok, C., Nawarat, J., Miyake, E., Aoki, K., Ohtsuka, H., Inaba, Y., Kagaya, Y., & Chaovalit, S. (2024). The Relationship between Components of Postural Control and Locomotive Syndrome in Older Adults. *International Journal of Environmental Research and Public Health*, *21*(10). https://doi.org/10.3390/ijerph21101349

Leporace, G., Batista, L. A., Metsavaht, L., & Nadal, J. (2015, 2015–08). Residual analysis of ground reaction forces simulation during gait using neural networks with different configurations. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),

Leporace, G., Batista, L. A., & Nadal, J. (2018). Prediction of 3D ground reaction forces during gait based on accelerometer data. *Research on Biomedical Engineering*, *34*(3), 211–216. https://doi.org/10.1590/2446-4740.06817

Liang, W., Wang, F., Fan, A., Zhao, W., Yao, W., & Yang, P. (2023). Extended Application of Inertial Measurement Units in Biomechanics: From Activity Recognition to Force Estimation. *Sensors (Basel)*, *23*(9), 4229. https://doi.org/10.3390/s23094229

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

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200209. https://doi.org/10.1098/rsta.2020.0209

Lim, H., Kim, B., & Park, S. (2019). Prediction of Lower Limb Kinetics and Kinematics during Walking by a Single IMU on the Lower Back Using Machine Learning. *Sensors (Basel)*, *20*(1), 130. https://doi.org/10.3390/s20010130

Liu, D., Ma, Y., Wang, J., Hou, M., & Zhang, C. (2024). Convolutional neural network based ground reaction forces and center of pressure estimation during stair walking using multi-level kinematics. *Expert Systems with Applications*, *243*, 122868. https://doi.org/10.1016/j.eswa.2023.122868

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2021, 2021/10/06). Pyraformer: Low-Complexity Pyramidal Attention for Long-Range Time Series Modeling and Forecasting. International Conference on Learning Representations,

Liu, T., Inoue, Y., & Shibata, K. (2010). A wearable ground reaction force sensor system and its application to the measurement of extrinsic gait variability. *Sensors (Basel)*, *10*(11), 10240–10255. https://doi.org/10.3390/s101110240

Lorenz, D. L., & van den Bogert, A. J. (2024). A comprehensive dataset on biomechanics and motor control during human walking with discrete mechanical perturbations. *PeerJ*, *12*, e17256. https://doi.org/10.7717/peerj.17256

Magnan, A., McFadyen, B. J., & St-Vincent, G. (1996). Modification of the sit-to-stand task with the addition of gait initiation. *Gait & Posture*, *4*(3), 232–241. https://doi.org/10.1016/0966-6362(95)01048-3

Mancini, M., & Horak, F. B. (2010). The relevance of clinical balance assessment tools to differentiate balance deficits. *European journal of physical and rehabilitation medicine*, *46*(2), 239–248.

McClay, I. S., & Cavanagh, P. R. (1994). Relationship between foot placement and mediolateral ground reaction forces during running. *Clin Biomech (Bristol)*, *9*(2), 117–123. https://doi.org/10.1016/0268-0033(94)90034-5

Mengarelli, A., Cardarelli, S., Tigrini, A., Marchesini, L., Strazza, A., Fioretti, S., & Verdini, F. (2019, 2019–06). Inertial Sensor Based Estimation of the Center of Pressure During Unperturbed Upright Stance. 2019 IEEE 23rd International Symposium on Consumer Technologies (ISCT),

Mezzarobba, S., Grassi, M., Valentini, R., & Bernardis, P. (2018). Postural control deficit during sit-to-walk in patients with Parkinson's disease and freezing of gait. *Gait & Posture*, *61*, 325–330. https://doi.org/10.1016/j.gaitpost.2018.01.032

Miller, M. F., van der Kruk, E., & Silverman, A. K. (2024). Age and initial position affect movement biomechanics in sit to walk transitions: Whole body balance and trunk control. *Journal of Biomechanics*, *175*, 112256. https://doi.org/10.1016/j.jbiomech.2024.112256

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & Group, P.-P. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Montero-Odasso, M., van der Velde, N., Martin, F. C., Petrovic, M., Tan, M. P., Ryg, J., Aguilar-Navarro, S., Alexander, N. B., Becker, C., Blain, H., Bourke, R., Cameron, I. D., Camicioli, R., Clemson, L., Close, J., Delbaere, K., Duan, L., Duque, G., Dyer, S. M.,…Task Force on Global Guidelines for Falls in Older, A. (2022). World guidelines for falls prevention and management for older adults: a global initiative. *Age Ageing*, *51*(9), afac205. https://doi.org/10.1093/ageing/afac205

Moon, J., Lee, D., Jung, H., Choi, A., & Mun, J. H. (2022). Machine Learning Strategies for Low-Cost Insole-Based Prediction of Center of Gravity during Gait in Healthy Males. *Sensors (Basel)*, *22*(9). https://doi.org/10.3390/s22093499

Morris, M. E., Iansek, R., Matyas, T. A., & Summers, J. J. (1994). The pathogenesis of gait hypokinesia in Parkinson's disease. *Brain: A Journal of Neurology*, *117 (Pt 5)*, 1169–1181. https://doi.org/10.1093/brain/117.5.1169

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

Ngoh, K. J., Gouwanda, D., Gopalai, A. A., & Chong, Y. Z. (2018). Estimation of vertical ground reaction force during running using neural network model and uniaxial accelerometer. *J Biomech*, *76*, 269–273. https://doi.org/10.1016/j.jbiomech.2018.06.006

Nguyen, H., Lebel, K., Boissy, P., Bogard, S., Goubault, E., & Duval, C. (2017). Auto detection and segmentation of daily living activities during a Timed Up and Go task in people with Parkinson's disease using multiple inertial sensors. *Journal of NeuroEngineering and Rehabilitation*, *14*(1), 26. https://doi.org/10.1186/s12984-017-0241-2

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022, 2022/09/29). A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. The Eleventh International Conference on Learning Representations,

Noamani, A., Riahi, N., Vette, A. H., & Rouhani, H. (2023). Clinical Static Balance Assessment: A Narrative Review of Traditional and IMU-Based Posturography in Older Adults and Individuals with Incomplete Spinal Cord Injury. *Sensors (Basel, Switzerland)*, *23*(21), 8881. https://doi.org/10.3390/s23218881

Ochoa, J., Sternad, D., & Hogan, N. (2017). Treadmill vs. overground walking: different response to physical interaction. *J Neurophysiol*, *118*(4), 2089–2102. https://doi.org/10.1152/jn.00176.2017

Ortega-Bastidas, P., Gómez, B., Aqueveque, P., Luarte-Martínez, S., & Cano-de-la-Cuerda, R. (2023). Instrumented Timed Up and Go Test (iTUG)—More Than Assessing Time to Predict Falls: A Systematic Review. *Sensors*, *23*(7), 3426.

Pagnacco, G., Silva, A., Oggero, E., & Berme, N. (2000). Inertially compensated force plate: a means for quantifying subject's ground reaction forces in non-inertial conditions. *Biomedical Sciences Instrumentation*, *36*, 397–402.

Park, S. H. (2018). Tools for assessing fall risk in the elderly: a systematic review and meta-analysis. *Aging Clin Exp Res*, *30*(1), 1–16. https://doi.org/10.1007/s40520-017-0749-0

Perera, C. K., Gopalai, A. A., Gouwanda, D., Ahmad, S. A., & Salim, M. S. B. (2023). Sit-to-walk strategy classification in healthy adults using hip and knee joint angles at gait initiation. *Sci Rep*, *13*(1), 16640. https://doi.org/10.1038/s41598-023-43148-0

Perera, C. K., Hussain, Z., Khant, M., Gopalai, A. A., Gouwanda, D., & Ahmad, S. A. (2024). A Motion Capture Dataset on Human Sitting to Walking Transitions. *Sci Data*, *11*(1), 878. https://doi.org/10.1038/s41597-024-03740-z

Peterson, C. L., Cheng, J., Kautz, S. A., & Neptune, R. R. (2010). Leg extension is an important predictor of paretic leg propulsion in hemiparetic walking. *Gait Posture*, *32*(4), 451–456. https://doi.org/10.1016/j.gaitpost.2010.06.014

Pillay, J., Gaudet, L. A., Saba, S., Vandermeer, B., Ashiq, A. R., Wingert, A., & Hartling, L. (2024). Falls prevention interventions for community-dwelling older adults: systematic review and meta-analysis of benefits, harms, and patient values and preferences. *Syst Rev*, *13*(1), 289. https://doi.org/10.1186/s13643-024-02681-3

Podobnik, J., Kraljić, D., Zadravec, M., & Munih, M. (2020). Centre of Pressure Estimation during Walking Using Only Inertial-Measurement Units and End-To-End Statistical Modelling. *Sensors*, *20*(21), 6136.

Podsiadlo, D., & Richardson, S. (1991). The timed "Up & Go": a test of basic functional mobility for frail elderly persons. *Journal of the American Geriatrics Society*, *39*(2), 142–148. https://doi.org/10.1111/j.1532-5415.1991.tb01616.x

Rawal, Y. R., & Singer, J. C. (2021). The influence of net ground reaction force orientation on mediolateral stability during walking. *Gait Posture*, *90*, 73–79. https://doi.org/10.1016/j.gaitpost.2021.08.009

Rayatdoost, S., Rudrauf, D., & Soleymani, M. (2020, October 22, 2020). Multimodal Gated Information Fusion for Emotion Recognition from EEG Signals and Facial Behaviors.*ICMI '20*

Razak, A. H., Zayegh, A., Begg, R. K., & Wahab, Y. (2012). Foot plantar pressure measurement system: a review. *Sensors (Basel)*, *12*(7), 9884–9912. https://doi.org/10.3390/s120709884

Riley, P. O., Paolini, G., Della Croce, U., Paylo, K. W., & Kerrigan, D. C. (2007). A kinematic and kinetic comparison of overground and treadmill walking in healthy subjects. *Gait Posture*, *26*(1), 17–24. https://doi.org/10.1016/j.gaitpost.2006.07.003

Robinovitch, S. N., Feldman, F., Yang, Y., Schonnop, R., Leung, P. M., Sarraf, T., Sims-Gould, J., & Loughin, M. (2013). Video capture of the circumstances of falls in elderly people residing in long-term care: an observational study. *The Lancet*, *381*(9860), 47–54. https://doi.org/10.1016/S0140-6736(12)61263-X

Rodda, J. M., Graham, H. K., Carson, L., Galea, M. P., & Wolfe, R. (2004). Sagittal gait patterns in spastic diplegia. *J Bone Joint Surg Br*, *86*(2), 251–258. https://doi.org/10.1302/0301-620x.86b2.13878

Roetenberg, D., Luinge, H. J., Baten, C. T., & 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*(3), 395–405. https://doi.org/10.1109/TNSRE.2005.847353

Rong, R., & Kuo, C. (2024). Dynamic Soft Tissue Artifacts during Impulsive Loads: Measurement Errors Vary With Wearable Inertial Measurement Unit Sensor Design. *IEEE Trans Biomed Eng*, *71*(11), 3275–3282. https://doi.org/10.1109/TBME.2024.3416378

Ronneberger, O., Fischer, P., & Brox, T. (2015, 2015). U-Net: Convolutional Networks for Biomedical Image Segmentation.

Rutz, E., Donath, S., Tirosh, O., Graham, H. K., & 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*(3), 455–460. https://doi.org/10.1016/j.gaitpost.2013.01.014

Salehi, A., Khandan, A., Arab Baniasad, M., Baghdadi, S., Farahmand, F., & Zohoor, H. (2018). P 043 – Center of pressure progression and ground reaction forces are altered in cerebral palsy crouch gait. *Gait & Posture*, *65*, 307–308. https://doi.org/10.1016/j.gaitpost.2018.06.198

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Med Inform Decis Mak*, *7*(1), 16. https://doi.org/10.1186/1472-6947-7-16

Scheltinga, B. L., Kok, J. N., Buurke, J. H., & Reenalda, J. (2023). Estimating 3D ground reaction forces in running using three inertial measurement units. *Front Sports Act Living*, *5*, 1176466. https://doi.org/10.3389/fspor.2023.1176466

Schilaty, N. D., Bates, N. A., Ueno, R., & Hewett, T. E. (2021). Filtration Selection and Data Consilience: Distinguishing Signal from Artefact with Mechanical Impact Simulator Data. *Ann Biomed Eng*, *49*(1), 334–344. https://doi.org/10.1007/s10439-020-02562-5

Schoene, D., Wu, S. M., Mikolaizak, A. S., Menant, J. C., Smith, S. T., Delbaere, K., & Lord, S. R. (2013). Discriminative ability and predictive validity of the timed up and go test in identifying older people who fall: systematic review and meta-analysis. *J Am Geriatr Soc*, *61*(2), 202–208. https://doi.org/10.1111/jgs.12106

Schrager, M. A., Kelly, V. E., Price, R., Ferrucci, L., & Shumway-Cook, A. (2008). The effects of age on medio-lateral stability during normal and narrow base walking. *Gait Posture*, *28*(3), 466–471. https://doi.org/10.1016/j.gaitpost.2008.02.009

Shafizadegan, Z., Karimi, M. T., Shafizadegan, F., & Rezaeian, Z. S. (2016). Evaluation of Ground Reaction Forces in Patients with Various Severities of Knee Osteoarthritis. *Journal of Mechanics in Medicine and Biology*, *16*(02), 1650003. https://doi.org/10.1142/s0219519416500032

Shahabpoor, E., Pavic, A., Brownjohn, J. M. W., Billings, S. A., Guo, L. Z., & 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*(6), 1243–1253. https://doi.org/10.1109/TNSRE.2018.2830976

Sharifmoradi, K., Farahpour, N., Bahram, A., Karimi, M. T., & Mazdeh, M. (2015). An assessment of gait spatiotemporal and ground reaction force characteristics of patients with Parkinson compared with normal elderly. *Journal of Research in Rehabilitation Sciences*, *10*(5), 676–686. https://doi.org/10.22122/jrrs.v10i5.1968

Sharma, D., Davidson, P., Muller, P., & Piche, R. (2021). Indirect Estimation of Vertical Ground Reaction Force from a Body-Mounted INS/GPS Using Machine Learning. *Sensors (Basel)*, *21*(4), 1553. https://doi.org/10.3390/s21041553

Shumway-Cook, A., Brauer, S., & Woollacott, M. (2000). Predicting the probability for falls in community-dwelling older adults using the Timed Up & Go Test. *Phys Ther*, *80*(9), 896–903.

Simpkins, C., Soysa, I. B., & Christey, G. (2024). Falls at home: hospital admissions in a health region of Aotearoa New Zealand. *N Z Med J*, *137*(1588), 47–56. https://doi.org/10.26635/6965.6264

Sinitski, E. H., Lemaire, E. D., & Baddour, N. (2015). Evaluation of motion platform embedded with dual belt treadmill instrumented with two force plates. *J Rehabil Res Dev*, *52*(2), 221–234. https://doi.org/10.1682/JRRD.2013.11.0244

Stats NZ-Tatauranga Aotearoa. (2024). *Population*. Retrieved Nov. 5, 2024 from https://www.stats.govt.nz/topics/population/

Sun, T., Li, D., Fan, B., Tan, T., & 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*(7), 3222–3233. https://doi.org/10.1109/JBHI.2023.3268239

van der Kruk, E., Silverman, A. K., Reilly, P., & Bull, A. M. J. (2021). Compensation due to age-related decline in sit-to-stand and sit-to-walk. *Journal of Biomechanics*, *122*, 110411. https://doi.org/10.1016/j.jbiomech.2021.110411

van der Kruk, E., Strutton, P., Koizia, L. J., Fertleman, M., Reilly, P., & Bull, A. M. J. (2022). Why do older adults stand-up differently to young adults?: investigation of compensatory movement strategies in sit-to-walk. *npj Aging*, *8*(1), 13. https://doi.org/10.1038/s41514-022-00094-x

Van Der Velde, G., Laloyaux, H., & Ronsse, R. (2025). Inducing asymmetric gait in healthy walkers: a review. *Front Rehabil Sci*, *6*, 1463382. https://doi.org/10.3389/fresc.2025.1463382

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. u., & Polosukhin, I. (2017). Attention is All you Need.

Vervoort, D., Vuillerme, N., Kosse, N., Hortobágyi, T., & Lamoth, C. J. C. (2016). Multivariate Analyses and Classification of Inertial Sensor Data to Identify Aging Effects on the Timed-Up-and-Go Test. *PLoS One*, *11*(6), e0155984. https://doi.org/10.1371/journal.pone.0155984

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., & Yu, P. (2022). Generalizing to Unseen Domains: A Survey on Domain Generalization. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 1–1. https://doi.org/10.1109/tkde.2022.3178128

Weiss, A., Herman, T., Plotnik, M., Brozgol, M., Maidan, I., Giladi, N., Gurevich, T., & Hausdorff, J. M. (2010). Can an accelerometer enhance the utility of the Timed Up & Go Test when evaluating patients with Parkinson's disease? *Med Eng Phys*, *32*(2), 119–125. https://doi.org/10.1016/j.medengphy.2009.10.015

Williams, S. E., Gibbs, S., Meadows, C. B., & 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*(3), 370–373. https://doi.org/10.1016/j.gaitpost.2011.06.003

Winter, D. A. (2009). Kinetics: Forces and Moments of Force. In *Biomechanics and Motor Control of Human Movement* (pp. 107–138). https://doi.org/10.1002/9780470549148.ch5

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/s0893-6080(05)80023-1

Wouda, F. J., Giuberti, M., Bellusci, G., Maartens, E., Reenalda, J., van Beijnum, B. F., & Veltink, P. H. (2018). Estimation of Vertical Ground Reaction Forces and Sagittal Knee Kinematics During Running Using Three Inertial Sensors. *Front Physiol*, *9*, 218. https://doi.org/10.3389/fphys.2018.00218

Wu, C. C., Chen, Y. J., Hsu, C. S., Wen, Y. T., & Lee, Y. J. (2020). Multiple Inertial Measurement Unit Combination and Location for Center of Pressure Prediction in Gait. *Front Bioeng Biotechnol*, *8*, 566474. https://doi.org/10.3389/fbioe.2020.566474

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, *31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Zanardi, A. P. J., da Silva, E. S., Costa, R. R., Passos-Monteiro, E., Dos Santos, I. O., Kruel, L. F. M., & Peyre-Tartaruga, L. A. (2021). Gait parameters of Parkinson's disease compared with healthy controls: a systematic review and meta-analysis. *Sci Rep*, *11*(1), 752. https://doi.org/10.1038/s41598-020-80768-2

Zarzeczny, R., Nawrat-Szołtysik, A., Polak, A., Maliszewski, J., Kiełtyka, A., Matyja, B., Dudek, M., Zborowska, J., & Wajdman, A. (2017). Aging effect on the instrumented Timed-Up-and-Go test variables in nursing home women aged 80–93 years. *Biogerontology*, *18*, 651–663.

Appendices

Appendix A. Supplementary Material to Chapter 2

A1. Systematic search strategy and yield by database

Table S2-1 Systematic search strategy and yield by database

Last Searched On	Database	Search Terms used	Years Searched	Publication type	No. Of Results	Language
12/04/2025	PUBMED	(((Center of pressure) OR (CoP) OR (Ground reaction force) OR (GRF))) AND ((Inertial Measurement Unit) OR (IMU))	2011-2025	Research Articles	161	English
12/04/2025	IEEE Xplore	("Center of Pressure" OR "CoP" OR "Ground Reaction Force" OR "GRF") AND ("Inertial Measurement Unit*" OR "IMU*")	2011-2025	Research Articles	207	English
13/04/2025	Science Direct	("Center of Pressure" OR "CoP" OR "Ground Reaction Force" OR "GRF") AND ("Inertial Measurement Unit" OR "IMU")	2011-2025	Research Articles	117	English
13/04/2025	Scopus	(((Center AND of AND pressure) OR (CoP) OR (Ground AND Reaction AND force) OR (GRF))) AND ((inertial AND measurement AND unit) OR (IMU))	2011-2025	Research Articles	143	English

A2. Quality Assessment Questions

1: Is the study's objective or research question clearly defined?

2: Is the study design evident and appropriate?

3: Are the characteristics of the subjects or the details of input variables and ground truth sufficiently described?

4: Are outcomes clearly defined, and assessment methods thoroughly reported?

5: Is the approach used for data processing and model training clearly described?

6: Did the study report any statistical analyses to evaluate model performance?

7: Are results reported in sufficient detail?

8: Are the conclusions supported by the results?

Table S2-2 Evaluation of each study against eight quality-assessment questions. Scores are 0 (no), 0.5 (partial) or 1 (yes)

Study	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total
(Guo et al., 2017)	1	1	1	1	1	1	1	1	8
(Shahabpoor et al., 2018)	1	1	0.5	1	0.5	1	1	0.5	6.5
(Cerfoglio et al., 2021)	1	1	1	1	1	1	1	1	8
(Jiang et al., 2020)	1	1	1	1	1	1	1	1	8
(Lee & Park, 2020)	1	1	1	1	1	1	1	1	8
(Wouda et al., 2018)	1	1	1	1	1	1	1	1	8
(Lim et al., 2019)	1	1	1	1	1	1	1	1	8
(Ngoh et al., 2018)	1	1	1	1	1	1	1	1	8
(Leporace et al., 2015)	1	0.5	1	1	1	0.5	1	0.5	6.5
(Leporace et al., 2018)	1	1	1	1	1	1	1	1	8
(Kim et al., 2020)	1	0.5	1	1	0.5	1	1	0.5	6.5
(Scheltinga et al., 2023)	1	1	1	1	1	1	1	1	8
(Havashinezhadian et al., 2023)	1	1	1	1	1	0.5	1	1	7.5
(Bach et al., 2022)	1	1	1	1	1	1	1	1	8
(Donahue & Hahn, 2023a)	1	1	1	1	1	0.5	0.5	0.5	6.5
(Donahue & Hahn, 2023b)	1	1	1	1	1	0.5	1	1	7.5
(Wu et al., 2020)	1	1	1	1	1	1	1	1	8
(Alcantara et al., 2022)	1	1	1	1	1	1	1	1	8
(Sun et al., 2023)	1	0.5	1	1	0.5	0.5	1	0.5	6
(Dorschky et al., 2020)	1	1	1	1	1	1	1	1	8
(Johnson et al., 2021)	1	1	1	0.5	0.5	1	0.5	0.5	6
(Hossain et al., 2023)	1	1	1	1	1	1	1	1	8

A3. Participant Characteristics of included studies

Table S2-3 Participant Characteristics of included studies

Study	Total	Male	Female	Mean Age	SD Age	Mean Height	SD Height	Mean Weight	SD Weight
(Guo et al., 2017)	9	6	3	28	3	Not Reported	Not Reported	Not Reported	Not Reported
(Shahabpoor et al., 2018)	6	6	0	21	1	182	8	77	16
(Cerfoglio et al., 2021)	11	6	5	24.5	1.7	170.5	10.3	65	11.3
(Jiang et al., 2020)	9	9	0	27	8	176	7	72	9
(Lee & Park, 2020)	20	12	8	24.7	3.2	166.7	7.8	60.1	8.9
(Wouda et al., 2018)	8	8	0	25.1	5.2	183.7	4.5	77.7	9.4
(Lim et al., 2019)	7	7	0	25	2.9	168.8	7.5	72.1	7.7
(Ngoh et al., 2018)	7	7	0	21.3	0.5	174.9	6.6	63	6.1
(Leporace et al., 2015)	17	11	6	27.1	3.4	Not Reported	Not Reported	84.3	4.5
(Leporace et al., 2018)	17	11	6	27.1	3.4	Not Reported	Not Reported	84.3	4.5
(Kim et al., 2020)	7	7	0	25	2.9	168.8	7.5	72.1	7.7
(Scheltinga et al., 2023)	12	8	4	31.6	9	178	11	73.7	17.5
(Havashinezhadian et al., 2023)	45	Not Reported	Not Reported	36.66 (Healthy) & 65.63 (MKOA)	4.24 (Healthy) & 7.32 (MKOA)	172 (Healthy) & 165 (MKOA)	8 (Healthy) & 12 (MKOA)	73.11 (Healthy) & 80.51 (MKOA)	16.45 (Healthy) & 15.27 (MKOA)
(Bach et al., 2022)	21	13	8	20.8	1	181.7	10.3	71.1	9.8
(Donahue & Hahn, 2023a)	13	8	5	23.2	Not Reported	167.8	Not Reported	65	Not Reported

(Donahue & Hahn, 2023b)	15	9	6	23.6	Not Reported	178.3	Not Reported	73.5	Not Reported
(Wu et al., 2020)	5	3	2	24	2	168	2	64	11
(Alcantara et al., 2022)	19	10	9	29	9	173	9	68.1	9.9
(Sun et al., 2023)	16	16	0	23.2	1.4	177	4	72.8	9.7
(Dorschky et al., 2020)	10	10	0	27	2.6	182	5	76.9	8.6
(Johnson et al., 2021)	458,372 motion trials	59.9%	40.1%	Not Reported	Not Reported	177	10.1	74.9	34.1
(Hossain et al., 2023)	20	12	8	21.7	3.65	170	7	68.21	11.52

Note: Lee and Park (2020) recruited 20 participants (12F, 8M), but 3 were excluded because of equipment failure; the analyzed-subset sex split was not reported separately.

A4. Sensor Characteristics of included studies

Table S2-4 Sensor Characteristics of included studies

Study	No. of IMUs used	Sensor Type	Location of IMUs	Sampling Rate of IMU	Components of IMU used	Reference Measurement
(Guo et al., 2017)	3	Triaxial (Opal, APDM)	Fifth lumbar vertebra (L5), seventh cervical vertebra (C7) and forehead.	128 Hz	Accelerometer	pressure-sensing insoles (F-Scan 3000E by Tekscan™) 128 Hz
(Shahabpoor et al., 2018)	12	Triaxial (Opal)	7th cervical vertebra (C7), 5th lumbar vertebrae (L5), upper arms, forearms, thighs, shanks, and fourth metatarsals	128 Hz to 100 Hz	Accelerometer	Custom instrumented treadmill equipped with two separate belts and six-axis force plates for each foot. 1000 Hz to 100Hz
(Cerfoglio et al., 2021)	3	Triaxial (Physilog® by Gait Up Ltd., Lausanne, Switzerland)	Sacrum, right thigh and shank	128 Hz	Accelerometer, Gyroscope	Two force platforms (AMTI OR6-7 Advanced Mechanical Technology, Inc., Watertown, MA, USA) 200 Hz to 128Hz
(Jiang et al., 2020)	4	Triaxial (MTw Awinda)	Foot, distal shank, distal thigh, proximal thigh	100 Hz	Accelerometer, Gyroscope	split-belt force plate-instrumented treadmill (Bertec Corporation, Columbus, OH, USA) 100 Hz
(Lee & Park, 2020)	1	Triaxial (TrignoTM Avanti sensors, Delsys, Natick, MA, USA)	Sacrum	148 downsampled to 100Hz	Accelerometer	two force plates (FP 6012®, Bertec, Columbus, OH, USA) underneath a split-belt treadmill - 200Hz downsampled to 100Hz

(Wouda et al., 2018)	3	Triaxial (Xsens MVN Link)	Lower legs and pelvis	240 Hz to 120 Hz	Accelerometer, Gyroscope	S-Mill instrumented treadmill (ForceLink, Culemborg, Netherlands) 1,000 Hz.
(Lim et al., 2019)	1	Triaxial (EBIMU-9DOFV4, Acceleration sensitivity: ±16 g Resolution: 0.001 g)	Sacrum	100Hz	Accelerometer	2 force plates (Bertec FP 6012®) 200 Hz
(Ngoh et al., 2018)	1	Triaxial (Opal APDM Inc.) (Acceleration: ± 6 g m/s², Angular velocity: ± 2000 deg/s, Magnetic field: ± 6 Gauss)	right shoe above the third metatarsal	128 Hz	Accelerometer	Instrumented treadmill equipped with two force plates (Mercury, H/P Cosmos Sports and Medical GmbH), sampled at 1000 Hz. GRF waveform was later time-normalized to 100 stance-phase data points during preprocessing.
(Leporace et al., 2015)	1	Triaxial (± 6 g, model MMA7260Q, Freescale, USA)	distal and anterior part shank	1 kHz	Accelerometer	A force plate (AccuGait, AMTI, USA) embedded in the middle of the walkway 1 kHz
(Leporace et al., 2018)	1	Triaxial (± 6 g, model MMA7260Q, Freescale, USA)	distal and anterior part of the shank	1000 Hz	Accelerometer	A force plate (AccuGait, AMTI, USA) was embedded into the middle of the walkway 1000 Hz
(Kim et al., 2020)	1	**1** Simulated from motion capture system (Motion Analysis Hawk)	Sacrum	100Hz	Not reported	Force plates embedded in a treadmill (Bertec FP 6012), sampled at 200 Hz. Stance phases were later resampled to 200 points during preprocessing.

(Scheltinga et al., 2023)	8	Triaxial (Xsens MTx, Xsens Technologies, Enschede, The Netherlands)	feet, (proximal) tibias, thighs, pelvis, and trunk)	240Hz	Accelerometer	(Y-mill, Motek Medical, Amsterdam, The Netherlands) - 2048Hz to 240Hz
(Havashinezhadian et al., 2023)	5	Triaxial (Physilog®4 and 5, Lausanne, Switzerland)	(1) the top of the shoe (TS); (2) the heel (H); (3) above the medial malleolus (MM); (4) the middle and front of the tibia (MFT); and (5) the medial of the shank close to the knee joint on SINGLE LEG	200 Hz	Accelerometer	Instrumented Treadmill (Bertec, Columbus, OH, USA, 1000 Hz
(Bach et al., 2022)	2	Triaxial ± 16 g (Mini Wave Plus, Zerowire; Cometa, Bareggio, Italy)	right and left tibia (shank)	143 Hz	Accelerometer	instrumented dual-belt treadmill 1 kHz to 143 Hz
(Donahue & Hahn, 2023a)	3	Triaxial (Casio, Tokyo, Japan)	2 on dorsal aspect of each foot and 1 on sacrum)	200Hz downsampled to 100Hz	Accelerometer, Gyroscope	Loadsol insole force sensors (Novel Electronics, St. Paul, MN, USA) bilaterally in between the foot and the shoe at 100Hz
(Donahue & Hahn, 2023b)	3	Triaxial (Casio, Tokyo, Japan) - Only Acc and Gyro	3 (2 on dorsal aspect of each foot and 1 on sacrum)	200Hz downsampled to 100Hz	Accelerometer, Gyroscope	Loadsol insole force sensors (Novel Electronics, St. Paul, MN, USA) bilaterally in between the foot and the shoe at 100Hz

(Wu et al., 2020)	4	Triaxial (Delsys Inc, USA)	mounted to four corners of the second force plate for measuring acceleration and gyroscope	148 Hz	Accelerometer, Gyroscope	Two force plates (Advanced Mechanical Technology Inc., USA) 1037 Hz re-sampled at 148 Hz
(Alcantara et al., 2022)	3	Biaxial	two on the right shoe and one on the sacrum.	2,000 Hz downsampled to 500 Hz	Accelerometer	Instrumented treadmill (Bertec, OH, USA) 2,000 Hz downsampled to 500 Hz
(Sun et al., 2023)	8	Triaxial (Xsens Technologies B.V., Enschede, The Netherlands)	Chest (trunk, at the fifth thoracic vertebra) Waist (mid-point between the left and right anterior superior iliac spine) Thighs (mid-point between the left anterior superior iliac spine and femur medial epicondyle, on both left and right legs) Shanks (one-third point between the femur medial epicondyle and tibia apex of the medial malleolus, near the proximal end of the tibia, on both left and	100Hz	Accelerometer, Gyroscope	2 force plates embedded in the ground - 1000 Hz to 100 Hz

			right legs) Feet (second metatarsal, on both left and right feet)			
(Dorschky et al., 2020)	4	Triaxial (Portabiles GmbH, Erlangen, Germany) ±16g and ±2000 deg/s & Simulated Data	Lower back Lateral right thigh Lateral right shank Over the second to fourth metatarsal on the right foot	1000 Hz to 100 Hz	Accelerometer, Gyroscope	Kistler force plate (Kistler Instruments Corp, Winterthur, CH), sampled at 1000 Hz. Gait cycles were later resampled to 100 time points during preprocessing.
(Johnson et al., 2021)	5	Triaxial (DTS-3D 518 (Noraxon Inc., Scottsdale, AZ)	pelvis, bilateral thighs, and bilateral shanks	IMUs 250 Hz	Accelerometer	Two force platforms (AMTI OR6-7 Advanced Mechanical Technology, Inc., Watertown, MA, USA)
(Hossain et al., 2023)	3	Triaxial (Yost, Ohio, USA)	thigh, shank, and foot	200 Hz	Accelerometer, Gyroscope	(force plates) 1000 Hz Resampled to 200Hz

A5. Preprocessing and Feature extraction methods used in the included studies

Table S2-5 Preprocessing and Feature extraction methods used in the included studies

Study	Sync Technique	Filters on IMU data	Filters on Ground Truth Data	Reported Preprocessing Steps	Features
(Guo et al., 2017)	vertical jump at the beginning of each trial and cross correlation	Not Applied	Not Applied	Synchronization Exclusion of Resting/Transitory Data Splitting of Acceleration Data: The acceleration data from the IMUs were split into two components for each foot (left and right) using membership functions.	64 features selected through iterative Orthogonal Forward Regression (iOFR) from Raw acceleration components (ax, ay, az) of each IMU Time-lagged versions of each acceleration component. Second-order polynomial combinations of these time-lagged acceleration components. From this extensive pool, the iOFR algorithm identified the 64 most important features
(Shahabpoor et al., 2018)	Not reported	low-pass, zero-lag, fourth-order Butterworth digital filter with a cutoff frequency of 12 Hz	low-pass, zero-lag, fourth-order Butterworth digital filter with a cutoff frequency of 12 Hz	**Data Reorientation:** The acceleration signals measured by each IMU were reoriented from the IMU's local coordinate system (SCS) to the laboratory-fixed coordinate system (LCS) using the IMU's orientation	Raw Accelerometer values

				data (quaternions). The IMU data at the 7th cervical vertebra (C7) was aligned to match the vertical (z-axis) direction of the laboratory coordinate system. The body-fixed coordinate system (BCS) was established to represent the walking directions (anteroposterior, vertical, and medial-lateral). **Resampling and Synchronization Filtering**	
(Cerfoglio et al., 2021)	Cross-correlation between IMU and GRF from force platform	fourth-order low-pass Butterworth filter with a cutoff frequency of 32 Hz	Not Applied	Synchronization Filtering Resampling **Cutting the data:** Both IMU and GRF signals were cut to isolate the first landing phase of each jump (defined as the period when the vertical GRF exceeded 10 N until the feet lifted off for the jump). **Data Normalization:** Ground truth data was normalized by the participants' bodyweight for further analysis.	**Raw IMU data** 19 (time + 3 IMU x (3 Acc x 3 Gy))

(Jiang et al., 2020)	Two synchronization impulse signals were generated: one at the beginning and one at the end of each trial.	median filter with a five-sample window	Low-pass filtered at a cutoff frequency of 10 Hz.	Filtering Feature Extraction: Normalization: The extracted features were normalized to a uniform range between 0 and 1	110-ms window (equivalent to 11 samples at a 100 Hz sampling rate) **10 temporal domain IMU features** Root Mean Square (RMS) Sum of Absolute Value Mean Absolute Deviation Variance Wavelength Slope Sign Changes Simple Square Integral Mean Wavelet with db7 Difference Absolute Standard Deviation Value Average Amplitude Change Log Detector Coefficients of Linear Fit Coefficients of Parabolic Fit
(Lee & Park, 2020)	simultaneous recording and data alignment during post-processing	fifth-order zero-phase Butterworth low-pass filter with a cutoff frequency of 10 Hz	fifth-order zero-phase Butterworth low-pass filter with a cutoff frequency of 20 Hz	**Filtering**, **Downsampling**, **Normalization** (IMU data (except for time) was normalized to a range of 0 to 1., GRF data was normalized by bodyweight), **Segmentation**:(The IMU data was segmented into the stance phases of the gait cycle), **Interpolation:** To	10 × 1 column vector was fed to 10 nodes of the input layer, which consisted of time t and the corresponding 3D kinematics data of the sacrum, which acted as a proxy of the CoM, such as (t, xm, ym, zm, vx, vy, vz, ax, ay, az); xm, ym, and zm are the ML, AP, and vertical positions, and v and a are the velocity and acceleration of the sacrum

				standardize the length of the dataset across all walking speeds and participants, the data for each stance phase was linearly interpolated to 201 points.	measured at a specific time t, respectively.
(Wouda et al., 2018)	using an external analog synchronization signal.	None	6th order Butterworth filter with a cutoff frequency of 20 Hz	Filtering, Downsampling **Segmentation**: Only steady-state running phase.	Raw IMU data
(Lim et al., 2019)	Manually aligning the IMU acceleration profiles with the GRF profiles	low-pass Butterworth filter with 10 Hz cutoff	fifth-order low-pass Butterworth filter with a cutoff frequency of 10 Hz	**Segmentation:** The IMU data was segmented into individual steps, focusing on the stance phase. This was done by identifying key gait events such as heel strike (HS) and toe-off (TO) based on IMU acceleration profiles. The CoM apex, HS, and TO were detected using characteristic features from the IMU data. **Filtering** **Drift Removal:** The velocity and displacement of the CoM were calculated by integrating the IMU	7 input nodes for the ANN, representing the CoM kinematics: (t, x, y, vx, vy, ax, ay).

				acceleration data, which required drift removal. This was done using the assumption of steady walking and random error. The drift was removed linearly over the duration from one CoM apex to the next. **Normalization:** All kinematic data, such as acceleration, velocity, and displacement, were normalized by their maximum values so that the input values to the neural network ranged from 0 to 1.	
(Ngoh et al., 2018)	using an external trigger	2nd order Butterworth low-pass filter with a cutoff frequency of 10 Hz	None	Filtering, Stance-phase normalization to 100 data points GRF data normalized to the participant's bodyweight	Raw acceleration (X-axis only)
(Leporace et al., 2015)	Simultaneous Data Collection using BIOPAC system	2nd order Butterworth low-pass filter with a 25 Hz cutoff frequency	2nd order Butterworth low-pass filter with a 25 Hz cutoff frequency applied in both the forward and reverse	Filtering: Normalization: The data was normalized by bodyweight	**14** 3D accelerometer data (normalized by bodyweight), Relative velocity and displacement (first and second integration of Acc data) First derivative of the accelerometer signal, Stance duration,

			directions to prevent phase shifts		Time point expressed as a percentage of stance time.
(Leporace et al., 2018)	Not reported	second-order Butterworth low-pass filter with a cutoff frequency of 25 Hz applied in both the forward and reverse directions to avoid phase shift	second-order Butterworth low-pass filter with a cutoff frequency of 25 Hz, applied in both the forward and reverse directions to avoid phase shift	**Filtering** **Stance Phase** Detection: period when the vertical GRF exceeded 10 N	**14** 3D acceleration (normalized by bodyweight). 3D velocity (obtained by single integration of the acceleration data). 3D displacement (obtained by double integration of the acceleration data). The first derivative of the 3D acceleration signals. Stance duration (in seconds). Time point of the stance phase (expressed as a percentage of stance time - The data from each gait cycle were interpolated into 51 values, representing 0% to 100% of the stance phase.
(Kim et al., 2020)	N/A	fifth-order low-pass Butterworth filter (with a cutoff frequency of 10 Hz)	fifth-order low-pass Butterworth filter (with a cutoff frequency of 10 Hz)	**Filtering,** **Compensation for Anteroposterior (AP) Displacement:** The anteroposterior displacement was adjusted by adding the product of treadmill speed and time to align the motion data with the corresponding	**Method 1: CoM Kinematics** **7** x(t): Anteroposterior displacement of the CoM. y(t): Vertical displacement of the CoM. v_x(t): Anteroposterior velocity of the CoM. v_y(t): Vertical velocity of the

				GRF data. **Normalization:** All input data was normalized using a 0-1 normalization Stance-phase data were resampled to 200 points.	CoM. a_x(t): Anteroposterior acceleration of the CoM. a_y(t): Vertical acceleration of the CoM. t: Time. **Method 2: Spring-Loaded Inverted Pendulum (SLIP) model-derived forces** F_x(t): Anteroposterior force derived from the SLIP model. F_y(t): Vertical force derived from the SLIP model. V_gait: Gait speed.
(Scheltinga et al., 2023)	The cross-correlation between the sum of the sensor free acceleration along the axis of both tibias and vertical GRF was used to find the temporal offset between the two systems	3rd order bidirectional Butterworth low-pass filter with a cutoff frequency at 10 Hz	sixth order, zero-phase shift low-pass Butterworth filter with a cutoff frequency of 30 Hz	**Filtering** **Downsampling** **3D sensor free acceleration:** gravity subtracted acceleration in the global frame **Synchronization** **Segmentation:** Segmentation into stance phase and flight phase	**Direct Model** Raw IMU Data **Hybrid Model** Hybrid model used the physical estimate as input, in addition to the sensor acceleration data physical estimate, a previously developed model based on Newton's second law was used to estimate the vertical GRF from IMU data from the pelvis and both tibias
(Havashinezhadian et al., 2023)	cross-correlation between IMU	-	-	**Synchronization** **Normalization:** Ground truth	**Raw Acc data** **Input combinations:** There

	and GRF from instrumented treadmill			data was normalized by each participant's bodyweight to standardize the data	were five different combinations of acceleration axes: AP,ML,V,AP-ML,ML-V,AP-V,AP-ML-V
(Bach et al., 2022)	Not reported	2nd order Butterworth High-pass filter Applied bi-directionally to remove low-frequency drift from the accelerometer signals	1st order Savitzky-Golay filter with Window size of 9 samples	**Principal Component Analysis (PCA):** The triaxial accelerometer data (from X, Y, Z axes) was transformed using PCA to align with the direction of maximum variance. Only the first principal component was retained. **High-Pass Filtering Integration:** The first principal component was integrated twice to obtain velocity and position data. **Normalization:** The resulting acceleration, velocity, and position data was normalized to a range of [0,1]. The vertical GRF data was normalized by converting it to a z-score, where each trial's GRF was normalized based on its mean and standard deviation.	3-dimensional matrix combining the: Normalized acceleration (a) Normalized velocity (v) Normalized position (p)
(Donahue & Hahn, 2023a)	Foot stomps between each trial	4th-order low-pass zero-lag Butterworth filter	4th-order low-pass zero-lag Butterworth	Synchronization Filtering **Drift Correction:** For	Raw 3D IMU and Gy values 1-sec temporal window

		with a cutoff frequency (fc) of 35 Hz	filter with a cutoff frequency (fc) of 20 Hz	footfalls with drifting baselines, corrections were made by identifying swing phases and setting forces during those phases to 0 BW **Normalization:** The kinetic data from the force-sensing insoles were normalized to each participant's bodyweight (BW) **Clock Drift Resolution:** To handle clock drift between the IMUs and the force-sensing insoles, an iterative corrections algorithm.	
(Donahue & Hahn, 2023b)	Foot stomps between each trial	4th-order low-pass zero-lag Butterworth filter with a cutoff frequency (fc) of 35 Hz	4th-order low-pass zero-lag Butterworth filter with a cutoff frequency (fc) of 20 Hz	**Coordinate System Alignment:** The IMU data was initially processed with a Kalman filter to align the local IMU coordinate system to gravity. **Synchronization Filtering** **Clock Drift Resolution:** To handle clock drift between the IMUs and the force-sensing insoles, an iterative corrections algorithm.	Raw 3D IMU and Gy values 4-sec temporal window
(Wu et al., 2020)	synchronized by the Vicon Nexus system (Vicon	third-level discrete wavelet transform	Not reported	DWT (IMU), Zero Offset Removal (IMU), Downsampling (Ground	**Raw IMU data** 24 (4 IMUs * 3Acc * 3Gy)

	Motion System, 2018)			truth) Synchronization	
(Alcantara et al., 2022)	N/A	4th order low-pass Butterworth filter with a 20 Hz cutoff.	4th order low-pass Butterworth filter with a 30 Hz cutoff	Downsampling Filtering Windowing: overlapping six-frame (12ms) windows	**13 features are:** Mean vertical sacral acceleration (over a 12ms window) Standard deviation of vertical sacral acceleration Range of vertical sacral acceleration Mean anteroposterior sacral acceleration (over a 12 ms window) Standard deviation of anteroposterior sacral acceleration Range of anteroposterior sacral acceleration Body mass of the subject Height of the subject Running speed Slope of the treadmill Percentage of steps classified as rearfoot strikes Percentage of steps classified as midfoot strikes Percentage of steps classified as forefoot strikes classification of foot strikes based on a previously

					validated method (Giandolini et al., 2014)
(Sun et al., 2023)	External synchronization, allowing all the systems to start recording data simultaneously	zero-lag second-order low-pass Butterworth filter with a cutoff frequency of 15 Hz.	zero-lag second-order low-pass Butterworth filter with a cutoff frequency of 15 Hz	**Downsampling** **Filtering** **Normalization:** The GRF values were normalized by dividing by the subject's bodyweight (BW).	Raw IMU data (Acc and Gy)
(Dorschky et al., 2020)	Two light barriers, placed 2 meters apart, were used to measure the participants' speed. This speed data helped synchronize the IMU data and force plate data by aligning the sensor signals and the measured GRF.	Not reported	Not reported	Gait cycles resampled to 100 time points using linear interpolation **Normalization**: GRFs were normalized to bodyweight (%)	100 x 12: 100 time points 12 features (AP acceleration, Longitudinal acceleration, Medial-Lateral angular velocity) x 4 sensors
(Johnson et al., 2021)	based on detecting foot-strike (FS) and cross- correlation	Not reported	Not reported	**Downsampling:** IMU data downsampled to 250Hz **Acceleration Alignment:** Both simulated (training) and recorded (test) accelerations required alignment. PCA	the aligned 4D (t + ax + ay + az) acceleration inputs were flattened into 2D images by representing the five sensor locations on the horizontal axis, stance-normalized time

				using Singular Value Decomposition (SVD) re-orientation was applied to align accelerations with the global coordinate system **Stance Phase** **Normalization:** The GRF and GRM data collected from the force plates were time-normalized to the stance phase **Smoothing:** Data smoothing was reported but the exact methods applied were not reported	frames upwards on the vertical axis, and by use of the Python SciPy imsave function to map the 3D accelerations onto the RGB colorspace
(Hossain et al., 2023)	Not reported	Not reported	Not reported	**Normalization**: Ground truth data was normalized by each participant's bodyweight to standardize the data **Segmentation:** Level-ground, ramp, and stair walking tasks, the IMU, joint moments, and GRF data was segmented for a full gait cycle.	Raw IMU data Window length of 100 samples,

A6. Details of Machine/Deep models used in the included studies

Table S2-6 Details of Machine/Deep models used in the included studies. N/A indicates information not reported by the study authors

Study	Model Type	Size of layers	Hyperparameters	Predicted Direction
(Guo et al., 2017)	NARMAX	N/A	Time Lag (L): Maximum time lag of 18 samples Polynomial Order: 2	V
(Shahabpoor et al., 2018)	NARMAX	Inputs: Acceleration signals from selected IMU placements (e.g., C7, L5, and thighs). Outputs: Estimated GRF in three directions (vertical, anteroposterior, and mediolateral).	**Polynomial Order** NARMAX model was tested with both first-order and second-order polynomial structures to capture nonlinearities in the relationship between input (IMU acceleration) and output (GRF). In the study, it was found that the first-order model provided satisfactory results, with minimal additional benefit from using a second-order polynomial. **Lag Terms:** The number of lagged terms of the input signals included in the model was another critical parameter. The study allowed up to five lagged terms to account for time dependencies between the input acceleration signals and the output GRF. Including more than one lag term did improve the model accuracy slightly, but the increase was marginal (about 1% improvement).	ML, AP, V

(Cerfoglio et al., 2021)	RNN (nonlinear autoregressive network)	**Input Layer** 19 input features: Including acceleration, angular velocity, and the frame reference index (from the 3 IMU sensors). **Hidden Layers** 2 fully connected hidden layers. Each hidden layer contained 1000 neurons. The activation function for the hidden layers was a linear activation function. **Output Layer** The output layer produced predictions for six target variables: Right knee flexion/extension (FE) moment Right knee adduction/abduction (AA) moment Right knee rotation (R) moment Vertical GRF AP GRF ML GRF	**Training Algorithm:** Resilient Backpropagation algorithm **epochs:** 5000 **Regularization:** regularization factor of 0.5 was applied to prevent overfitting and improve the model's ability to generalize	ML, AP, V
(Jiang et al., 2020)	Random Forest Regression model	N/A	Number of trees: 50 Number of leaves: 5	V
(Lee & Park, 2020)	ANN	**1 input layer**. **1 hidden layer** had 20 nodes with a sigmoidal function as the activation transfer function **1 output layer**	**Loss Function:** Mean Squared Error (MSE) **Optimization Algorithm:** Levenberg-Marquardt optimization for backpropagation. **Stopping Criterion:** Training stopped if the validation set MSE failed to decrease or remained constant for 20 epochs.	ML, AP, V

(Wouda et al., 2018)	ANN	**Hidden Layers:** 2 The first layer had 250 neurons. The second layer had 100 neurons.	**Stopping criteria:** Trained for up to 2,000 iterations or if the gradient did not decrease for 6 consecutive iterations, or The gradient became smaller than 1×10^{-6}	V
(Lim et al., 2019)	ANN	**Input Layer:** 7 input nodes representing the time and kinematic features from the IMU (t, x, y, vx, vy, ax, ay). **Hidden Layer:** 20 nodes with sigmoid activation. **Output Layer:** 11 output nodes 6 segment angles (thigh, shank, foot for both stance and swing legs). 3 joint torques (hip, knee, ankle for the stance leg). horizontal and vertical GRFs	Loss Function: Mean Square Error (MSE) Optimizer: Adam optimizer learning rate: Not mentioned batch size: Not mentioned number of epochs: Not mentioned	AP, V
(Ngoh et al., 2018)	Feed-forward neural network (NN)	**Input layer:** The input consisted of 100 normalized acceleration data points from the IMU. **Hidden layer:** 10 sigmoid neurons. **Output layer:** Consisted of 100 linear neurons, corresponding to the estimated vertical GRF over the 100 data points representing the stance phase.	**Training algorithm:** Levenberg-Marquardt backpropagation algorithm. **Loss Function:** Mean Squared Error (MSE)	V

(Leporace et al., 2015)	MLP	**Single MLP Network with Three Outputs:** **Input Layer:** 14 neurons **Hidden Layer:** 8 neurons (determined as the optimal number based on minimizing the mean absolute deviation (MAD)). Hyperbolic tangent (tanh) activation function **Output Layer:** 3 neurons (corresponding to the three components of GRF). Linear activation function **Three Separate MLP Networks with One Output Each:** **Input Layer:** 14 neurons **Hidden Layer:** 8 neurons for AP, 5 neurons for V, 10 neurons for ML. Hyperbolic tangent (tanh) activation function **Output Layer:** 1 neuron for each network (predicting one GRF). Linear activation function	Training Algorithm: Levenberg-Marquardt backpropagation algorithm Stopping Criteria: Generalization error (the performance on the validation set) was monitored to prevent overfitting. Minimum gradient was used as an additional stop criterion during training.	ML, AP, V
(Leporace et al., 2018)	Multilayer perceptron (MLP)	**Input Layer** 14 features **Hidden Layer** 1 hidden layer. The number of neurons in the hidden layer was optimized by testing different configurations, with the optimal model having 10 neurons. activation function: tanh **Output Layer:**	**Training Algorithm:** Levenberg-Marquardt backpropagation algorithm **Stopping Criteria:** The training process was stopped based on the generalization error observed in the validation set and the minimum gradient	ML, AP, V

		The output layer had three neurons, one for each of the predicted GRF components AP, ML and V activation function: linear		
(Kim et al., 2020)	ANN	**Input Layer:** For Method 1 (CoM kinematics): The input layer had 7 nodes For Method 2 (SLIP transform): The input layer had 3 nodes **Hidden Layer:** 20 nodes in both methods. Sigmoid activation function **Output Layer:** The output layer had 2 nodes: one for the vertical GRF and one for AP GRF	**Loss Function:** Mean square error (MSE) **Optimizer:** Adam optimizer **Learning Rate:** 0.05. **Epochs**: 1000.	AP, V
(Scheltinga et al., 2023)	ANN and hybrid ANN (physics informed)	Both models had 2 layers with 100 neurons and ReLU activation function.	Loss function: Mean squared error Optimizer: Adam Epochs: 1,000 epochs, Batch size: 250. Stopping Criteria: If validation loss did not decrease for more than 100 epochs.	ML, AP, V
(Havashine zhadian et al., 2023)	echo states network (ESN)	**Input Layer:** Raw Acc data **Reservoir Layer:** 100 neurons, activation function: tanh, recurrent connections within the reservoir, with a feedback weight matrix (W) connecting reservoir nodes to past output states, enabling the model to retain memory of previous inputs. Leaking rate (α) of 0.1053 **Output Layer**	Number of nodes (N): 100 Leaking rate (α): 0.1053 Spectral radius (ρ): 0.7471 (controls the stability of the reservoir's dynamics) Sparsity (ρ): 0.21 (controls the connectivity of the reservoir) Regularization parameter (γ): 1×10^{-6} (for ridge regression to avoid overfitting)	ML, AP, V

(Bach et al., 2022)	Echo State Network	**Input Layer:** 3 nodes (acceleration, velocity, position) **Reservoir Layer:** 1,000 nodes. Sparsely connected in a random manner, with connectivity determined by the spectral radius set to 0.5. **Output Layer:** 1 node (predicted vGRF).	Leak Rate (γ): 0.5 Noise (ε): Small noise of 10^-4 was added to the reservoir during training to avoid overfitting Training Algorithm: Ridge Regression Time Constant (τ): 1 Connectivity (C): sparse random matrix, with its weights normalized	V
(Donahue & Hahn, 2023a)	BD - LSTM	**LSTM Layer:** The LSTM layer had 19 hidden units **Fully Connected Layer:** sigmoid activation function. **Regression Layer:** linear activation function.	**Learning Rate:** The initial learning rate was set to 0.001, which is the default learning rate for the ADAM optimizer used in the model. **Batch Size:** Batch size of 50 **Number of Epochs:** 100 epochs.	V
(Donahue & Hahn, 2023b)	LSTM	LSTM Layer: The LSTM layer had 42 hidden units Fully Connected Layer: 1 - sigmoid activation function. Regression Layer: linear activation function.	**Initial Learning Rate, Gradient Decay Factor, Squared Gradient Decay Factor, L2 Regularization:** All Default values recommended by MATLAB **4-second temporal window.** The temporal window for input data was tested in ranges from 0.5 to 5 seconds. The 4-second window was found to be the most accurate for estimating GRF waveforms.	V

(Wu et al., 2020)	LSTM	**Input:** The model used 24 features, which included triaxial acceleration and angular velocity data from four IMU sensors placed under the walking surface. **Hidden Layers** The model for predicting ML and AP GRF had three hidden layers with 150, 250, and 100 neurons, respectively. The model for predicting vGRF used a simpler architecture with 100, 200, and 100 neurons. **Batch Normalization Layer**	**Epochs**: 40 for ML and AP, and 110 for vertical GRF. **Loss Function:** Mean Squared Error (MSE) **Optimizer:** RMSprop	ML, AP, V
(Alcantara et al., 2022)	BD-LSTM	Bidirectional LSTM Layer: **Multilayer Perceptron (MLP):** First Fully Connected Layer: 128 neurons Second Fully Connected Layer: 384 neurons Third Fully Connected Layer: 320 neurons Each layer in the MLP used a rectified linear unit (ReLU) activation function, and a dropout rate was applied to prevent overfitting: Input Layer Dropout: 20% Bidirectional LSTM Output Dropout: 40% **output layer**	Batch Size: 32 Learning Rate: 0.001 Loss Function: Mean Squared Error (MSE) Optimizer: Adaptive Moment Estimation (Adam optimizer) Epochs: A maximum of 1,000 epochs was allowed. Training stopped early if the mean squared error did not decrease by 0.001 BW (bodyweight) for 30 consecutive epochs (early stopping criterion). Dropout Rate: Input Layer: 20% Output of Bidirectional LSTM Layer: 40% These hyperparameters were selected using the Hyperband hyperparameter optimization algorithm,	V
(Sun et al., 2023)	LSTM	**LSTM Layer:** 130 units **Fully Connected (FC) Layers:**	**Weight Initialization:** The weights in the LSTM and FC layers were initialized using the Glorot normal initializer (Xavier initialization)	V

		3 fully connected (dense) layers: First FC Layer: 60 neurons with ReLU activation. Second FC Layer: 30 neurons with ReLU activation. Third FC Layer: 1 neuron with tanh activation	**Optimizer:** The Stochastic Gradient Descent (SGD) with momentum optimizer	
(Dorschky et al., 2020)	CNN	**2 convolutional layers for feature extraction:** Layer 1: Convolutional layer with a filter size of 5 × 3 and 64 filters. Layer 2: Convolutional layer with a filter size of 5 × 3 and 128 filters. Both layers applied a rectified linear unit (ReLU) activation function. **max-pooling layer** Pooling size was 2 × 2, reducing the dimensionality of the feature maps. **2 dense layers** The first dense layer had 100 nodes with a ReLU activation function. The output layer had 100 nodes (corresponding to the 100 time points per gait cycle) with a linear activation function	Loss function: mean squared error (MSE) L2 regularization Factor: 0.001 Batch Size: 64 Learning Rate: 0.001 Number of Epochs: 1000 Optimizer: ADAM	AP, V
(Johnson et al., 2021)	CNN CaffeNet ResNet-50	**Hidden Layers:** Not reported **Output Layer:** output layer of each CNN was modified from a SoftMax binary to a Euclidean loss layer, which turned the CNN into a multivariate regression network.	Not reported	ML, AP, V

(Hossain et al., 2023)	Kinetics-FM-DLR-Ensemble-Net, (bagging)	**GRU-Net:** GRU-1: 512 units, Dropout: 0.1 (0.5 for Pearson r loss function) GRU-2: 256 units, Dropout: 0.1 (0.5 for Pearson r loss function) **GRU-Conv1D-Net:** **Conv1D Block:** Conv1D-1: 256 filters, Kernel size: 3, Max-pooling: 2 Conv1D-2: 256 filters, Kernel size: 3, Max-pooling: 2 Conv1D-3: 512 filters, Kernel size: 3, Max-pooling: 2 Conv1D-4: 512 filters, Kernel size: 3, Max-pooling: 2 **Fully Connected (FC) layers:** FC-1: 64 units, Dropout: 0.25 FC-2: 32 units, Dropout: 0.25 **GRU-Conv2D-Net:** **Conv2D Block:** Conv2D-1: 256 filters, Kernel size: 5x3, Max-pooling: 2x2 Conv2D-2: 256 filters, Kernel size: 5x3, Max-pooling: 2x2 Conv2D-3: 512 filters, Kernel size: 5x3, Max-pooling: 2x2 Conv2D-4: 512 filters, Kernel size: 5x3, Max-pooling: 2x1 **Fully Connected (FC) layers:** FC-1: 64 units, Dropout: 0.25 FC-2: 32 units, Dropout: 0.25 **Fusion Module (FM):**	**Optimizer:** Adam **Batch Size:** 64 **Epochs:** 40 **Loss Function:** A weighted combination of RMSE and Pearson r	ML, AP, V

	FC-1: 128 units, Dropout: 0.4 (for Pearson r loss function) FC-2: 6 units

A7. Tasks performed, validation strategies, and GRF waveform evaluation metrics used in the included studies

Table S2-7 Tasks performed, validation strategies, reporting window (stance phase or whole gait cycle), and GRF waveform evaluation metrics used in the included studies

Study	Task Performed	Cross validation or dataset split approach	Accuracy reported	GRF Waveform evaluation method
(Guo et al., 2017)	**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	dataset split approach 50% Training, 50% testing	entire gait cycle	rRMSE, Cross-Correlation Coefficient
(Shahabpoor et al., 2018)	Treadmill walking	Model trained using data from participants 1-4, and validation was performed using two methods: Intra-subject validation: Three tests from S1-S4 were held out and used for validation. Inter-subject validation: Two tests from participants S5 and S6, who were not part of the training set, were used to validate the model's ability to generalize to new subjects.	entire gait cycle	NRMSE

(Cerfoglio et al., 2021)	Vertical drop jumps	seven participants' data and tested on the data from four participants. 20-fold cross-validation	-	RMSE
(Jiang et al., 2020)	Treadmill Walking for 1 min at five different speeds: 0.4 m/s, 0.7 m/s, 1.0 m/s, 1.3 m/s, 1.6 m/s	**Intra-Participant Testing:** The data from each participant was split into two halves: First half of the data (from all five walking speeds) was used for training. Second half of the data was used for testing. **Inter-Participant Testing:** A leave-one-out cross-validation method was used: Data from 8 participants was used for training the model. Data from the remaining 1 participant was used for testing.	Stance phase only	RMSE, NRMSE, Pearson correlation
(Lee & Park, 2020)	Treadmill Walking at four different speeds (0.7, 1.0, 1.3, and 1.6 m/s) for 2 minutes at each speed.	leave-one-subject-out (LOO) validation	Stance Only	NRMSE, Coefficient of determination (R^2)
(Wouda et al., 2018)	Treadmill Running at three different speeds—10, 12, and 14 km/h for 3 minutes at each speed.	Two evaluation scenarios: **Single-subject evaluation:** trained on 10 and 14 km/h and tested on 12 km/h within the same participant **Multi-subject evaluation:** trained on multiple participants and	Stance Only	RMSE, Mean Pearson's correlation

		evaluated on a different participant		
(Lim et al., 2019)	One-minute treadmill walking 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.	Leave-One-Out Cross-Validation (LOO-CV) 6 subjects were used for training, and the remaining 1 subject was used for validation	Stance Only	NRMSE, R^2
(Ngoh et al., 2018)	Treadmill Walking and Running	random dataset split approach **Training set:** 280 randomly selected stance phases were used to train the neural network model. **Validation and testing sets:** 120 stance phases were used for both validation and testing.	Stance Only	RMSE, Pearson correlation
(Leporace et al., 2015)	Overground Walking - 8m	k-fold cross-validation approach: Total folds **4**	stance phase of walking	Mean Absolute Deviation (MAD), Pearson correlation
(Leporace et al., 2018)	8m walk overground	Leave-One-Out Cross-Validation (LOO-CV)	stance phase only.	MAD, Normalized MAD, Pearson correlation
(Kim et al., 2020)	treadmill walking at three speed (1.0 to 2.3 m/s)	Leave-One-Out Cross-Validation (LOO-CV) 6 subjects were used for training, and the remaining 1 subject was used for validation	stance phase only	NRMSE, Pearson correlation

(Scheltinga et al., 2023)	Treadmill Running	Leave one-subject-out cross-validation	stance phase only	RMSE, rRMSE, Pearson correlation
(Havashinezhadian et al., 2023)	Instrumented treadmill Walking	70% of the data for training, 30% testing	entire gait cycle	MAE, NRMSE
(Bach et al., 2022)	Treadmill walking and running Walking speed: 1.24 ± 0.12 m/s Running speed: 2.20 ± 0.14 m/s	**Random trial-based split repeated 100 times:** Training Set: 50% Validation Set: 25% Testing Set: 25% **Leave-M-Out Cross-Validation:** repeated for M = 1 to M = 6. cross-validation was repeated 42 times	entire gait cycle	NRMSE, R^2
(Donahue & Hahn, 2023a)	Outdoor Running: Level ground, Incline and decline	Leave-One-Out Cross-Validation (LOOCV) 13 FOLD	Stance phase only	RMSE
(Donahue & Hahn, 2023b)	Outdoor Running - 400m. 5 trials with increasing velocity (2.33 - 5.36 m/s)	Leave One Out Cross Validation (LOOCV) for training For the hyperparameter optimization process, 70% of the data was used for training, 15% for validation, and 15% for testing.	entire gait cycle	RMSE, Bland-Altman Analysis
(Wu et al., 2020)	Overground walking	70% of the data for training, 30% testing	Stance Only	NRMSE
(Alcantara et al., 2022)	Treadmill Running five different slopes: 0°, +5°, -5°, +10°, and -10°.	Leave-One-Subject-Out (LOSO) cross-validation method,		RMSE, rRMSE

	three speeds: 2.5 m/s, 3.33 m/s, and 4.17 m/s for each slope.			
(Sun et al., 2023)	single-leg and double-leg drop landing tasks from a height of 30 cm	Leave-One-Out Cross-Validation (LOO-CV) 15 subjects were used for training, and the remaining 1 subject was used for validation	stance phase of the drop landings	RMSE, rRMSE, R^2
(Dorschky et al., 2020)	Overground walking and running at six speed bands. **Walking Speeds:** 0.9-1.0 m/s, 1.2-1.4 m/s, 1.8-2.0 m/s **Running Speeds:** 3.1-3.3 m/s, 3.9-4.1 m/s, 4.7-4.9 m/s	Fixed independent 3-subject test set. Model selection and cross-validation were performed within the 7-subject training set	stance phase only	RMSE, Pearson correlation
(Johnson et al., 2021)	**Running:** slow (2-3 m/s), moderate (4-5 m/s), and fast (>6 m/s) 90-degree **sidestep**	The training dataset was generated from a large biomechanics archive collected over 17 years from The University of Western Australia (UWA). This dataset consisted of 458,372 motion trials The test dataset came from a separate, independent data collection performed at Liverpool John Moores University (LJMU), using real sensor accelerations from five athletes performing running and sidestepping movements.	stance phase only	Cross correlation coefficient

		Validation: Single-fold validation on the independent test set.		
(Hossain et al., 2023)	Treadmill walking, Level-ground walking, ramp walking, and Stair Climb	leave-one-subject-out cross-validation	whole gait cycle	NRMSE, Pearson correlation

Appendix B. Supplementary Material to Chapter 3

B1. Per-Fold Stability Analysis

The per-fold cross-validation results revealed substantial inter-fold variability for certain architectures. Table S3-1 to Table S3-3 show the mean RMSE (%BW) across subjects within each fold for the best variant from each architectural family.

Table S3-1 Per-fold RMSE (%BW) stability across the five cross-validation folds for vertical GRF prediction.

Model	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Range
DS-Conformer	3.66	3.99	4.25	4.21	4.13	0.59
ConformerUNet	3.83	4.95	4.38	4.59	4.45	1.13
TCN	5.21	5.60	5.58	5.94	5.61	0.73
MultiResUNet	6.22	6.29	9.72	5.80	6.96	3.92
BiGRU	6.18	10.02	7.95	5.92	5.73	4.30
Hybrid-Concat	5.08	5.30	6.84	10.25	10.17	5.17
Hybrid-Gated	5.52	4.33	6.52	13.79	9.65	9.47
DS-Transformer	10.02	10.33	7.72	9.80	9.39	2.61
ResCorr (frozen)	7.91	8.07	15.23	14.66	13.42	7.32
MRU-AttnGRU	34.08	5.24	6.29	9.36	6.99	28.84
ResCorr (joint)	4.81	50.94	6.29	14.48	4.76	46.18
BiLSTM	19.51	27.00	36.34	29.94	22.94	16.83
ShallowCNN	23.86	24.74	37.54	28.38	35.08	13.69

Note. Values are mean RMSE (%BW) across subjects within each fold. Range = max - min across folds.

Table S3-2 Per-fold RMSE (%BW) stability across the five cross-validation folds for anteroposterior GRF prediction.

Model	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Range
ResCorr (joint)	1.59	1.65	1.72	1.60	1.83	0.24
DS-Conformer	1.56	2.09	1.65	1.69	1.77	0.53
BiGRU	2.24	2.08	2.89	2.58	2.28	0.81
DS-Transformer	3.04	2.78	2.42	2.35	2.60	0.69
Hybrid-Gated	1.51	1.88	1.67	6.31	1.86	4.81
Hybrid-Concat	1.40	1.72	4.93	1.43	3.81	3.53
TCN	2.22	2.34	5.07	1.98	1.88	3.19
ConformerUNet	2.37	2.22	6.18	1.79	2.06	4.39
ResCorr (frozen)	3.00	2.84	3.72	2.96	3.00	0.87
MultiResUNet	5.98	1.80	1.98	2.31	7.69	5.89
BiLSTM	4.45	4.73	4.39	4.58	4.50	0.34
ShallowCNN	4.87	5.00	5.11	4.89	5.11	0.24
MRU-AttnGRU	7.27	6.16	6.13	6.51	9.02	2.89

Note. Values are mean RMSE (%BW) across subjects within each fold. Range = max - min across folds.

Table S3-3 Per-fold RMSE (%BW) stability across the five cross-validation folds for mediolateral GRF prediction.

Model	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Range
ResCorr (joint)	0.78	0.90	1.55	1.07	0.84	0.76
DS-Conformer	1.07	1.12	1.07	1.15	1.02	0.13
DS-Transformer	2.32	1.48	1.34	1.49	1.33	0.99
TCN	1.18	1.07	4.28	1.41	1.12	3.21
Hybrid-Gated	2.89	3.67	0.94	2.12	1.14	2.73
ConformerUNet	2.65	2.43	2.70	1.06	2.72	1.66
Hybrid-Concat	0.98	2.43	2.74	2.88	2.68	1.90
ResCorr (frozen)	2.57	1.96	2.76	2.08	2.64	0.80
BiLSTM	2.38	2.30	2.64	2.75	2.60	0.44
ShallowCNN	2.46	2.34	2.61	2.77	2.59	0.43
BiGRU	2.59	2.40	2.68	2.86	2.62	0.46
MultiResUNet	0.91	3.19	5.17	3.73	8.64	7.73
MRU-AttnGRU	5.90	3.34	5.63	5.17	4.50	2.56

Note. Values are mean RMSE (%BW) across subjects within each fold. Range = max - min across folds.

For vertical GRF prediction, DS-Conformer was the most stable model across folds, with a fold-to-fold range of only 0.59 %BW. ResCorr (joint) exhibited the largest fold-to-fold variability (range 46.18 %BW). Hybrid-Gated produced a single outlier fold (fold 3) with RMSE of 13.8 %BW, roughly 2× their median; MRU-AttnGRU produced a single outlier fold (fold 0) with RMSE of 34.1 %BW, roughly 5× their median, while their remaining folds clustered tightly.

For anteroposterior GRF prediction, ResCorr (joint) was the most stable model across folds, with a fold-to-fold range of only 0.24 %BW. MultiResUNet exhibited the largest fold-to-fold variability (range 5.89 %BW). Overall, the AP axis showed lower absolute fold-to-fold variability than the vertical axis, reflecting the smaller amplitude of AP forces.

For mediolateral GRF prediction, DS-Conformer was the most stable model across folds, with a fold-to-fold range of only 0.13 %BW. MultiResUNet exhibited the largest fold-to-fold variability (range 7.73 %BW).

B2. Median and Interquartile Range Summary of GRF Prediction Performance

Although the per-fold stability analysis indicated that the overall ranking of the leading models was generally maintained across folds, it also uncovered significant variability in performance for some force components and architectures. Several metrics exhibited wide variation across folds and held-out subjects, suggesting that the distributions were not consistently well represented by a symmetric mean-based summary. In such cases, mean values can be disproportionately affected by a small number of unstable folds or poorly predicted subjects.

Therefore, the GRF results are also summarized using the median and interquartile range (IQR). The median offers a reliable estimate of typical prediction performance, while the IQR measures the spread of the middle 50% of observations and is less affected by extreme values than the standard deviation. Reporting median [Q1, Q3] thus gives a clearer picture of the typical subject-level accuracy for each model and complements the per-fold stability analysis by showing how consistently each model performed across the entire cross-validation sample.

Table S3-4 Five-Fold Subject-Wise Cross-Validation Results for Vertical GRF Prediction Reported as Median [Q1, Q3]

Model	RMSE (%BW)	MAE (%BW)	R^2	Pearson r	CCC
DS-Conformer	2.65 [2.25, 5.22]	1.93 [1.65, 3.46]	1.00 [0.98, 1.00]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]
ConformerUNet	3.01 [2.31, 5.26]	2.27 [1.55, 3.57]	0.99 [0.98, 1.00]	1.00 [0.99, 1.00]	1.00 [0.99, 1.00]
TCN	4.39 [3.79, 5.46]	3.29 [2.75, 4.08]	0.99 [0.98, 0.99]	1.00 [0.99, 1.00]	0.99 [0.99, 1.00]
BiGRU	5.51 [4.43, 7.28]	4.18 [3.27, 5.04]	0.98 [0.97, 0.99]	0.99 [0.98, 1.00]	0.99 [0.98, 0.99]
Hybrid (concat)	5.44 [4.36, 9.15]	3.94 [3.10, 5.31]	0.98 [0.95, 0.99]	0.99 [0.98, 1.00]	0.99 [0.97, 0.99]
MultiResUNet	4.78 [3.85, 5.64]	3.50 [2.95, 4.21]	0.99 [0.98, 0.99]	0.99 [0.99, 1.00]	0.99 [0.99, 1.00]
Hybrid (gated)	5.04 [3.51, 8.29]	3.48 [2.58, 4.37]	0.98 [0.96, 0.99]	0.99 [0.98, 1.00]	0.99 [0.98, 1.00]
DS-Transformer	7.36 [6.33, 11.55]	4.66 [3.88, 5.92]	0.97 [0.91, 0.98]	0.98 [0.96, 0.99]	0.98 [0.96, 0.99]
MRU-AttnGRU	6.13 [4.26, 14.95]	5.06 [3.15, 9.21]	0.98 [0.87, 0.99]	0.99 [0.93, 1.00]	0.99 [0.93, 0.99]
ResCorr (frozen)	7.62 [6.24, 13.21]	4.73 [3.84, 6.79]	0.96 [0.89, 0.98]	0.98 [0.95, 0.99]	0.98 [0.95, 0.99]
ResCorr (joint)	7.10 [4.00, 16.63]	4.46 [3.00, 9.50]	0.97 [0.82, 0.99]	0.98 [0.91, 1.00]	0.98 [0.91, 1.00]
BiLSTM	23.79 [19.52, 29.81]	14.74 [10.87, 20.07]	0.69 [0.45, 0.77]	0.86 [0.72, 0.89]	0.80 [0.68, 0.87]
ShallowCNN	27.47 [23.59, 32.44]	16.16 [12.41, 20.27]	0.55 [0.39, 0.66]	0.76 [0.64, 0.82]	0.74 [0.61, 0.81]

Table S3-5 Five-Fold Subject-Wise Cross-Validation Results for AP GRF Prediction Reported as Median [Q1, Q3]

Model	RMSE (%BW)	MAE (%BW)	R^2	Pearson r	CCC
DS-Conformer	1.35 [0.99, 2.24]	0.91 [0.68, 1.44]	0.95 [0.86, 0.97]	0.98 [0.94, 0.99]	0.97 [0.93, 0.98]
ResCorr (joint)	1.10 [0.77, 2.05]	0.91 [0.60, 1.35]	0.96 [0.88, 0.98]	0.99 [0.94, 0.99]	0.98 [0.94, 0.99]
TCN	1.96 [1.68, 2.33]	1.36 [1.16, 1.71]	0.89 [0.84, 0.92]	0.95 [0.94, 0.97]	0.94 [0.92, 0.96]
BiGRU	2.18 [1.71, 2.82]	1.49 [1.25, 1.91]	0.86 [0.78, 0.91]	0.94 [0.89, 0.96]	0.93 [0.88, 0.95]
DS-Transformer	2.21 [1.66, 3.05]	1.40 [0.99, 1.88]	0.84 [0.73, 0.93]	0.93 [0.87, 0.96]	0.92 [0.86, 0.96]
ResCorr (frozen)	2.84 [2.41, 3.35]	1.92 [1.62, 2.32]	0.79 [0.66, 0.84]	0.90 [0.85, 0.92]	0.88 [0.82, 0.92]
Hybrid (concat)	2.41 [0.92, 4.30]	1.59 [0.68, 2.72]	0.83 [0.51, 0.98]	0.93 [0.72, 0.99]	0.91 [0.66, 0.99]
Hybrid (gated)	1.77 [0.91, 4.06]	1.15 [0.69, 2.76]	0.91 [0.53, 0.98]	0.95 [0.77, 0.99]	0.95 [0.75, 0.99]
ConformerUNet	2.03 [1.53, 3.35]	1.49 [1.10, 2.22]	0.88 [0.63, 0.94]	0.95 [0.86, 0.97]	0.94 [0.85, 0.97]
MultiResUNet	2.96 [1.42, 6.61]	1.79 [1.02, 4.40]	0.78 [-0.07, 0.94]	0.90 [0.15, 0.97]	0.89 [0.09, 0.97]
BiLSTM	4.41 [4.02, 4.84]	2.85 [2.64, 3.29]	0.44 [0.36, 0.50]	0.70 [0.65, 0.75]	0.61 [0.51, 0.66]
ShallowCNN	4.84 [4.49, 5.22]	3.03 [2.83, 3.39]	0.32 [0.23, 0.40]	0.59 [0.52, 0.65]	0.53 [0.43, 0.58]
MRU-AttnGRU	6.74 [6.02, 7.69]	4.56 [4.12, 5.59]	-0.21 [-0.60, -0.01]	0.08 [-0.05, 0.21]	0.03 [-0.01, 0.06]

Table S3-6 Five-Fold Subject-Wise Cross-Validation Results for ML GRF Prediction Reported as Median [Q1, Q3]

Model	RMSE (%BW)	MAE (%BW)	R^2	Pearson r	CCC
DS-Conformer	0.99 [0.82, 1.26]	0.65 [0.56, 0.80]	0.87 [0.79, 0.91]	0.95 [0.91, 0.96]	0.93 [0.89, 0.96]
ResCorr (joint)	0.62 [0.36, 1.29]	0.45 [0.27, 0.82]	0.95 [0.78, 0.98]	0.98 [0.90, 0.99]	0.97 [0.88, 0.99]
TCN	1.10 [0.99, 1.37]	0.71 [0.65, 0.88]	0.83 [0.72, 0.88]	0.93 [0.89, 0.94]	0.91 [0.87, 0.94]
DS-Transformer	1.40 [1.13, 2.08]	0.86 [0.73, 1.17]	0.71 [0.35, 0.84]	0.87 [0.60, 0.92]	0.84 [0.57, 0.92]
Hybrid (gated)	1.90 [1.11, 3.10]	1.15 [0.74, 2.16]	0.48 [-0.47, 0.86]	0.76 [0.26, 0.95]	0.76 [0.24, 0.93]
ConformerUNet	2.45 [2.11, 2.77]	1.71 [1.38, 1.91]	0.08 [-0.01, 0.21]	0.39 [0.30, 0.51]	0.23 [0.18, 0.32]
ResCorr (frozen)	2.33 [2.04, 2.76]	1.60 [1.29, 1.79]	0.12 [0.05, 0.42]	0.48 [0.36, 0.67]	0.19 [0.14, 0.64]
BiLSTM	2.55 [2.22, 2.84]	1.64 [1.44, 1.88]	0.18 [0.06, 0.26]	0.53 [0.41, 0.60]	0.29 [0.21, 0.36]
ShallowCNN	2.57 [2.33, 2.84]	1.63 [1.48, 1.89]	0.15 [0.06, 0.20]	0.43 [0.36, 0.50]	0.26 [0.22, 0.31]
Hybrid (concat)	2.52 [2.05, 2.93]	1.63 [1.26, 1.96]	0.06 [0.00, 0.12]	0.34 [0.20, 0.49]	0.10 [0.04, 0.16]
MultiResUNet	3.77 [2.92, 5.52]	2.76 [2.14, 4.61]	-0.84 [-2.85, -0.26]	0.07 [-0.01, 0.20]	0.03 [-0.01, 0.13]
BiGRU	2.63 [2.33, 2.92]	1.74 [1.53, 2.01]	0.09 [0.03, 0.11]	0.41 [0.32, 0.51]	0.14 [0.11, 0.16]
MRU-AttnGRU	5.08 [4.30, 5.70]	4.41 [3.54, 5.01]	-2.22 [-3.19, -1.47]	0.03 [-0.08, 0.10]	0.01 [-0.02, 0.04]

B3. Statistical Comparisons between the models

Pairwise statistical comparisons were performed between the best-performing model and all other architectural families for each GRF axis. The Wilcoxon signed-rank test and paired t-test were applied to per-subject RMSE (%BW) values, with Cohen's d as the effect size measure.

Table S3-7 Pairwise statistical comparisons of DS-Conformer versus all other models based on RMSE (%BW) for vertical GRF. A positive mean difference indicates that DS-Conformer outperforms the comparator. Effect sizes: small (d < 0.2), medium (0.2-0.8), large (d > 0.8).

Comparator Model	Mean Diff (%BW)	Cohen's d	p (Wilcoxon)	p (t-test)
ConformerUNet	0.39	0.18	0.042	0.150
TCN	1.54	0.63	1.9×10^{-7}	3.3×10^{-6}
MultiResUNet	2.95	0.57	9.0×10^{-10}	2.1×10^{-5}
BiGRU	3.12	0.67	3.7×10^{-9}	9.8×10^{-7}
Hybrid-Concat	3.48	0.80	5.9×10^{-10}	1.5×10^{-8}
Hybrid-Gated	3.92	0.40	2.3×10^{-8}	0.002
DS-Transformer	5.40	1.41	2.2×10^{-11}	4.9×10^{-17}
ResCorr (frozen)	7.81	0.87	3.5×10^{-12}	1.6×10^{-9}
MRU-AttnGRU	8.35	0.72	5.5×10^{-10}	2.4×10^{-7}
ResCorr (joint)	12.21	0.65	2.4×10^{-10}	1.9×10^{-6}
BiLSTM	23.10	2.06	2.4×10^{-12}	6.6×10^{-25}
ShallowCNN	25.87	2.69	2.4×10^{-12}	3.3×10^{-31}

Table S3-8 Pairwise statistical comparisons of ResCorr (joint) versus all other models based on RMSE (%BW) for anteroposterior GRF. A positive mean difference indicates that ResCorr (joint) outperforms the comparator. Effect sizes: small (d < 0.2), medium (0.2-0.8), large (d > 0.8).

Comparator Model	Mean Diff (%BW)	Cohen's d	p (Wilcoxon)	p (t-test)
DS-Conformer	0.08	0.13	0.043	0.309
BiGRU	0.74	0.90	2.9×10^{-8}	6.8×10^{-10}
DS-Transformer	0.96	1.70	5.5×10^{-12}	9.4×10^{-21}
Hybrid-Gated	0.97	0.48	2.2×10^{-4}	2.3×10^{-4}
Hybrid-Concat	0.98	0.64	2.6×10^{-4}	2.9×10^{-6}
TCN	1.02	0.24	1.1×10^{-4}	0.057
ConformerUNet	1.25	0.67	5.1×10^{-8}	1.0×10^{-6}
ResCorr (frozen)	1.43	1.32	1.4×10^{-11}	8.1×10^{-16}
MultiResUNet	2.28	0.89	1.7×10^{-10}	1.0×10^{-9}
BiLSTM	2.85	2.65	3.8×10^{-12}	8.2×10^{-31}
ShallowCNN	3.32	3.36	2.4×10^{-12}	8.3×10^{-37}
MRU-AttnGRU	5.34	3.21	2.4×10^{-12}	1.4×10^{-35}

Table S3-9 Pairwise statistical comparisons of ResCorr (joint) versus all other models based on RMSE (%BW) for mediolateral GRF. A positive mean difference indicates that ResCorr (joint) outperforms the comparator. Effect sizes: small (d < 0.2), medium (0.2-0.8), large (d > 0.8).

Comparator Model	Mean Diff (%BW)	Cohen's d	p (Wilcoxon)	p (t-test)
DS-Conformer	0.06	0.05	2.1×10^{-4}	0.710
DS-Transformer	0.57	0.43	3.1×10^{-9}	0.001
TCN	0.78	0.22	1.1×10^{-6}	0.087
Hybrid-Gated	1.13	0.64	1.0×10^{-10}	2.8×10^{-6}
ConformerUNet	1.29	0.84	3.8×10^{-10}	4.1×10^{-9}
Hybrid-Concat	1.31	0.81	2.2×10^{-10}	1.2×10^{-8}
ResCorr (frozen)	1.38	0.98	6.0×10^{-11}	4.2×10^{-11}
BiLSTM	1.51	1.05	4.8×10^{-11}	5.7×10^{-12}
ShallowCNN	1.53	1.05	5.2×10^{-11}	4.3×10^{-12}
BiGRU	1.60	1.05	5.7×10^{-11}	4.9×10^{-12}
MultiResUNet	3.30	1.23	7.0×10^{-12}	1.6×10^{-14}
MRU-AttnGRU	3.88	2.30	4.0×10^{-11}	2.2×10^{-27}

For vertical GRF, DS-Conformer significantly outperformed most architectures, with large effect sizes against Hybrid-Concat, DS-Transformer, ResCorr (frozen), ShallowCNN, BiLSTM, and medium effects against TCN, MultiResUNet, BiGRU, Hybrid-Gated, MRU-AttnGRU, and ResCorr (joint). Small but significant effects were observed against ConformerUNet.

For AP GRF, ResCorr (joint) significantly outperformed most architectures, with large effect sizes against MultiResUNet, BiGRU, DS-Transformer, MRU-AttnGRU, ResCorr (frozen), ShallowCNN, BiLSTM, and medium effects against ConformerUNet, TCN, Hybrid-Concat, and Hybrid-Gated. Small but significant effects were observed against DS-Conformer.

For ML GRF, ResCorr (joint) significantly outperformed most architectures, with large effect sizes against ConformerUNet, MultiResUNet, BiGRU, Hybrid-Concat, MRU-AttnGRU, ResCorr (frozen), ShallowCNN, BiLSTM, and medium effects against TCN, Hybrid-Gated, DS-Transformer. Small but significant effects were observed against DS-Conformer.

B4. Biomechanical Event Metrics

In addition to continuous waveform metrics, discrete biomechanical event metrics were extracted to evaluate clinically relevant features of the predicted GRF waveforms. Peak force error measured the absolute difference between the maximum predicted and measured GRF. Time-to-peak error captured the timing accuracy of the force onset. Impulse error reflected cumulative force accuracy. The loading rate error quantified the accuracy of force-development dynamics. Onset error measured the temporal offset of predicted force initiation.

Table S3-10 Biomechanical event metrics for the best model from each architectural family (vertical GRF). Values are mean ± SD across 65 subjects.

Model	Peak Force Error (BW)	Time-to-Peak Error (s)	Impulse Error (BW·s)	Loading Rate Error (BW/s)	Onset Error (s)
DS-Conformer	0.030 ± 0.020	0.856 ± 0.763	0.304 ± 0.289	0.197 ± 0.201	-0.003 ± 0.113
ConformerUNet	0.038 ± 0.023	1.208 ± 1.044	0.278 ± 0.347	0.216 ± 0.182	-0.049 ± 0.246
TCN	0.056 ± 0.190	1.222 ± 0.998	0.381 ± 0.264	0.269 ± 0.208	-0.078 ± 0.518
MultiResUNet	0.055 ± 0.045	1.451 ± 1.196	0.471 ± 0.615	0.412 ± 0.423	-0.302 ± 1.637
BiGRU	0.057 ± 0.033	1.348 ± 1.102	0.440 ± 0.294	0.494 ± 0.375	-0.081 ± 0.355
Hybrid-Concat	0.072 ± 0.046	1.255 ± 0.915	0.533 ± 0.486	0.340 ± 0.320	-0.241 ± 1.015
Hybrid-Gated	0.460 ± 2.723	1.369 ± 1.173	0.420 ± 0.469	3.125 ± 20.167	-0.053 ± 1.451
DS-Transformer	0.049 ± 0.028	1.729 ± 1.181	0.397 ± 0.253	1.612 ± 1.688	-0.021 ± 0.162
ResCorr (frozen)	0.051 ± 0.031	1.855 ± 1.314	0.774 ± 1.085	0.477 ± 0.336	-0.413 ± 1.007
MRU-AttnGRU	0.174 ± 0.260	1.681 ± 1.171	1.209 ± 1.567	3.266 ± 12.902	-2.822 ± 4.494
ResCorr (joint)	0.232 ± 0.419	1.580 ± 1.248	1.434 ± 2.168	2.108 ± 12.054	-2.412 ± 4.556
BiLSTM	0.219 ± 0.145	2.433 ± 1.464	1.824 ± 2.139	1.763 ± 1.571	-0.868 ± 1.947
ShallowCNN	0.078 ± 0.042	2.463 ± 1.336	1.651 ± 2.101	1.957 ± 1.546	-2.787 ± 2.862

Table S3-10 presents discrete biomechanical event metrics for the best model from each architectural family, evaluating clinically relevant features of the predicted vertical GRF waveforms. DS-Conformer achieved the lowest peak-force error (0.030 BW), indicating superior reproduction of peak-loading dynamics during the STW transition. DS-Conformer also achieved the lowest loading-rate error (0.197 BW/s), indicating accurate reproduction of the slope dynamics during the rapid force-onset phase. TCN exhibited unusually high variance in peak force error (SD 0.190), likely driven by occasional convergence failures in individual folds.

B5. Robustness Experiments

To evaluate model resilience to realistic sensor perturbations encountered in wearable deployments, a robustness evaluation suite was applied to all models. Three categories of perturbation were systematically applied to the held-out test data.

Additive Gaussian noise: Independent zero-mean Gaussian noise was added to each IMU channel at two severity levels ($\sigma = 0.01$ and $\sigma = 0.02$, relative to the normalized signal amplitude). This perturbation simulates electronic noise, vibration artifacts, and thermal drift.

Axis misalignment (sensor rotation): A random three-dimensional rotation was applied to the triaxial accelerometer channels at two angular magnitudes (10° and 20°), simulating imprecise sensor placement or strap loosening during wear.

Signal dropout (packet loss): Contiguous segments of the IMU signal were replaced with zeros at two severity levels (5% and 10% of the total signal length) to simulate data packet loss during wireless transmission.

Together with the clean (unperturbed) baseline, these seven conditions (clean, noise $\sigma = 0.01$, noise $\sigma = 0.02$, rotation 10°, rotation 20°, dropout 5%, dropout 10%) provide a systematic characterization of each architecture's sensitivity to the three principal degradation modes encountered in wearable sensor deployments.

Table S3-11 to Table S3-13 present the robustness analysis results for vertical, AP, and ML GRF prediction. Each table reports RMSE (%BW) under clean and perturbed conditions, with Δ indicating the change from the clean baseline, averaged across the five cross-validation folds.

Table S3-11 Robustness of vertical GRF prediction under sensor perturbations. RMSE (%BW) under clean and perturbed conditions; Δ = change from clean baseline, averaged across five folds.

Model	Clean	Noise 0.01 (Δ)	Noise 0.02 (Δ)	Rot 10° (Δ)	Rot 20° (Δ)	Drop 5% (Δ)	Drop 10% (Δ)
DS-Conformer	4.05	4.05 (+0.00)	4.05 (+0.01)	4.35 (+0.30)	5.22 (+1.17)	7.58 (+3.54)	9.40 (+5.36)
ConformerUNet	4.44	4.50 (+0.06)	4.67 (+0.23)	4.66 (+0.22)	5.49 (+1.05)	10.80 (+6.36)	14.55 (+10.11)
TCN	5.59	5.60 (+0.01)	5.63 (+0.04)	5.70 (+0.11)	6.25 (+0.66)	10.85 (+5.26)	15.00 (+9.41)
MultiResUNet	7.00	8.09 (+1.09)	14.43 (+7.44)	8.45 (+1.46)	12.84 (+5.84)	37.36 (+30.36)	43.01 (+36.01)
BiGRU	7.16	7.31 (+0.15)	7.31 (+0.15)	7.54 (+0.37)	8.96 (+1.79)	24.07 (+16.91)	27.41 (+20.25)
Hybrid-Concat	7.53	7.44 (-0.09)	7.57 (+0.04)	8.01 (+0.48)	10.48 (+2.95)	38.47 (+30.94)	43.02 (+35.50)
Hybrid-Gated	7.96	7.84 (-0.12)	8.61 (+0.65)	9.07 (+1.10)	10.80 (+2.84)	43.17 (+35.21)	48.40 (+40.44)
DS-Transformer	9.45	9.46 (+0.01)	9.49 (+0.04)	9.62 (+0.17)	10.30 (+0.85)	13.64 (+4.19)	16.24 (+6.79)
ResCorr (frozen)	11.86	11.90 (+0.05)	11.99 (+0.13)	14.80 (+2.94)	20.95 (+9.10)	23.76 (+11.90)	31.03 (+19.17)
MRU-AttnGRU	12.39	12.51 (+0.12)	13.85 (+1.46)	12.91 (+0.52)	14.14 (+1.75)	29.52 (+17.13)	38.81 (+26.42)
ResCorr (joint)	16.26	15.48 (-0.77)	15.40 (-0.86)	17.00 (+0.75)	20.30 (+4.04)	43.94 (+27.68)	50.52 (+34.27)
BiLSTM	27.15	27.25 (+0.10)	27.25 (+0.10)	27.78 (+0.64)	31.54 (+4.40)	27.44 (+0.30)	27.62 (+0.47)
ShallowCNN	29.92	30.15 (+0.23)	31.35 (+1.43)	32.82 (+2.90)	37.31 (+7.39)	32.66 (+2.74)	35.10 (+5.18)

Table S3-12 Robustness of anteroposterior GRF prediction under sensor perturbations. RMSE (%BW) under clean and perturbed conditions; Δ = change from clean baseline, averaged across five folds.

Model	Clean	Noise 0.01 (Δ)	Noise 0.02 (Δ)	Rot 10° (Δ)	Rot 20° (Δ)	Drop 5% (Δ)	Drop 10% (Δ)
ResCorr (joint)	1.68	1.77 (+0.09)	2.04 (+0.36)	2.03 (+0.35)	2.45 (+0.77)	9.53 (+7.85)	12.96 (+11.28)
DS-Conformer	1.75	1.75 (0.00)	1.75 (0.00)	1.87 (+0.11)	2.12 (+0.37)	2.81 (+1.05)	3.48 (+1.73)
BiGRU	2.41	2.41 (0.00)	2.41 (0.00)	2.48 (+0.06)	2.67 (+0.25)	5.73 (+3.32)	7.37 (+4.95)
DS-Transformer	2.64	2.64 (+0.01)	2.66 (+0.02)	2.73 (+0.09)	3.01 (+0.37)	3.71 (+1.07)	4.48 (+1.85)
Hybrid-Gated	2.65	2.75 (+0.11)	3.03 (+0.38)	2.88 (+0.24)	3.17 (+0.53)	4.87 (+2.22)	5.82 (+3.17)
Hybrid-Concat	2.66	2.79 (+0.13)	3.21 (+0.55)	2.87 (+0.21)	3.38 (+0.72)	6.29 (+3.63)	7.61 (+4.95)
TCN	2.70	2.69 (0.00)	2.69 (0.00)	2.67 (-0.02)	2.67 (-0.03)	3.27 (+0.57)	3.75 (+1.05)
ConformerUNet	2.92	3.00 (+0.08)	3.22 (+0.29)	2.97 (+0.05)	3.11 (+0.18)	5.07 (+2.15)	6.34 (+3.41)
ResCorr (frozen)	3.10	3.11 (0.00)	3.12 (+0.01)	3.14 (+0.04)	3.29 (+0.18)	7.31 (+4.21)	10.61 (+7.51)
MultiResUNet	3.95	4.01 (+0.06)	4.17 (+0.22)	4.12 (+0.17)	4.40 (+0.45)	5.93 (+1.98)	6.90 (+2.95)
BiLSTM	4.53	4.53 (0.00)	4.53 (0.00)	4.60 (+0.07)	4.94 (+0.41)	4.75 (+0.22)	4.99 (+0.46)
ShallowCNN	5.00	5.00 (0.00)	5.01 (+0.01)	5.13 (+0.14)	5.50 (+0.50)	5.29 (+0.30)	5.55 (+0.55)
MRU-AttnGRU	7.02	7.02 (+0.01)	7.03 (+0.02)	7.01 (-0.01)	6.98 (-0.04)	7.02 (0.00)	7.01 (-0.01)

Table S3-13 Robustness of mediolateral GRF prediction under sensor perturbations. RMSE (%BW) under clean and perturbed conditions; Δ = change from clean baseline, averaged across five folds.

Model	**Clean**	**Noise 0.01 (Δ)**	**Noise 0.02 (Δ)**	**Rot 10° (Δ)**	**Rot 20° (Δ)**	**Drop 5% (Δ)**	**Drop 10% (Δ)**
ResCorr (joint)	1.03	1.08 (+0.06)	1.34 (+0.31)	1.28 (+0.26)	1.73 (+0.70)	7.42 (+6.39)	10.36 (+9.33)
DS-Conformer	1.09	1.09 (0.00)	1.09 (0.00)	1.15 (+0.06)	1.38 (+0.29)	1.21 (+0.12)	1.30 (+0.21)
DS-Transformer	1.59	1.59 (0.00)	1.59 (0.00)	1.66 (+0.07)	1.81 (+0.22)	1.86 (+0.27)	2.01 (+0.42)
TCN	1.81	1.81 (0.00)	1.81 (0.00)	1.83 (+0.02)	1.98 (+0.17)	2.23 (+0.42)	2.58 (+0.77)
Hybrid-Gated	2.15	2.28 (+0.13)	2.52 (+0.37)	2.31 (+0.16)	2.74 (+0.58)	7.50 (+5.35)	9.79 (+7.64)
ConformerUNet	2.31	2.38 (+0.07)	2.57 (+0.25)	2.36 (+0.05)	2.54 (+0.23)	8.27 (+5.95)	11.30 (+8.99)
Hybrid-Concat	2.34	2.34 (0.00)	2.35 (+0.01)	2.37 (+0.03)	2.47 (+0.13)	2.47 (+0.13)	2.55 (+0.21)
ResCorr (frozen)	2.40	2.41 (+0.00)	2.43 (+0.03)	2.46 (+0.06)	2.63 (+0.23)	4.53 (+2.13)	6.13 (+3.73)
BiLSTM	2.53	2.53 (+0.00)	2.53 (+0.00)	2.71 (+0.18)	3.07 (+0.54)	2.54 (+0.01)	2.56 (+0.03)
ShallowCNN	2.55	2.56 (+0.00)	2.56 (+0.00)	2.64 (+0.09)	2.83 (+0.27)	2.61 (+0.06)	2.68 (+0.12)
BiGRU	2.63	2.63 (+0.00)	2.63 (+0.00)	2.64 (+0.01)	2.69 (+0.07)	2.64 (+0.01)	2.65 (+0.03)
MultiResUNet	4.33	4.34 (+0.01)	4.36 (+0.03)	4.40 (+0.08)	4.58 (+0.25)	4.83 (+0.50)	5.13 (+0.80)
MRU-AttnGRU	4.91	4.90 (-0.01)	4.89 (-0.02)	4.94 (+0.03)	4.96 (+0.05)	4.93 (+0.02)	4.94 (+0.03)

Three distinct robustness profiles emerged across all three force axes.

Noise robustness: Most models were invariant to σ = 0.01 noise (Δ < 0.1 %BW), though attention-based and hybrid models showed higher sensitivity. At σ = 0.02, DS-Conformer and TCN models remained stable across all axes, showing near-zero degradation. MultiResUNet was the most noise-sensitive on the vertical axis (+7.44 %BW at σ = 0.02).

Rotation robustness: At 10° rotation, all models experienced modest degradation across all three axes. At 20°, a clear separation appeared. DS-Conformer was consistently among the least affected models across all axes (vertical: +1.17 %BW at 20°). ResCorr (frozen) showed the largest rotation sensitivity (+9.10 %BW).

Signal dropout robustness: This was the most discriminating perturbation across all three axes. At 10% dropout on the vertical axis, models were separated into distinct tiers. DS-Conformer was the most robust among competitive architectures (+5.36 %BW), while Hybrid-Gated showed the largest dropout sensitivity (+40.44 %BW).

ShallowCNN and BiLSTM showed paradoxically high dropout robustness across all axes because their baseline predictions were already poor. The perturbation could not substantially degrade an already inaccurate prediction.

B6. Model size and inference costs

Inference was benchmarked on the New Zealand eScience Infrastructure (NeSI) HPC system accessed through Mahuika, using one NVIDIA L4 GPU (23,034 MiB reported device memory) and a host with two AMD EPYC 9634 84-core processors. The operating system reported 1510.93 GiB host memory; the job was allocated 32 GiB RAM and four logical CPUs. The software environment comprised Python 3.10.12 and TensorFlow/Keras 2.14.0 (CUDA 12.3 build; NVIDIA driver 580.95.05).

One trained checkpoint from the first cross-validation fold represented each distinct selected model variant. Timings used model.predict() with batch sizes of one and five, with 10 warm-up calls followed by 50 timed calls per condition. Mixed precision and TF32 were disabled. Timing included prediction-API overhead and completed host–device transfers, but excluded input generation, checkpoint loading and warm-up. Each model was tested in a separate process, using two TensorFlow intra-operation threads and one inter-operation thread.

Table S3-14 Parameter counts and inference costs of all the used models..

Model	Total parameters	Trainable parameters	Single-trial latency (ms)	Batched inference (ms/trial)	Peak GPU memory (MiB)
DS-Conformer	556,289	555,777	48.47 [43.49, 54.38]	11.81 [11.47, 12.25]	15.4
ConformerUNet	948,769	948,257	48.20 [44.47, 53.38]	16.70 [16.23, 17.47]	54.7
TCN	413,761	413,761	40.83 [40.37, 41.50]	10.32 [9.79, 10.71]	6.4
DS-Transformer	363,265	363,265	45.89 [42.80, 49.29]	8.47 [7.73, 9.38]	5.7
MultiResUNet	16,451,847	16,400,567	51.99 [50.45, 57.96]	13.32 [12.80, 13.56]	90.2
MRU-AttnGRU	20,778,019	20,722,751	194.57 [190.25, 199.85]	38.55 [37.92, 39.13]	123.2
Hybrid-Concat	16,509,577	16,458,297	184.17 [183.09, 188.03]	36.54 [35.83, 37.18]	98.3
Hybrid-Gated	16,509,515	16,458,235	181.73 [178.45, 184.28]	36.67 [36.11, 37.10]	98.4
ResCorr (joint)	16,514,632	16,463,352	57.98 [56.85, 59.85]	14.55 [14.16, 15.18]	90.2
ResCorr (frozen)	16,514,632	62,785	51.02 [50.30, 55.64]	14.67 [14.35, 14.95]	90.2
BiGRU	57,665	57,665	161.50 [159.66, 164.55]	32.68 [32.34, 33.13]	23.9
BiLSTM	2,352,257	2,352,257	108.35 [106.58, 157.77]	28.59 [28.24, 28.98]	121.0

Model	Total parameters	Trainable parameters	Single-trial latency (ms)	Batched inference (ms/trial)	Peak GPU memory (MiB)
ShallowCNN	20,513	20,513	43.43 [41.85, 45.00]	8.80 [8.64, 8.97]	3.0

Note. Timings are median [Q1, Q3] across repeated calls. Single-trial latency uses batch size one; batched inference is the duration of a five-trial batch divided by five and represents throughput. GPU memory is the TensorFlow-tracked peak during the batch-size-one timing block, measured after resetting the peak statistic following warm-up; it includes resident model weights but excludes untracked driver/context memory. Total parameter counts include non-trainable model state and exclude optimizer state. Separate rows identify component-specific variants.

Appendix C. Supplementary Material to Chapter 4

C1. Median and Interquartile Range Summary

Table S4-1 Leave-One-Out Cross-Validation Results for Ground Reaction Force, Moment, and Center of Pressure Prediction Reported as Median [Q1, Q3]

Component	RMSE (%BW)	NRMSE (%)	MAE (%BW)	R^2	Pearson r	CCC
GRF (V)	4.77 [3.79, 6.35]	3.39 [2.73, 4.70]	3.65 [2.93, 5.11]	0.985 [0.975, 0.991]	0.996 [0.992, 0.997]	0.993 [0.987, 0.995]
GRF (AP)	2.06 [1.72, 2.53]	5.38 [4.27, 6.13]	1.53 [1.28, 1.88]	0.879 [0.808, 0.907]	0.952 [0.922, 0.966]	0.938 [0.902, 0.955]
GRF (ML)	1.14 [1.00, 1.31]	7.22 [6.05, 8.98]	0.78 [0.70, 0.91]	0.801 [0.703, 0.870]	0.918 [0.894, 0.948]	0.901 [0.857, 0.937]
GRM (Sagittal)	4.81 [4.09, 6.36]	2.71 [2.36, 3.25]	3.14 [2.58, 3.97]	0.989 [0.980, 0.991]	0.996 [0.994, 0.997]	0.995 [0.991, 0.996]
GRM (Frontal)	2.12 [1.78, 2.43]	5.86 [4.92, 6.90]	1.50 [1.20, 1.81]	0.941 [0.914, 0.960]	0.981 [0.971, 0.986]	0.971 [0.962, 0.980]
GRM (Transverse)	1.18 [0.99, 1.47]	5.17 [4.35, 7.14]	0.71 [0.61, 0.83]	0.842 [0.665, 0.891]	0.942 [0.895, 0.960]	0.915 [0.853, 0.943]
COP (AP)	0.065 [0.052, 0.081]	3.96 [3.37, 4.98]	0.043 [0.038, 0.061]	0.975 [0.964, 0.985]	0.994 [0.990, 0.997]	0.988 [0.982, 0.992]
COP (ML)	0.030 [0.025, 0.036]	11.38 [8.92, 13.93]	0.021 [0.018, 0.026]	0.522 [0.298, 0.678]	0.829 [0.721, 0.888]	0.771 [0.680, 0.837]

Note. Values are reported as median [Q1, Q3] across 65 held-out subjects from leave-one-out cross-validation. RMSE and MAE are reported in percent bodyweight (%BW) for GRF, in meters (m) for CoP, and in %BW·m for GRM components. NRMSE = RMSE normalized by the ground-truth signal range. R^2 = coefficient of determination; r = Pearson correlation coefficient; CCC = Lin's concordance correlation coefficient

C2. Model size and inference costs

Inference was benchmarked on the New Zealand eScience Infrastructure (NeSI) HPC system accessed through Mahuika, using one NVIDIA L4 GPU (23,034 MiB reported device memory) and a host with two AMD EPYC 9634 84-core processors. The operating system reported 1510.93 GiB host memory; the job was allocated 32 GiB RAM and four logical CPUs. The software environment comprised Python 3.10.12 and TensorFlow/Keras 2.14.0 (CUDA 12.3 build; NVIDIA driver 580.95.05).

A trained checkpoint from the first cross-validation fold was used to benchmark one DS-Conformer component model. Timings used model.predict() with batch sizes of one and five, with 10 warm-up calls followed by 50 timed calls per condition. Mixed precision and TF32 were disabled. Timing included prediction-API overhead and completed host–device transfers, but excluded input generation, checkpoint loading and warm-up. The benchmark ran in a separate process, using two TensorFlow intra-operation threads and one inter-operation thread.

Table S4-2 Parameter counts and inference costs of the DS-Conformer model

Model	Total parameters	Trainable parameters	Single-trial latency (ms)	Batched inference (ms/trial)	Peak GPU memory (MiB)
DS-Conformer	1,742,785	1,741,633	111.61 [110.54, 115.53]	25.83 [25.54, 27.63]	33.2

Note. Timings are median [Q1, Q3] across repeated calls. Single-trial latency uses batch size one; batched inference is the duration of a five-trial batch divided by five and represents throughput. Peak GPU memory is TensorFlow-tracked memory during the batch-size-one timing block, including resident model weights but excluding untracked driver/context memory.

C3. Exploratory GRF estimation on external dataset

This supplementary analysis examines whether the GRF modelling framework can be extended to rising movements recorded in an external dataset. It complements Chapter 4 by evaluating anterior–posterior (AP), mediolateral (ML) and vertical (V) forces during sit-to-stand, using virtual accelerations reconstructed from optical motion capture. The analysis concerns GRF only; external prediction of ground reaction moments or center of pressure was not evaluated.

Data and analysis

Data were drawn from the Falisse and Uhlrich cohorts in AddBiomechanics (Falisse et al., 2017; Uhlrich et al., 2023; Werling et al., 2024). The retained sample comprised 129 rises from 18 participants: 91 rises from 12 development participants and 38 rises from six test participants. One development rise failed trunk-marker quality control. All repetitions from each participant remained in the same partition. These recordings supplied measured foot forces and optical markers, but no measured IMU input for this analysis.

Virtual pelvis or trunk accelerations were derived from marker trajectories in physical time, expressed in sensor coordinates with gravity retained. Evaluation covered 10–90% of the pelvis-height rise; inputs included up to 0.5 s of surrounding context. Measured foot forces were normalized to bodyweight and filtered using a fourth-order causal 6 Hz Butterworth filter. This analysis follows the archived acceleration-only implementation, whose source recordings were identified as anterior-trunk signals. These details differ from the sacrum placement and 10 Hz filtering described in Chapter 4; the present results therefore concern the archived implementation and its external adaptations.

Four participant-grouped development folds selected models by RMSE, with concordance correlation coefficient (CCC) as a tie breaker. The selected AP model was a temporal convolutional network trained from scratch on virtual pelvis input at 100 Hz. ML and V used adapted DS-Conformer models with 4096-point inputs from the virtual pelvis and trunk, respectively. Final predictions averaged three independently seeded fits. A development-set average waveform, calculated only from development participants, provided the same baseline prediction for every test participant.

Results and interpretation

Table S4-3 External GRF prediction compared with the development-set average waveform

Component	Model RMSE (%BW)	Pearson r	CCC	Development-set average waveform RMSE (%BW)
GRF (V)	5.54 ± 1.44	0.965 ± 0.035	0.941 ± 0.028	15.05 ± 7.70
GRF (AP)	1.87 ± 0.60	0.943 ± 0.021	0.854 ± 0.075	3.02 ± 1.32
GRF (ML)	1.19 ± 0.53	0.455 ± 0.408	0.198 ± 0.141	0.96 ± 0.36

Note. Values are mean ± standard deviation across six test participants. Metrics were calculated for each valid rise, averaged within participant, and then averaged equally across participants. RMSE is expressed as percent bodyweight (%BW); r = Pearson correlation coefficient; CCC = Lin's concordance correlation coefficient. This baseline uses development data only. All entries use the same causal 6 Hz force reference and the same 38 test rises.

Table S4-3 shows strong AP and V waveform tracking, with correlations of 0.943 and 0.965, respectively. Their RMSEs were 38.1% and 63.2% lower than the development-set average waveform baseline. ML estimation remained weak: despite its small absolute RMSE, correlation was variable and CCC was only 0.198. Its RMSE exceeded this baseline by 24.5%, indicating that the model did not improve lateral force accuracy over a generic rising waveform. Figure S4-1 illustrates the AP and V agreement and the residual ML offset. These findings support further investigation of AP and V estimation after adaptation, while ML force recovery remains unresolved.

Limitations

The six test participants had been examined in an earlier experiment and informed the follow-up questions, so these findings are exploratory despite development-only model selection. Virtual accelerations do not establish equivalence to physical IMUs. Different movements, sensor representations, preprocessing and fitted models also prevent a direct accuracy comparison with Chapter 4. In particular, the AP model was trained from scratch rather than transferred from the original checkpoint. Marker-defined windows and access to future context make this an offline analysis. Independent participants with measured IMU and force data are needed before claiming wearable generalization; GRM, CoP and clinical applicability remain untested externally.

Measured and predicted waveforms

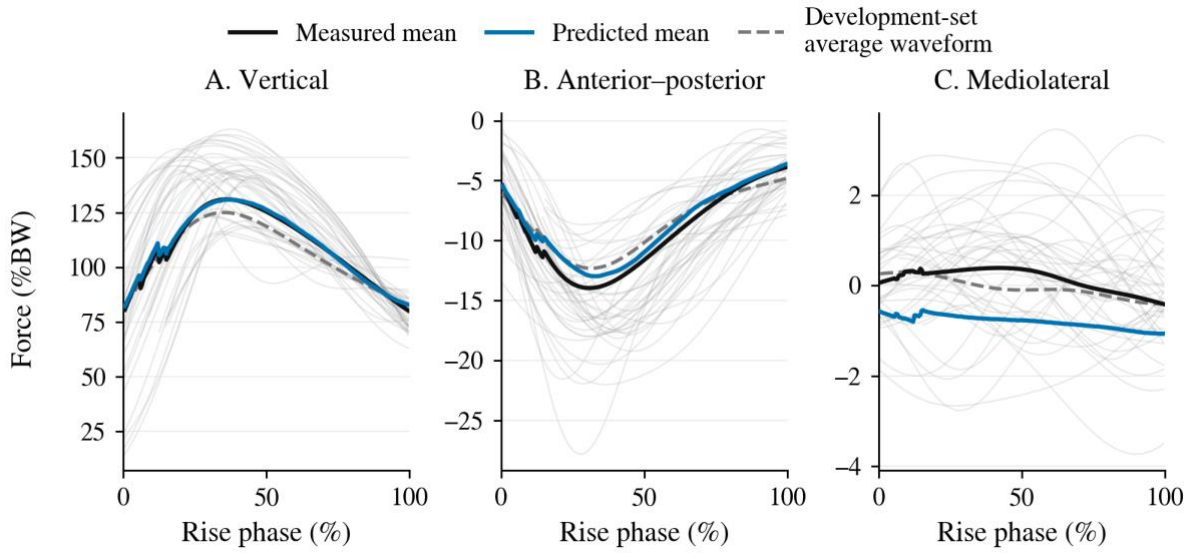


Figure S4-1 Measured and predicted external GRF waveforms. Black and blue curves show participant-balanced measured and predicted means; dashed grey curves show the development-set average waveform. Faint grey lines represent individual measured rises. At the start of the rise, the curves represent 30 rises from three participants. All 38 rises from six participants contribute between 16% and 99.5% of the rise phase. The exact endpoints (0% and 100%) are omitted because rounding can select samples outside the valid scoring interval. No additional smoothing was applied. These averages illustrate waveform patterns; prediction accuracy in Table S4-3 was calculated for individual rises using only valid samples.

Original and external dataset performance

Table S4-4 compares original and external GRF results descriptively, with prediction error and waveform correlation reported under their respective evaluation protocols.

Table S4-4 Original and external GRF prediction performance for descriptive comparison

Component	Original RMSE (%BW)	External RMSE (%BW)	Original Pearson r	External Pearson r
GRF (V)	6.55 ± 4.76	5.54 ± 1.44	0.982 ± 0.033	0.965 ± 0.035
GRF (AP)	2.24 ± 0.76	1.87 ± 0.60	0.927 ± 0.074	0.943 ± 0.021
GRF (ML)	1.22 ± 0.34	1.19 ± 0.53	0.893 ± 0.097	0.455 ± 0.408

Note. Values are mean ± SD across participants. Original results (Table 4-1) cover 325 complete sit-to-walk trials from 65 participants under leave-one-participant-out cross-validation. External results (Table S4-3) cover 38 sit-to-stand rises from six participants, scored over 10–90% of the pelvis-height rise. Trial/rise metrics were averaged within participants and then across participants with equal weighting. RMSE = root mean square error (%

bodyweight); r = Pearson's correlation coefficient; AP = anteroposterior; ML = mediolateral; V = vertical.

The lower external vertical and AP RMSE may reflect differences in movement and evaluation interval, training or adaptation on external development data, and averaging three fitted models. These factors were not isolated, so the numerical reductions do not establish improved generalization.

ML correlation was weak and highly variable ($r = 0.455 \pm 0.408$), indicating inconsistent waveform recovery across participants. Its small RMSE (1.19% BW) must be interpreted relative to the small-amplitude ML signal: CCC was only 0.198, and RMSE exceeded the average-waveform baseline (0.96% BW). Low absolute error therefore did not indicate reliable lateral force prediction.

Consistent with the discussion in Chapter 5, this pattern highlights a potential limitation of the thesis's single midline-mounted IMU approach: central motion may represent forward and upward dynamics more clearly than left–right load redistribution. Reliable external ML estimation remains unestablished, motivating evaluation of additional sensor locations. However, the external inputs were virtual accelerations, so this analysis cannot isolate sensor placement as the cause of the failure.

References

Falisse, A., Van Rossom, S., Jonkers, I., & De Groote, F. (2017). EMG-driven optimal estimation of subject-specific Hill model muscle-tendon parameters of the knee joint actuators. *IEEE Transactions on Biomedical Engineering, 64*(9), 2253–2262. https://doi.org/10.1109/TBME.2016.2630009

Uhlrich, S. D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A. S., Hicks, J. L., & Delp, S. L. (2023). OpenCap: Human movement dynamics from smartphone videos. *PLOS Computational Biology, 19*(10), e1011462. https://doi.org/10.1371/journal.pcbi.1011462

Werling, K., Kaneda, J., Tan, A., Agarwal, R., Skov, S., Van Wouwe, T., Uhlrich, S., Bianco, N., Ong, C., Falisse, A., Sapkota, S., Chandra, A., Carter, J., Preatoni, E., Fregly, B., Hicks, J., Delp, S., & Liu, C. K. (2024). *AddBiomechanics dataset: Capturing the physics of human motion at scale*. arXiv. https://doi.org/10.48550/arXiv.2406.18537