

Article

When Mixture-of-Experts Fails: An Empirical Study of Expert Routing Limitations in Long-Tailed Gait Regression

Hyeran Koo

SW-Centered University Project Group, Hanyang University, ERICA Campus, Ansan 15588, Republic of Korea; jasminekoo@hanyang.ac.kr

Featured Application

This study provides a diagnostic framework for evaluating whether Mixture-of-Experts routing is appropriate before deploying wearable plantar-pressure-based gait regression models in subject-independent settings.

Abstract

This study presents a methodological evaluation of the conditions required for successful Mixture-of-Experts (MoE) deployment in continuous biomechanical regression, rather than proposing a new predictive architecture. Using plantar-pressure-based stride length estimation as a case study, we examined whether MoE routing provides stable performance gains over a single convolutional neural network (CNN) under a long-tailed target distribution. Plantar pressure data from nine force-sensitive resistor sensors were collected from 22 male participants, including five older adults and 17 young adults, and represented as 36-channel time-series inputs comprising raw pressure, first-order differences, second-order differences, and cumulative integrals. Under leave-one-subject-out evaluation with five repeated runs, a CNN baseline was compared with four MoE variants using statistical gating, BiLSTM encoder gating, fully connected autoencoder gating, and zone-specific expert architectures. The CNN achieved the most stable overall performance (RMSE: 7.54 ± 0.14 cm). MoE-AEG significantly improved short-stride prediction but significantly deteriorated long-stride prediction. In MoE-FCAE, expert-only analyses confirmed zone-specific specialization, whereas gating weights remained predominantly allocated to the middle-stride expert across all zones. These findings show that expert specialization alone is insufficient for improved MoE regression; input separability, expert specialization, and gating alignment should be verified separately, together with subject-wise robustness under subject-independent evaluation.



Academic Editor: Juan Francisco De Paz Santana

Received: 17 June 2026

Revised: 21 July 2026

Accepted: 24 July 2026

Published: 27 July 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: time-series data; Mixture-of-Experts; long-tailed regression; gating alignment; expert specialization; convolutional neural network; plantar pressure; stride length

1. Introduction

Gait analysis plays a central role in clinical assessment, rehabilitation monitoring, and the evaluation of mobility decline associated with aging and neurological disorders. Among the various spatiotemporal gait parameters, stride length is a key variable reflecting gait efficiency, lower-extremity motor control, and compensatory locomotor strategies [1,2]. Stride length is also included among normative spatiotemporal gait parameters used to assess and interpret gait dysfunction in older adults [3]. Previous evidence further

suggests that stride length is clinically relevant for predicting adverse outcomes in older adults [4]. Accordingly, technologies capable of continuously estimating stride length in both laboratory and free-living environments are of substantial importance for clinical decision-making and longitudinal monitoring.

While optical motion capture systems and instrumented walkways provide highly accurate stride length measurements, they require specialized laboratory settings, expensive equipment, and constrained measurement environments [5]. These limitations have driven the adoption of wearable sensor-based gait analysis, including inertial measurement units (IMUs) and instrumented insoles [6,7]. Deep learning approaches have also been used to map stride-specific wearable sensor signals directly to stride length, supporting the feasibility of sensor-based stride-length estimation outside conventional laboratory systems [8]. Whereas IMUs directly capture the kinematics of body segments, plantar pressure insoles measure localized load distribution between the foot and the ground [9]. Consequently, plantar pressure signals reflect foot–ground interaction dynamics during gait but do not directly measure stride length as a whole-body translational distance. Stride length is determined by a combination of lower-extremity kinematics, walking speed, stance duration, push-off mechanics, center-of-pressure trajectory, and individual locomotor strategies, of which plantar pressure signals encode only a subset of biomechanical information indirectly [10,11]. Estimating continuous stride length values from plantar pressure signals therefore constitutes an indirect and nonlinear regression problem rather than a simple linear transformation.

Recent gait-analysis studies have increasingly applied deep learning and machine-learning methods to wearable-sensor and vision-based data for gait classification, clinical impairment assessment, and spatiotemporal gait-parameter estimation [12]. For example, Zadka et al. estimated step length from a single lower-back inertial sensor in a large cohort of young adults, older adults, and individuals with neurological disorders [7]. Although their model achieved a single-step RMSE of 6.08 cm, shorter steps tended to be overestimated and longer steps underestimated, indicating that regression-to-the-mean bias can persist even in relatively large and heterogeneous datasets. Deep-learning studies using plantar-pressure signals have also focused primarily on clinical event detection and classification rather than continuous spatiotemporal gait-parameter regression. Shalin et al. applied long short-term memory networks to pressure-sensing insole data for the detection and prediction of freezing of gait in individuals with Parkinson's disease [13]. More recently, Park et al. developed a temporal convolutional neural network using foot-pressure sensing insoles for freezing-of-gait detection [14]. These studies demonstrate that plantar-pressure time series contain clinically informative temporal gait patterns; however, they addressed discrete clinical events and did not examine continuous stride-length regression, region-specific expert specialization, or gating alignment. In contrast, the present study investigates whether plantar-pressure representations support functionally specialized experts and whether gating decisions appropriately transfer expert-level advantages to continuous stride-length prediction. However, previous studies have primarily focused on improving overall predictive accuracy or clinical discrimination.

Beyond gait-specific applications, recent long-sequence forecasting studies have investigated integrated time–frequency representation learning for continuous regression. CSA-WTConvLSTM was designed to jointly capture low- and high-frequency components, whereas ST-TFFM-Observer enhanced the modeling of temporal–frequency dependencies within a dynamic-system framework [15,16]. Although these models were developed for ionospheric total electron content forecasting rather than gait biomechanics, they demonstrate methodological strategies for extracting multi-scale coupled time–frequency features from long sequential signals. Such strategies may also be relevant to plantar-

pressure signals, which contain both gradual loading patterns and rapid pressure transitions; however, their effectiveness for subject-independent stride-length regression has not yet been established.

This challenge is further compounded when the target distribution of stride lengths is imbalanced. In the present dataset, stride lengths were concentrated around a central range, with relatively fewer observations at the short- and long-stride extremes. This setting corresponds to a deep imbalanced regression problem, in which continuous target values are unevenly represented across the target space [17]. In such long-tailed continuous target distributions, single-model regressors trained to minimize overall mean error are prone to exhibiting regression-to-the-mean bias, in which predictions are pulled toward the central value in the tail regions [17]. In a preliminary conference-level analysis preceding this study [18], a CNN achieved a mean error of 0.08 ± 4.8 cm overall, yet errors increased substantially in the short and long stride groups, with short strides overestimated (0.18 ± 8.0 cm) and long strides underestimated (-0.17 ± 8.3 cm).

Established approaches for mitigating imbalanced regression include resampling, data augmentation, and loss reweighting [17,19,20]. However, these strategies primarily adjust sample distribution or the relative contribution of training samples to the loss function, without directly modeling the possibility that the functional relationship between input and output may differ across stride length regions. In plantar-pressure-based stride estimation, the characteristic patterns of stance duration, loading trajectory, push-off, and center-of-pressure displacement may differ substantially across short, middle, and long strides. It is therefore not self-evident that a single global regression function is equally well-suited to all stride length ranges.

This consideration motivates the use of local modeling approaches that learn distinct prediction functions for different stride length regions. Mixture-of-Experts (MoE) is a representative framework for implementing such local modeling: multiple expert networks are each responsible for mapping a distinct subset of the input or target space, while a gating network modulates the contribution of each expert for a given input [21,22]. Under conditions in which the input data are meaningfully separable into distinct subregions, each expert learns a specialized mapping for its assigned region, and the gating network assigns high weight to the appropriate expert, MoE may offer advantages over a single global model.

However, the theoretical merits of MoE do not necessarily translate into improved regression performance in practice. Plantar-pressure-based stride prediction is not a categorically separable classification problem; rather, it is a continuous and overlapping biomechanical regression problem. For MoE to function effectively, it is insufficient merely to introduce multiple experts; the conditions of input separability, expert specialization, and gating alignment must be simultaneously satisfied.

Accordingly, the primary objective of this study was not to propose a new predictive architecture or simply identify the best-performing deep learning model, but to evaluate whether the conditions required for successful MoE deployment are satisfied in subject-independent plantar-pressure-based continuous stride length regression. Specifically, three structural conditions were examined separately: input separability, expert specialization, and gating alignment. Subject-wise robustness was additionally evaluated under LOSO validation to determine whether any observed performance advantage was consistently maintained across unseen participants.

2. Materials and Methods

2.1. Study Design

This study compared a single CNN model against MoE-based models for plantar-pressure-based stride length regression and examined whether MoE architectures provide stable performance gains under a long-tailed stride length distribution. Because MoE operates on the principle that a gating network determines per-input expert contributions, and that local experts are each responsible for a distinct data subset or input region [21,22], the study evaluated whether these structural premises hold in a continuous gait regression task from three perspectives. First, input separability refers to whether plantar pressure features or learned latent representations contain sufficient information to distinguish stride length zones (short, middle, and long strides). Second, expert specialization refers to whether each expert achieves relatively lower prediction error within its designated stride length zone. Third, gating alignment refers to whether the gating network assigns high weight to the appropriate expert for a given true stride length zone or the locally best-performing expert.

2.2. Dataset and Leave-One-Subject-Out Evaluation

The dataset comprised 22 male participants, including five older adults aged 65 years or above and 17 young adults. The older adults performed walking trials at a comfortable speed, whereas the young adults performed trials at slow, comfortable, and fast walking speeds. These different walking conditions were included to obtain a broad distribution of stride lengths and were not intended for independent comparisons of age or walking-speed effects. Plantar pressure data were recorded using a custom-fabricated insole equipped with nine force-sensitive resistor (FSR) sensors (Tekscan Inc., Norwood, MA, USA) [9], and the target variable for each trial was stride length measured using a VICON 460 motion capture system (Vicon Motion Systems Ltd., Oxford, UK).

Model generalization to unseen individuals was evaluated using leave-one-subject-out (LOSO) cross-validation, an established approach for assessing subject-independent performance in wearable gait analysis research. In each fold, one participant was held out as the test subject, and the remaining participants' data were used exclusively for training.

To account for stochastic variability inherent in deep learning optimization, all models were trained and evaluated five times under identical LOSO partitioning conditions. Overall performance was summarized as the mean $\pm$ standard deviation across the five repeated runs.

2.3. Signal Preprocessing and Feature Construction

Each stride-level plantar pressure sequence consisted of nine FSR sensor channels. A second-order Butterworth low-pass filter with a cutoff frequency of 5 Hz was applied to attenuate high-frequency noise in the raw FSR signals [23]. Zero-phase forward-backward filtering was employed to minimize phase distortion.

Four feature types were derived from each sensor channel: raw pressure, first-order differences, second-order differences, and cumulative integrals. First-order differences captured the rate of pressure change over time, while second-order differences represented the curvature of the pressure time series—i.e., the acceleration of pressure change. Cumulative integrals encoded the accumulated load over time [24]. These four feature types were concatenated across all nine sensor channels to yield a 36-channel time-series input.

To assess the sensitivity of the CNN baseline to the difference features, an additional ablation analysis compared the original 36-channel input with an 18-channel input retaining only the raw and cumulative integral features. All other model architecture, training, and LOSO evaluation settings were maintained.

To preserve temporal information associated with stride duration, cadence, and walking speed without complete phase normalization, sequences were not normalized to 100% of the gait cycle [25]. Instead, zero-padding was applied to standardize all sequences to a fixed length of 256 time steps.

2.4. Stride Length Zone Definitions

Stride length targets were stratified into three zones based on true stride length values. Zone A (short strides) was defined as stride lengths below 125 cm, Zone B (middle strides) as 125–135 cm, and Zone C (long strides) as stride lengths exceeding 135 cm. These boundaries were empirically selected to approximate the 25th and 75th percentiles of the stride length distribution in the present dataset (Figure 1). These thresholds were introduced solely as distribution-based analytical partitions for examining lower-tail, central, and upper-tail model behavior. They do not represent clinically validated stride-length cutoff values and should not be interpreted as diagnostic or normative gait standards. Consequently, the numerical boundaries may not be directly transferable to other demographic or clinical populations. To partially compensate for the sparsity of training samples in the tail zones, data augmentation was applied within the training set of each LOSO fold to Zone A and Zone C samples only. Augmentation procedures included jittering, magnitude scaling, and temporal warping [26,27]. Augmentation was not applied to the held-out test subject's data in any fold.

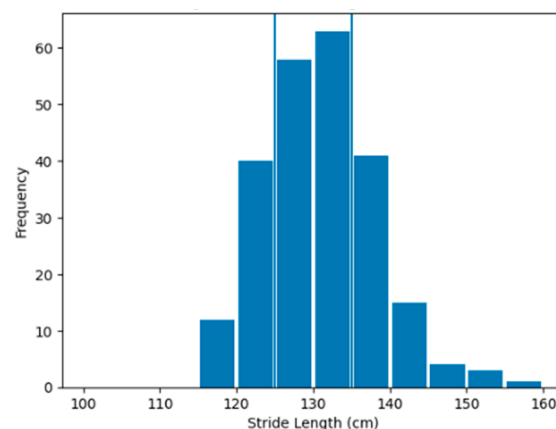


Figure 1. Histogram of Stride Length Distribution. The vertical lines indicate the boundaries used to define Zone A (<125 cm), Zone B (125–135 cm), and Zone C (>135 cm).

2.5. Baseline CNN Model

A one-dimensional CNN was selected as the reference model, as CNNs are capable of learning local temporal patterns from multichannel time-series data and have been established as appropriate baseline models in wearable-sensor-based gait analysis [8,28]. The CNN baseline served as the reference against which the additional structural complexity of MoE-based models could be assessed in terms of realized performance gains.

The CNN baseline accepted 36-channel stride-level time-series input and produced a single continuous stride length prediction. The architecture comprised two 1D convolutional layers with 128 and 256 filters and kernel sizes of 7 and 5, respectively, followed by batch normalization [29], ReLU activation, dropout (rate = 0.3) [30], global average pooling [31], and a fully connected regression output layer. Batch normalization was included to enhance training stability, dropout to mitigate overfitting, and global average pooling to produce a compact temporal feature representation while reducing the parameter count of the subsequent fully connected layer. The model was trained using the Adam optimizer with a learning rate of 0.001, a mini-batch size of 32, and a maximum of 50 epochs. Before

the final repeated experiments, five predefined CNN configurations were compared using LOSO evaluation. The candidate configurations varied the numbers of convolutional filters, kernel sizes, and the inclusion of dropout. The configuration comprising 128 and 256 filters, kernel sizes of 7 and 5, and a dropout rate of 0.3 achieved the lowest aggregate prediction error and was selected as the final CNN baseline. The optimizer-related settings were fixed across the candidate configurations. This procedure represented limited candidate-based architecture selection rather than exhaustive hyperparameter optimization or nested cross-validation.

2.6. Mixture-of-Experts Architecture

The MoE framework consisted of three expert regressors and one gating network. Experts A, B, and C were trained to specialize in Zones A, B, and C, respectively. Each expert independently produced a stride length prediction for the given input sequence. The gating network accepted either handcrafted features or a learned latent representation as input and produced softmax weights over the three experts.

The final MoE prediction was computed as the weighted sum of the three expert predictions:

$$\hat{y} = w_A \cdot \hat{y}_A + w_B \cdot \hat{y}_B + w_C \cdot \hat{y}_C$$

where $\hat{y}_A$, $\hat{y}_B$ and $\hat{y}_C$ denote the predictions of Experts A, B, and C, respectively, and w_A , w_B and w_C denote softmax weights output by the gating network. In this study, the MoE framework was not treated as a simple ensemble but rather as a structure for empirically examining whether expert specialization and gating alignment are formed in practice across stride length zones.

2.7. Model Variants

This study compared the CNN baseline against four MoE-based model variants.

The CNN baseline directly regressed the full input space using a single CNN regression function, as described above.

MoE-SG employed statistical gating. For each 36-channel input sequence, the mean, standard deviation, and skewness were computed along the temporal dimension and used as the gating network input. The expert architecture consisted of identical CNN experts. This model was designed to evaluate whether simple handcrafted statistical features are sufficient for stride zone discrimination and effective expert routing.

MoE-AEG employed a BiLSTM sequence encoder to extract a 32-dimensional latent feature vector from the 36-channel stride-level time-series input [32,33]. The encoder comprised a sequence input layer (36 channels), a BiLSTM layer (64 hidden units, output mode: last hidden state), and a fully connected projection layer (32 dimensions). The extracted latent vector was supplied to the same gating classifier architecture used in MoE-SG to produce softmax weights over Zones A, B, and C. Expert architecture was identical to MoE-SG. This model was designed to assess whether a learned sequence-level temporal latent representation provides more informative routing signals than handcrafted statistical features.

MoE-FCAE flattened the 36×256 time-series input into a 9216-dimensional vector, which was then compressed into a 128-dimensional latent representation by a fully connected autoencoder (FC-AE) [34,35]. This global latent vector was used as the gating classifier input. Expert architecture was identical to that of MoE-SG and MoE-AEG. This model was designed to evaluate whether a FC-AE-derived global latent representation is sufficient to facilitate expert specialization and gating alignment.

MoE-zSE applied zone-specific expert architectures under the same FC-AE-based gating conditions as MoE-FCAE. Expert A used two Conv1D layers (kernel = 3, filters = 32); Expert B used a Conv1D layer (kernel = 5, filters = 32) followed by another (kernel = 5, filters = 64); Expert C used three Conv1D layers (kernel = 7, filters = 64; kernel = 5, filters = 64; kernel = 3, filters = 32). All experts employed ReLU activation, global average pooling, and a fully connected regression output layer. This model was designed to evaluate whether zone-specific expert architecture design, in conjunction with FC-AE-based gating, produces additional improvements in final regression performance.

The configurations and primary comparison objectives of the CNN baseline and the four MoE variants are summarized in Table 1.

Table 1. Model configurations and primary comparison objectives.

Model	Gating Input	Expert Architecture	Primary Comparison Objective
CNN baseline	None	Single CNN	Direct regression reference model
MoE-SG	Mean, SD, skewness per channel	Identical CNN experts	Assess statistical feature routing
MoE-AEG	BiLSTM encoder 32-d latent	Identical CNN experts	Assess sequence-level temporal latent routing
MoE-FCAE	FC-AE 128-d global latent	Identical CNN experts	Assess FC-AE global latent routing
MoE-zSE	FC-AE 128-d global latent	Zone-specific CNN experts	Assess zone-specific expert architecture effect

2.7.1. Distinction Between BiLSTM Sequence Encoder and FC-AE Latent Representations

Both MoE-AEG and MoE-FCAE were designed to supply the gating network with encoder-derived latent representations rather than handcrafted statistical features, based on the premise that simple statistics such as mean, standard deviation, and skewness may be insufficient to capture the temporal dynamics and inter-zone differences embedded in the plantar pressure time series.

Despite both employing learned latent representations as gating inputs, the two models differ in how these representations are generated. MoE-AEG preserves the sequential structure of the 36×256 input and extracts a 32-dimensional temporal latent feature through a BiLSTM-based sequence encoder, maintaining the temporal ordering of the signal. By contrast, MoE-FCAE flattens the same input into a 9216-dimensional vector and compresses it into a 128-dimensional global latent vector through a fully connected autoencoder, without preserving sequential structure. MoE-AEG thus evaluates the utility of sequence-level temporal representations for routing, whereas MoE-FCAE evaluates the utility of a globally compressed latent representation derived from the flattened input.

2.7.2. Expert Training, Routing, and Model Selection Protocol

No separate validation set was reserved within each LOSO fold, and validation-based early stopping was not used. All networks were therefore trained for a predefined number of epochs using the training settings reported in Table 2. In each LOSO fold, the held-out participant was used only for testing, while model training and target normalization were conducted using data from the remaining participants.

Table 2. Architecture and training settings.

Component	Model	Architecture	Training Settings
Baseline CNN	CNN baseline	36-ch input $\rightarrow$ Conv1D($k = 7, f = 128$) $\rightarrow$ Conv1D($k = 5, f = 256$) $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ GAP $\rightarrow$ FC regression	Adam, lr = 0.001, epochs = 50, batch = 32; fixed-epoch training without validation-based early stopping
Common CNN expert	MoE-SG, MoE-AEG, MoE-FCAE	36-ch input $\rightarrow$ Conv1D($k = 3, f = 64$, same) $\rightarrow$ ReLU $\rightarrow$ BN $\rightarrow$ GAP $\rightarrow$ FC regression	Adam, epochs = 25, batch = 16
FC-AE	MoE-FCAE, MoE-zSE	9216-d input $\rightarrow$ FC encoder $\rightarrow$ 128-d latent $\rightarrow$ FC decoder $\rightarrow$ 9216-d output	Adam, trained within each LOSO training fold
Statistical gating	MoE-SG	Mean, SD, skewness $\rightarrow$ FC classifier (softmax, 3 classes)	Adam, epochs = 20, batch = 16
BiLSTM encoder gating	MoE-AEG	SeqInput(36) $\rightarrow$ BiLSTM(64, last) $\rightarrow$ FC(32) $\rightarrow$ FC classifier (softmax, 3 classes)	Encoder: Adam, epochs = 10, batch = 16; Classifier: Adam, epochs = 20, batch = 16
FC-AE gating	MoE-FCAE, MoE-zSE	128-d FC-AE latent $\rightarrow$ FC classifier (softmax, 3 classes)	Adam, trained within each LOSO training fold
Zone-specific experts	MoE-zSE	Expert A: Conv1D($k = 3, f = 32$) $\times 2$; Expert B: Conv1D($k = 5, f = 32$) $\rightarrow$ ($k = 5, f = 64$); Expert C: Conv1D($k = 7, f = 64$) $\rightarrow$ ($k = 5, f = 64$) $\rightarrow$ ($k = 3, f = 32$); all with ReLU, GAP, FC regression	Adam, batch = 16, trained within each LOSO fold
Evaluation	All models	LOSO cross-validation; 5 repeated runs per fold	Results: mean $\pm$ SD across 5 repetitions

Note: BN = batch normalization; GAP = global average pooling; FC = fully connected; k = kernel size; f = number of filters. MoE-SG, MoE-AEG, and MoE-FCAE shared the same CNN expert architecture; their primary differences lay in the gating input representation. MoE-zSE applied zone-specific expert designs under otherwise identical FC-AE gating conditions. The baseline CNN architecture was selected from five predefined candidate configurations through preliminary LOSO comparisons. No separate validation set or nested cross-validation procedure was used for architecture selection.

For expert training, the samples remaining in each LOSO training fold were assigned to Zones A, B, and C according to their true stride lengths, and Experts A, B, and C were trained independently using samples from their corresponding zones. The Zone A and Zone C training sets were supplemented using the augmentation procedures described in Section 2.4, whereas the held-out participant's data were never augmented or used for expert training. A common target mean and standard deviation were calculated from the expert-training samples within each fold and were applied to normalize the targets of all three experts. At inference, each test sequence was provided to all three experts. The gating network generated three softmax weights from the corresponding gating representation, and the final prediction was calculated as the weighted sum of the three normalized expert outputs, followed by inverse transformation to the original stride-length scale. All expert training, gating-network training, target normalization, and test prediction procedures were conducted independently within each LOSO fold.

For the baseline CNN, a limited set of five predefined architectural configurations was evaluated in preliminary LOSO experiments. The candidate configurations varied the numbers of convolutional filters, kernel sizes, and the use of dropout. The evaluated combinations were: (1) filters of 64 and 128 with kernel sizes of 5 and 3 without dropout; (2) filters of 128 and 256 with kernel sizes of 7 and 5 without dropout; (3) filters of 64 and 128 with kernel sizes of 5 and 3 and a dropout rate of 0.3; (4) filters of 128 and 256 with

kernel sizes of 7 and 5 and a dropout rate of 0.3; and (5) filters of 256 and 256 with kernel sizes of 5 and 5 without dropout.

The fourth configuration yielded the lowest aggregate LOSO prediction error and was therefore selected as the final CNN baseline architecture. The optimizer-related settings were held constant across all candidate configurations: Adam optimization, an initial learning rate of 0.001, a mini-batch size of 32, and 50 training epochs. Thus, the procedure constituted limited candidate-based architecture selection rather than exhaustive hyperparameter optimization or nested cross-validation.

For the MoE models, the architecture- and component-specific settings shown in Table 2 were fixed across all LOSO folds and repeated runs and were not individually optimized for each held-out participant.

For the MoE models, the hyperparameters were predefined before the repeated LOSO experiments and held constant across all folds and runs to avoid fold-specific tuning. The BiLSTM encoder and FC-AE were based on established sequence-encoding and autoencoder formulations [32–35], while the overall expert-gating structure followed the conventional MoE framework [21,22]. The 32-dimensional BiLSTM projection and the 128-dimensional FC-AE latent representation were selected as reduced representations relative to the original 36×256 input. The common expert architecture was intentionally kept identical across MoE-SG, MoE-AEG, and MoE-FCAE so that differences among these models primarily reflected the gating representation rather than expert capacity. MoE-zSE was included separately to examine whether heterogeneous expert architectures altered specialization. The epoch numbers, mini-batch sizes, and latent dimensions were not determined through exhaustive hyperparameter optimization and are therefore interpreted as fixed experimental settings rather than task-optimal values.

2.8. Performance Evaluation and Statistical Analysis

Model performance was quantified using RMSE, mean absolute error (MAE), mean error (ME), and error standard deviation (Error SD). RMSE and MAE were used as measures of overall prediction accuracy. ME was computed as the mean signed difference between predicted and actual stride lengths and was used to assess directional bias; positive ME indicates overestimation and negative ME indicates underestimation. Error SD reflected the dispersion of prediction errors.

Zone-wise analyses stratified trials into Zones A, B, and C based on true stride length, and RMSE and ME were computed separately within each zone. To quantify the degree of zone-specific improvement or deterioration relative to the CNN baseline, ΔRMSE was computed as $\text{RMSE}_{\text{CNN}} - \text{RMSE}_{\text{MoE}}$; positive ΔRMSE indicates that the MoE model achieved lower RMSE than the CNN (improvement), and negative ΔRMSE indicates higher RMSE (deterioration).

Primary statistical comparisons were conducted at the subject level, consistent with the LOSO evaluation structure. For each test subject, per-model overall RMSE values were computed, and zone-wise analyses further decomposed these into subject-wise RMSE estimates within each of the three zones. The subject-wise RMSE, rather than individual trial error, served as the primary unit of comparison.

Subject-wise RMSE values derived from the five repeated runs were first averaged within each combination of SubjectID, Model, and Zone, then used in paired comparisons between the CNN baseline and each MoE model. The normality of pairwise differences was assessed, and paired *t*-tests were applied when the normality assumption was satisfied; otherwise, Wilcoxon signed-rank tests were used. Multiple comparison correction was performed using the Holm procedure [36]. Both unadjusted *p*-values and Holm-adjusted *p*-values were reported for each paired comparison. Effect sizes were calcu-

lated as Cohen's d_z for paired t -tests and matched-pairs rank-biserial correlation (r_{rb}) for Wilcoxon signed-rank tests. The sign of each effect size followed the definition of ΔRMSE ($\text{CNN RMSE} - \text{MoE RMSE}$), with positive values favoring the MoE model and negative values favoring the CNN baseline.

To quantify uncertainty in the paired model differences, two-sided 95% confidence intervals for the mean ΔRMSE were estimated using percentile bootstrap resampling of the paired subject-level differences with 100,000 resamples. These confidence intervals were not adjusted for multiple comparisons; Holm-adjusted p -values were retained as the primary multiplicity-controlled inference.

To evaluate whether model effects varied across stride length zones, a linear mixed-effects model (LME) was additionally fitted with subject-wise zone RMSE as the dependent variable [37]. Model, Zone, and the Model $\times$ Zone interaction were included as fixed effects, and SubjectID and Repeat were included as random intercepts:

$$\text{RMSE}_{\text{subject-zone}} \sim \text{Model} \times \text{Zone} + (1 \mid \text{SubjectID}) + (1 \mid \text{Repeat})$$

Computational complexity was additionally evaluated descriptively using the number of trainable parameters, inference-active parameters, and CPU-based inference time per stride. For each model, inference latency was measured using the same 18 test strides for each of the five independently trained runs and was summarized as the mean $\pm$ standard deviation across the five runs. Detailed measurement procedures and results are provided in Supplementary Materials S2.

All signal processing, model development, training, prediction, and statistical analyses were performed using MATLAB R2026a (MathWorks, Natick, MA, USA).

2.9. Analysis Sample and Diagnostic Indices

2.9.1. Analysis Sample

The final analysis included 52 trials from 17 subjects in Zone A, 121 trials from 20 subjects in Zone B, and 64 trials from 14 subjects in Zone C (Table 3). Performance summary statistics were reported as mean $\pm$ SD across five repeated runs. Primary statistical comparisons between the CNN baseline and each MoE model were conducted using subject-wise RMSE or subject-wise zone RMSE as the unit of analysis, consistent with the LOSO structure.

Table 3. Sample sizes by stride length zone.

Zone	Definition	N Trials	N Subjects
Zone A (Short)	<125 cm	52	17
Zone B (Middle)	125–135 cm	121	20
Zone C (Long)	>135 cm	64	14

2.9.2. Routing Diagnostic Indices

To characterize the internal routing behavior of MoE models beyond standard performance metrics, a set of task-specific diagnostic indices was defined. These indices were not intended as primary outcome measures for statistical model comparison but rather as mechanistic diagnostic tools to elucidate why expert-level performance advantages may fail to propagate to the final MoE prediction.

For each trial i , the true stride length was denoted by y_i , the independent predictions of Experts A, B, and C as $\hat{y}_{A,i}$, $\hat{y}_{B,i}$, and $\hat{y}_{C,i}$, respectively, and the gating network's softmax weights as $w_{A,i}$, $w_{B,i}$, and $w_{C,i}$. The final MoE prediction was given by:

$$\hat{y}_{\text{MoE},i} = w_{A,i} \cdot \hat{y}_{A,i} + w_{B,i} \cdot \hat{y}_{B,i} + w_{C,i} \cdot \hat{y}_{C,i}$$

Gating confidence was defined as the maximum softmax weight for trial i : $\text{confidence}_i = \max(w_{A,i}, w_{B,i}, w_{C,i})$. Gating entropy was computed as

$$H_i = - \sum_K \omega_{k,i} \log(\omega_{k,i} + \varepsilon)$$

where ε is a small constant to avoid numerical instability. Lower entropy and higher confidence indicate a more concentrated gating distribution, suggesting more decisive routing behavior by the gating network.

The routed expert was defined as the expert receiving the highest gating weight for trial i . The true-zone alignment rate for zone z was defined as the proportion of trials in zone z for which the routed expert matched the expected expert (Expert A for Zone A, Expert B for Zone B, Expert C for Zone C):

$$\text{TrueZoneAlign}_Z = \frac{1}{N_Z} \sum_{i \in Z} 1[\text{routed}_i = e_Z]$$

where N_Z is the number of trials in zone z and e_Z is the expected expert for zone Z .

The best expert for each trial was defined as the expert with the lowest absolute prediction error among the three expert-only predictions. The best-expert alignment rate was defined as the proportion of trials for which the routed expert coincided with the best expert. Routing regret was computed as the difference in absolute error between the final MoE prediction and the best expert prediction:

$$\text{regret}_i = |\hat{y}_{\text{MoE},i} - y_i| - \min_k |\hat{y}_{k,i} - y_i|$$

Positive routing regret indicates that the final MoE prediction incurred greater absolute error than would have been achieved by routing to the best expert.

The high-confidence misalignment rate was defined as the proportion of trials with gating confidence $\geq \tau$ for which the routed expert did not match the best expert, where $\tau = 0.8$ was used as the high-confidence threshold. This index quantifies cases in which the gating network committed to a specific expert with high confidence yet selected an expert that was not the locally optimal choice.

3. Results

3.1. Overall Prediction Performance

Across all models, the CNN baseline achieved the lowest and most stable overall error (RMSE: 7.54 ± 0.14 cm; MAE: 6.14 ± 0.16 cm). Among the MoE-based models, MoE-AEG yielded the lowest RMSE (8.65 ± 0.41 cm); however, no MoE model achieved a lower overall RMSE than the CNN baseline. In paired comparisons of subject-wise RMSE, MoE-FCAE and MoE-zSE showed significantly higher RMSE than the CNN baseline after Holm correction, whereas the differences for MoE-SG and MoE-AEG did not reach statistical significance (Table 4). These results indicate that the introduction of expert routing did not yield consistent overall performance gains relative to the single CNN in the present dataset.

As a supplementary sensitivity analysis, the CNN baseline was additionally evaluated after removing the first- and second-order difference features. The resulting 18-channel CNN yielded a numerically higher overall RMSE than the original 36-channel CNN (7.95 ± 0.48 cm vs. 7.54 ± 0.14 cm across five repeated LOSO runs). However, the subject-wise paired difference was not statistically significant (Wilcoxon signed-rank test, unadjusted $p = 1.000$; matched-pairs rank-biserial correlation, $r_{rb} = -0.004$). Therefore, the analysis did not identify a systematic effect of the difference features on subject-independent prediction accuracy.

Table 4. Overall stride length prediction performance (mean $\pm$ SD across five repeated runs).

Model	RMSE (cm)	MAE (cm)	Mean Error (cm)	Error SD (cm)
CNN baseline	7.54 $\pm$ 0.14	6.14 $\pm$ 0.16	+0.44 $\pm$ 0.41	7.53 $\pm$ 0.13
MoE-SG	10.88 $\pm$ 1.50	7.87 $\pm$ 0.61	−5.59 $\pm$ 0.58	9.32 $\pm$ 1.67
MoE-AEG	8.65 $\pm$ 0.41	6.66 $\pm$ 0.27	−3.86 $\pm$ 0.22	7.76 $\pm$ 0.39
MoE-FCAE	9.10 $\pm$ 0.47 *	7.29 $\pm$ 0.33	+0.33 $\pm$ 0.64	9.10 $\pm$ 0.49
MoE-zSE	10.64 $\pm$ 1.19 ***	8.16 $\pm$ 0.64	−0.88 $\pm$ 0.83	10.61 $\pm$ 1.14

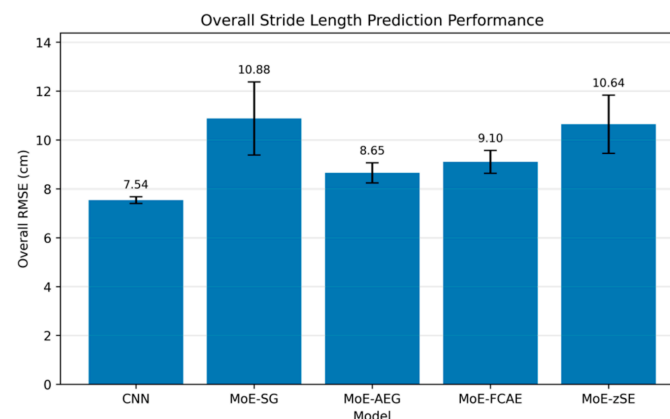
Note: Values represent mean $\pm$ SD across five repeated runs. Asterisks in the RMSE column denote statistically significant differences from the CNN baseline in subject-wise paired comparisons after Holm correction, as reported in Table 5. In this table, asterisks indicate significant deterioration for MoE-FCAE and MoE-zSE. * Holm- p < 0.05; *** Holm- p < 0.001.

Table 5. Paired comparison of subject-wise overall RMSE: CNN baseline vs. each MoE model.

Comparison	CNN RMSE (cm)	MoE RMSE (cm)	Δ RMSE	Test	Unadjusted p	Holm-Adjusted p	Effect Size (Cohen's d_z)	95% CI for Δ RMSE
CNN vs. MoE-SG	7.22 $\pm$ 1.78	9.01 $\pm$ 4.93	−1.79	Paired t -test	0.0896	0.1792	−0.379	[−3.80, 0.05]
CNN vs. MoE-AEG	7.22 $\pm$ 1.78	7.68 $\pm$ 3.30	−0.47	Paired t -test	0.5031	0.5031	−0.145	[−1.85, 0.77]
CNN vs. MoE-FCAE	7.22 $\pm$ 1.78	8.53 $\pm$ 2.54	−1.32 *	Paired t -test	0.0158	0.0474	−0.560	[−2.34, −0.42]
CNN vs. MoE-zSE	7.22 $\pm$ 1.78	9.76 $\pm$ 2.92	−2.54 ***	Paired t -test	0.0001	0.0004	−1.013	[−3.60, −1.55]

Note: RMSE values in this table reflect subject-wise RMSE averaged across subjects, which differs from the trial-pooled RMSE reported in Table 4 due to differences in aggregation unit. Δ RMSE = CNN RMSE − MoE RMSE; positive values indicate lower MoE RMSE (improvement) and negative values indicate higher MoE RMSE (deterioration). Asterisks indicate statistically significant differences after Holm correction. * Holm- p < 0.05; *** Holm- p < 0.001.

The overall RMSE comparison among the CNN baseline and the four MoE variants is presented in Figure 2, the corresponding overall prediction performance is summarized in Table 4. In paired comparisons of subject-wise RMSE, MoE-FCAE and MoE-zSE showed significantly higher RMSE than the CNN baseline after Holm correction, whereas the differences for MoE-SG and MoE-AEG did not reach statistical significance (Table 5).

**Figure 2.** Overall RMSE comparison between the CNN baseline and MoE variants. Bars and error bars represent the mean and standard deviation across five repeated LOSO experiments, respectively.

The 95% confidence intervals represent unadjusted percentile bootstrap intervals for the mean paired Δ RMSE based on subject-level differences. Statistical significance after multiple-comparison correction should be interpreted using the Holm-adjusted p -values.

The supplementary augmentation ablation showed no statistically significant difference in overall or zone-specific CNN performance after Holm correction. Therefore, augmentation was not interpreted as an independent source of performance improvement.

3.2. Zone-Wise Prediction Performance and Model $\times$ Zone Interaction

Zone-wise analyses revealed that prediction errors were not uniformly distributed across stride length regions (Figure 3 and Table 6). Even for the CNN baseline, Zone B yielded the lowest RMSE (5.52 ± 0.49 cm), while Zones A and C exhibited higher errors of 8.76 ± 0.38 cm and 9.49 ± 0.41 cm, respectively, confirming that the tail regions represent more challenging regression targets in plantar-pressure-based stride estimation.

MoE-AEG significantly reduced RMSE in Zone A to 6.59 ± 0.58 cm relative to the CNN baseline in subject-wise zone comparisons (Holm- $p = 0.0357$). A trend toward improvement was observed in Zone B as well, though this did not reach statistical significance after Holm correction (Holm- $p = 0.438$). Conversely, Zone C RMSE increased to 14.10 ± 0.62 cm, representing a statistically significant deterioration relative to the CNN (Holm- $p = 0.0092$). MoE-SG similarly showed a numerically lower Zone A RMSE, but this improvement was not statistically significant (Holm- $p = 0.135$), whereas Zone C deterioration was highly significant (Holm- $p = 0.0014$). Neither MoE-FCAE nor MoE-zSE demonstrated statistically significant improvement in any zone; MoE-zSE showed significant deterioration in Zone B (Holm- $p = 0.0014$) (Table 7).

The LME analysis confirmed significant main effects of Model ($F(4, 1260) = 22.52$, $p < 0.001$) and Zone ($F(2, 1260) = 16.91$, $p < 0.001$), as well as a significant Model $\times$ Zone interaction ($F(8, 1260) = 24.21$, $p < 0.001$). This interaction indicates that MoE effects varied substantially across stride length zones, and no single MoE architecture provided uniformly superior performance across all zones.

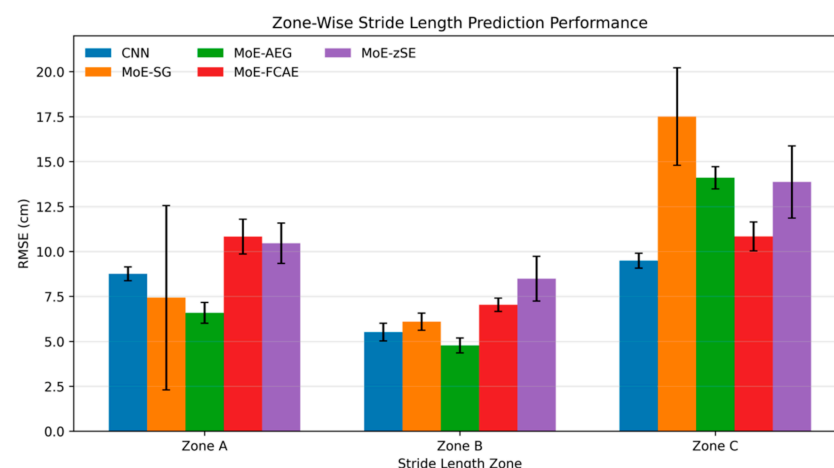


Figure 3. Zone-wise RMSE comparison between the CNN baseline and MoE variants. Bars and error bars represent the mean and standard deviation across five repeated LOSO experiments, respectively. Zone assignments were based on true stride length.

Table 6. Zone-wise RMSE performance (mean $\pm$ SD across five repeated runs, cm).

Model	Zone A RMSE	Zone B RMSE	Zone C RMSE	Overall RMSE
CNN baseline	8.76 $\pm$ 0.38	5.52 $\pm$ 0.49	9.49 $\pm$ 0.41	7.54 $\pm$ 0.14
MoE-SG	7.43 $\pm$ 5.13	6.09 $\pm$ 0.47	17.51 $\pm$ 2.72 **	10.88 $\pm$ 1.50
MoE-AEG	6.59 $\pm$ 0.58 *	4.77 $\pm$ 0.41	14.10 $\pm$ 0.62 **	8.65 $\pm$ 0.41
MoE-FCAE	10.83 $\pm$ 0.97	7.03 $\pm$ 0.37	10.84 $\pm$ 0.81	9.10 $\pm$ 0.47 *
MoE-zSE	10.46 $\pm$ 1.13	8.49 $\pm$ 1.25 **	13.87 $\pm$ 2.01	10.64 $\pm$ 1.19 ***

Note: Asterisks in Zone A–C columns indicate statistically significant differences from the CNN baseline in subject-wise zone RMSE paired comparisons after Holm correction; asterisks in the Overall RMSE column reflect subject-wise overall RMSE comparisons (Table 5). Asterisks denote both significant improvements and significant deteriorations; the direction should be interpreted in conjunction with the RMSE values and Table 7. * Holm- $p < 0.05$; ** Holm- $p < 0.01$; *** Holm- $p < 0.001$.

Table 7. Zone-wise Δ RMSE relative to the CNN baseline.

Zone	Comparison (CNN vs.)	Δ RMSE	Test	Unadjusted p	Holm-Adjusted p	Effect Size	95% CI for Δ RMSE
A	MoE-SG	+3.12	Paired t -test	0.0176	0.1350	+0.642	[0.77, 5.25]
	MoE-AEG	+2.98 *	Paired t -test	0.0040	0.0357	+0.815	[1.38, 4.76]
	MoE-FCAE	−0.83	Wilcoxon signed-rank	0.3812	0.7623	−0.242	[−2.90, 1.29]
B	MoE-zSE	−1.52	Paired t -test	0.0212	0.1350	−0.620	[−2.64, −0.38]
	MoE-SG	−0.23	Paired t -test	0.7227	0.7623	−0.081	[−1.44, 0.98]
	MoE-AEG	+0.89	Paired t -test	0.1142	0.4376	+0.370	[−0.13, 1.91]
	MoE-FCAE	−1.56	Wilcoxon signed-rank	0.0169	0.1350	−0.610	[−2.57, −0.64]
C	MoE-zSE	−2.33 **	Wilcoxon signed-rank	0.0001	0.0014	−0.981	[−3.24, −1.53]
	MoE-SG	−6.98 **	Wilcoxon signed-rank	0.0001	0.0014	−1.000	[−9.85, −4.44]
	MoE-AEG	−4.56 **	Paired t -test	0.0009	0.0092	−1.140	[−6.66, −2.64]
	MoE-FCAE	−1.19	Paired t -test	0.1094	0.4376	−0.459	[−2.49, 0.14]
	MoE-zSE	−3.06	Paired t -test	0.0248	0.1350	−0.678	[−5.50, −0.94]

Note: Δ RMSE = CNN RMSE − MoE RMSE, computed from subject-wise zone RMSE. Values may therefore differ slightly from the arithmetic difference of mean values in Table 6 due to the subject-level aggregation approach. Positive values indicate improvement (lower MoE RMSE); negative values indicate deterioration. Statistical tests were conducted as subject-wise paired comparisons with Holm correction. * Holm- $p < 0.05$; ** Holm- $p < 0.01$. Effect sizes are reported as Cohen's d_z for paired t -tests and matched-pairs rank-biserial correlation (r_{rb}) for Wilcoxon signed-rank tests. The sign follows the definition of Δ RMSE; positive values favor the MoE model, whereas negative values favor the CNN baseline. Confidence intervals are unadjusted 95% percentile bootstrap intervals based on paired subject-level Δ RMSE values. Holm-adjusted p -values should be used when determining statistical significance after multiple-comparison correction.

3.3. Directional Bias Analysis

The CNN baseline exhibited positive ME in Zone A (+7.73 cm), indicating systematic overestimation of short strides, and negative ME in Zone C (−7.37 cm), indicating systematic underestimation of long strides (Table 8). This pattern is consistent with the regression-to-the-mean phenomenon that motivated the present study.

Table 8. Zone-wise mean error (cm).

Model	Zone A ME	Zone B ME	Zone C ME	Overall ME
CNN baseline	+7.73	+1.44	−7.37	+0.44
MoE-SG	+4.44	−4.54	−15.75	−5.59
MoE-AEG	+5.48	−3.12	−12.84	−3.86
MoE-FCAE	+8.22	+1.35	−8.02	+0.33
MoE-zSE	+7.23	+0.21	−9.52	−0.88

MoE-SG and MoE-AEG partially attenuated the positive bias in Zone A; however, they substantially amplified the negative bias in Zone C. For example, MoE-AEG reduced Zone A ME to +5.48 cm relative to the CNN, yet Zone C ME increased to −12.84 cm, representing a considerably larger underestimation of long strides. The apparent improvement in Zone A for these models therefore reflects a region-wise trade-off rather than simultaneous correction of both tail regions. Given that MoE-SG, MoE-AEG, and MoE-FCAE shared identical expert architectures, the observed inter-model differences are primarily attributable to differences in the gating input representation.

3.4. Subject-Wise Robustness

The CNN baseline demonstrated the greatest subject-wise stability, with a worst-subject RMSE of 11.04 ± 1.96 cm (Table 9). By contrast, MoE-SG exhibited a worst-subject RMSE of 27.90 ± 12.11 cm, indicating severe performance collapse for certain individuals. MoE-AEG, MoE-FCAE, and MoE-zSE all showed higher worst-case indices than the CNN baseline, suggesting that localized improvements observed in specific zones or subjects did not generalize robustly to unseen individuals.

Table 9. Subject-wise robustness summary.

Model	Mean RMSE (cm)	Worst Top 10% RMSE (cm)	Worst-Subject RMSE (cm)
CNN baseline	7.22 ± 0.19	10.14 ± 0.56	11.04 ± 1.96
MoE-SG	9.01 ± 0.91	20.56 ± 5.20	27.90 ± 12.11
MoE-AEG	7.68 ± 0.37	13.61 ± 1.42	16.02 ± 3.39
MoE-FCAE	8.53 ± 0.48 *	13.60 ± 1.19	15.29 ± 1.22
MoE-zSE	9.76 ± 0.76 ***	17.08 ± 3.16	19.70 ± 3.83

Note: Asterisks in the Mean RMSE column reflect subject-wise overall RMSE paired comparisons with Holm correction (Table 5). Worst-case indices are presented as descriptive statistics. * Holm- $p < 0.05$; *** Holm- $p < 0.001$.

Among the three models sharing identical CNN expert architectures (MoE-SG, MoE-AEG, MoE-FCAE), differences in subject-wise robustness are attributable primarily to variations in the gating input representation. MoE-AEG exhibited a relatively lower worst-case error compared to MoE-SG, consistent with the hypothesis that richer latent representations may contribute to more stable routing. Nevertheless, neither MoE-AEG nor MoE-FCAE achieved a statistically significant performance advantage over the CNN baseline in subject-wise comparisons, and MoE-FCAE showed significantly higher subject-wise RMSE than the CNN despite using a learned latent representation for gating.

3.5. Separability Analysis

The separability summary also showed that better routing features did not automatically produce better final prediction (Table 10). MoE-SG showed extremely poor region separability (accuracy 0.219, macro-F1 0.120), consistent with its collapsed routing behavior. MoE-AEG improved separability to some extent (accuracy 0.499, macro-F1 0.253), but this did not translate into strong tail-region performance. MoE-FCAE showed the highest separability metrics (accuracy 0.754, macro-F1 0.734), yet its overall RMSE still remained worse than CNN.

Table 10. Region separability summary.

Model	Accuracy	Macro-F1	Purity
MoE-SG	0.219	0.120	0.511
MoE-AEG	0.499	0.253	0.511
MoE-FCAE	0.754	0.734	0.754
MoE-zSE	0.369	0.227	0.511

This finding reveals a critical mismatch between classification separability and regression utility. Although the gating model in MoE-FCAE achieved the highest region-separability performance, this did not translate into the best continuous prediction accuracy. Therefore, accurate region classification alone was not sufficient to guarantee improved regression performance. Even when the gating network could separate regions reasonably well, the final MoE still failed to outperform the simpler single CNN baseline in a stable and statistically convincing manner (Table 11).

Table 11. Gating summary.

Model	MaxProb Mean	Entropy Mean	Assign A	Assign B	Assign C
MoE-SG	0.447	1.066	1	0.000	0.000
MoE-AEG	0.446	1.061	0.040	0.959	0.001
MoE-FCAE	0.932	0.166	0.185	0.539	0.276
MoE-zSE	0.941	0.144	0.246	0.480	0.274

Note: Interpretation—MoE-SG: Degenerate routing to A, MoE-AEG: Degenerate routing to B, MoE-FCAE: Confident but not beneficial, MoE-zSE: Confident but not beneficial. Notably, routing confidence alone did not translate into improved prediction performance.

The corresponding row-normalized confusion matrices are presented in Figure 4 to illustrate the zone-specific assignment patterns of each gating model.

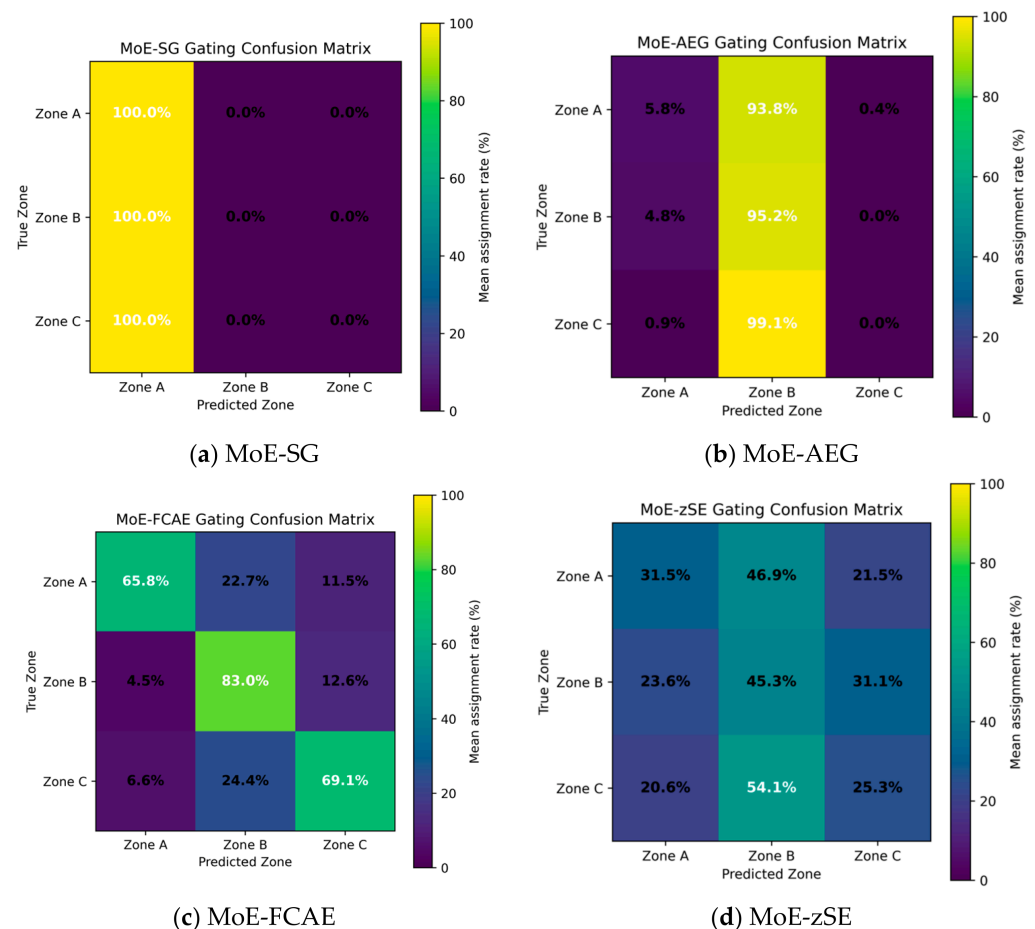


Figure 4. Row-normalized confusion matrices of gating-based stride-zone assignments: (a) MoE-SG, (b) MoE-AEG, (c) MoE-FCAE, and (d) MoE-zSE. Each cell represents the mean assignment rate across five repeated LOSO experiments. Rows indicate true zones and columns indicate predicted zones.

3.6. Separation of Expert Specialization and Gating Alignment in MoE-FCAE

Expert-only analyses were conducted independently of the final weighted prediction to assess whether expert networks developed distinct predictive functions through zone-specific training (Table 12). In MoE-FCAE, all three experts shared the same CNN architecture but were trained on zone-specific training samples.

Table 12. Expert-only RMSE matrix for MoE-FCAE (cm).

Test Zone	Expert A	Expert B	Expert C	Best Expert	Expected Expert
Zone A	2.96	8.71	16.60	Expert A	A
Zone B	8.32	3.99	9.54	Expert B	B
Zone C	17.64	10.85	5.31	Expert C	C

Expert A achieved its lowest expert-only RMSE in Zone A (2.96 cm), Expert B in Zone B (3.99 cm), and Expert C in Zone C (5.31 cm). The diagonal dominance of the expert-only RMSE matrix confirms that functional expert specialization emerged through zone-specific training, even without architectural differentiation across experts. Despite this, the final routed predictions did not reach the levels of performance observed in the expert-only evaluations, and overall and zone-wise performance did not significantly improve relative to the CNN baseline.

Gating diagnostic results revealed a systematic misalignment between expert-only performance and final routed predictions (Table 13). Across all true zones, mean gating weight was highest for Expert B: $w_B = 0.430$ in Zone A, 0.451 in Zone B, and 0.535 in Zone C. True-zone alignment rates were low across all zones (Zone A: 0.292; Zone B: 0.446; Zone C: 0.275), as were best-expert alignment rates (Zone A: 0.288; Zone B: 0.398; Zone C: 0.284). Mean routing regret was positive in all zones, indicating that the final routed predictions consistently incurred greater absolute error than the best available expert-only prediction. These findings suggest that the zone-specific predictive advantages observed at the expert level were not adequately transferred to the final weighted prediction.

Table 13. True-zone-wise gating alignment and routing regret for MoE-FCAE.

True Zone	Mean w_A	Mean w_B	Mean w_C	Dominant	True-Zone Align	Best-Expert Align	Routing Regret	High-Conf Misalign
Zone A	0.304	0.430	0.266	Expert B	0.292	0.288	6.55 cm	0.592
Zone B	0.240	0.451	0.309	Expert B	0.446	0.398	3.13 cm	0.514
Zone C	0.189	0.535	0.276	Expert B	0.275	0.284	5.52 cm	0.641

Note: w_A , w_B , w_C = mean softmax weights assigned to Experts A, B, and C, respectively. True-zone alignment = proportion of trials in which the routed expert matched the expected expert for the true stride length zone. Best-expert alignment = proportion of trials in which the routed expert matched the expert with the lowest expert-only absolute error. Routing regret = mean difference in absolute error between the final MoE prediction and the best expert-only prediction. High-confidence misalignment = proportion of trials with gating confidence ≥ 0.8 , for which the routed expert did not match the best expert. Positive routing regret indicates the final prediction was less accurate than the best available expert.

4. Discussion

4.1. Integrated Interpretation

This study showed that the effectiveness of MoE in plantar-pressure-based stride length regression depended on the joint satisfaction of input separability, expert specialization, and gating alignment [21,22]. Although individual MoE variants partially satisfied one or more of these conditions, no model satisfied them simultaneously or maintained a consistent performance advantage across unseen subjects.

These findings indicate that expert-level specialization cannot be interpreted as evidence of effective MoE operation unless the gating network reliably transfers the corresponding expert advantage to the final prediction. In a continuous and overlapping biomechanical regression problem, target-space partitioning may therefore produce region-specific trade-offs rather than stable overall performance gains.

4.2. Tail-Region Trade-Off and Regression-to-the-Mean

The CNN baseline exhibited positive ME in Zone A and negative ME in Zone C, consistent with the regression-to-the-mean phenomenon that motivated the present study. This directional bias pattern is commonly observed in single-model regressors trained on long-tailed continuous targets with high density in the central range [17].

Although some MoE variants reduced prediction error in one tail region, these gains were accompanied by deterioration in the opposite tail. This pattern indicates that expert routing redistributed prediction error across the target distribution rather than providing balanced correction of both extremes.

Accordingly, claiming tail-region improvement for MoE in similar settings requires evidence of simultaneous improvement across both tails, not merely partial improvement in one zone. Any evaluation of tail correction should accompany assessment of the complementary tail region and overall subject-wise robustness.

4.3. Misalignment Between Expert Specialization and Gating Decisions

The most important mechanistic finding of this study is the dissociation between expert specialization and final prediction performance observed in MoE-FCAE. The expert-only RMSE matrix confirmed diagonal dominance—Experts A, B, and C each achieved their lowest error within the corresponding zone—indicating that zone-specific training successfully induced functional specialization even without architectural differentiation. Despite this, the final MoE predictions did not surpass the CNN baseline.

Zone B contained 121 trials—substantially more than Zone A (52 trials) or Zone C (64 trials). The systematic assignment of the highest gating weight to Expert B across all true zones represents a plausible mechanism underlying the dissociation between expert specialization and final regression performance. This routing pattern may be related to the greater number of Zone B samples and the use of a zone-classification objective. However, because regression-aware and load-balancing gating objectives were not evaluated, the present results cannot determine whether the observed routing pattern was caused by the classification loss, the imbalanced zone distribution, the gating representation, or interactions among these factors. Therefore, the Expert B-dominant routing should be interpreted as evidence of gating misalignment rather than as evidence that the classification objective directly caused the MoE performance degradation.

Critically, high-confidence misalignment rates were substantial across all zones (Zone A: 0.592; Zone B: 0.514; Zone C: 0.641), indicating that the gating network committed to suboptimal routing with considerable certainty rather than through uncertainty. This pattern underscores the limitation of relying solely on expert-level performance evaluation: a gating network can be both confident and miscalibrated [38], and expert-only advantages cannot be assumed to translate into final prediction gains without verified gating alignment.

4.4. Scope of Model Comparisons

The MoE variants in this study were not designed as a performance ranking exercise but rather as a controlled comparison aimed at decomposing the conditions required for effective MoE operation. MoE-SG and MoE-AEG maintained identical expert architectures to isolate the effect of gating input representation. MoE-FCAE and MoE-zSE introduced FC-AE-based latent representations and zone-specific expert designs, respectively, and their results cannot therefore be attributed solely to differences in gating input.

The primary interpretive units of this study were the three structural conditions: whether plantar pressure features or latent representations provide sufficient information for stride zone discrimination; whether experts achieve lower error within their designated zones; and whether gating decisions are aligned with the true stride zone or the locally

optimal expert. Partial satisfaction of these conditions was observed across models, but no single model fulfilled all three simultaneously, which explains the absence of consistent overall improvement.

Compared with previous wearable-sensor studies that primarily evaluated the accuracy of gait-parameter estimation, the contribution of the present study is mechanistic rather than performance-oriented. Hannink et al. demonstrated that deep CNNs can estimate stride length from wearable inertial signals, while Zadka et al. showed that machine-learning-based step-length estimation can be extended to a large and clinically heterogeneous cohort [7,8]. Nevertheless, these studies were not designed to determine why local or region-specific models succeed or fail across sparsely represented target regions. Similarly, deep imbalanced regression methods, including label and feature distribution smoothing and ranking-based regularization, address continuous-target imbalance at the representation or loss-function level [17,39], but do not examine whether region-specific experts develop functional specialization or whether gating weights align with the locally best-performing expert. The present study therefore provides a complementary diagnostic perspective by separately evaluating input separability, expert specialization, gating alignment, and subject-wise robustness.

4.5. Practical Implications

Prior to deploying MoE architectures in wearable gait analysis applications, the structural preconditions for effective routing should be verified rather than assumed. Based on the present findings, at least four diagnostic evaluations are recommended before applying MoE in similar regression settings. First, the degree to which stride length zones are separable in the latent or feature space should be assessed. Second, the expert-only RMSE matrix should be examined for diagonal dominance. Third, true-zone-wise gating weight distributions should be evaluated to confirm alignment with expected or best-performing experts. Fourth, whether these alignment properties and associated performance gains are maintained under subject-independent LOSO evaluation should be verified.

When gating collapses toward a single expert such that entropy is low yet true-zone alignment is poor, or expert-level advantages fail to propagate to the final prediction, the introduction of MoE is likely to increase model complexity and subject-wise instability without commensurate performance improvement. The practical conclusion of this study is therefore not to abandon MoE as an architectural option, but to condition its application on the joint verification of input separability, expert specialization, gating alignment, and subject-wise robustness.

In wearable rehabilitation monitoring, consistent performance across individuals may be more important than isolated improvements within a specific stride-length region because unstable routing could produce inconsistent longitudinal estimates. Under the present experimental conditions, the single CNN therefore offers a practical advantage through its simpler architecture, lower overall prediction error, and greater subject-wise stability. MoE-based models should be considered only when their additional complexity produces robust and reproducible improvements under subject-independent evaluation.

This practical advantage was also supported by the supplementary computational analysis. The CNN showed the lowest measured mean inference time (26.23 ms per stride), whereas the MoE variants required 30.78–49.70 ms per stride. Moreover, the FC-AE-based models contained approximately 6.0 million trainable parameters compared with 197,505 parameters for the CNN. These measurements were interpreted descriptively as implementation-level computational indicators rather than as inferential hardware benchmarks (Supplementary Materials S2 and Table S3).

The present study was designed to evaluate the structural applicability of MoE routing rather than to establish clinical accuracy thresholds for stride length estimation. Therefore, the reported prediction errors should be interpreted as relative indicators of model performance, and their clinical acceptability requires validation in independent clinical populations.

4.6. Limitations and Future Directions

Several limitations of the present study should be acknowledged. First, the dataset comprised 22 male participants, including five older adults and 17 young adults. Although this sample permits a subject-independent feasibility evaluation, its modest size and male-only composition limit broad generalizations regarding the stability of complex MoE architectures. In addition, the cohort did not include individuals with clinically diagnosed gait disorders. Therefore, external validation using larger, sex-balanced, and clinically diverse cohorts is warranted [40,41]. A separate validation set or nested subject-wise cross-validation procedure was not used for architecture selection. The baseline CNN was selected from a limited set of predefined candidates based on preliminary LOSO performance on the same dataset; therefore, some degree of model-selection bias cannot be excluded.

Second, the Zone A/B/C boundaries of 125 cm and 135 cm were derived from the empirical distribution of the present dataset and approximately corresponded to its 25th and 75th percentiles. These boundaries were used as analytical partitions for investigating tail-region prediction and routing behavior and were not based on established clinical gait standards. Accordingly, conclusions associated with specific numerical thresholds should be interpreted cautiously. Future studies should examine whether the observed expert specialization and gating misalignment remain consistent across alternative zone definitions and independent datasets.

Third, the fixed-length sequence conversion using zero-padding preserves stride-duration-related temporal information but may introduce length-dependent confounds across individuals with different walking speeds or movement characteristics.

Fourth, the supplementary feature-ablation analysis did not demonstrate a statistically significant effect of removing the first- and second-order difference features. Therefore, the present results do not support a systematic increase in prediction error caused by residual-noise amplification from the difference operations. However, because removing these features simultaneously eliminates both temporal-transition information and any associated noise, the independent contribution of residual FSR noise could not be isolated.

Fifth, all MoE variants were evaluated using a predefined zone-classification gating objective. Therefore, the present study characterizes MoE behavior under this routing framework but does not establish whether alternative gating objectives would improve continuous regression performance. A rigorous comparison of classification-based, regression-aware, and load-balancing objectives would require separate loss-function development, auxiliary-loss coefficient selection, and subject-independent validation. These evaluations are therefore reserved for future work.

5. Conclusions

This study demonstrated that MoE-based expert routing does not universally outperform a single CNN in plantar-pressure-based long-tailed stride length estimation. Under the present dataset and design conditions, the CNN baseline achieved the most stable overall RMSE and subject-wise robustness, while no MoE variant consistently outperformed it across all evaluation criteria.

The central conclusion of this study is therefore not that MoE is generally ineffective, but that expert specialization alone is insufficient to guarantee improved final regression performance in MoE frameworks applied to continuous biomechanical regression. Target-space partitioning, zone classification, expert specialization, and continuous regression improvement are distinct conditions that should be separately verified. For MoE to be effectively deployed in wearable gait analysis, the joint satisfaction of input separability, expert specialization, gating alignment, and subject-wise stability must be empirically confirmed prior to deployment.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app16157476/s1>, Table S1, trial-pooled CNN performance with and without data augmentation; Table S2, subject-level paired comparison of CNN RMSE; Table S3, computational complexity and CPU-based inference cost.

Funding: This research received no external funding.

Institutional Review Board Statement: No formal ethical review was conducted because the data were collected in 2011 prior to the enforcement of mandatory Institutional Review approvals. The study was conducted in accordance with the ethical standards applicable at the time of data collection. Written informed consent was obtained from all participants before participation, and the study was conducted in accordance with the ethical standards applicable at the time of data collection.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. All participants voluntarily agreed to participate and provided written consent prior to data collection.

Data Availability Statement: The data presented in this study are available upon reasonable request from the corresponding author due to privacy and ethical considerations.

Acknowledgments: The author would like to thank all participants who volunteered for this study. During the preparation of this manuscript, the author used multiple versions of ChatGPT (OpenAI), including OpenAI o3 and GPT-5 series models up to GPT-5.5, for English language editing, manuscript structuring, assistance in organizing preliminary result interpretations, and generating graphical representations based on data provided and verified by the author. All statistical analyses were conducted and verified by the author. The author reviewed, edited, and verified all AI-assisted outputs and takes full responsibility for the accuracy, integrity, and content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AE	Autoencoder
BiLSTM	Bidirectional long short-term memory
CNN	Convolutional neural network
FC-AE	Fully connected autoencoder
FSR	Force-sensitive resistor
GAP	Global average pooling
LOSO	Leave-one-subject-out
LME	Linear mixed-effects model
MAE	Mean absolute error
ME	Mean error
MoE	Mixture-of-Experts
RMSE	Root mean square error

References

1. Whittle, M.W. *Gait Analysis: An Introduction*, 4th ed.; Butterworth-Heinemann: Oxford, UK, 2007.
2. Zijlstra, W.; Hof, A.L. Assessment of spatio-temporal gait parameters from trunk accelerations during human walking. *Gait Posture* **2003**, *18*, 1–10. [[CrossRef](#)] [[PubMed](#)]
3. Hollman, J.H.; McDade, E.M.; Petersen, R.C. Normative spatiotemporal gait parameters in older adults. *Gait Posture* **2011**, *34*, 111–118. [[CrossRef](#)] [[PubMed](#)]
4. Bytçi, I.; Henein, M.Y. Stride Length Predicts Adverse Clinical Events in Older Adults: A Systematic Review and Meta-Analysis. *J. Clin. Med.* **2021**, *10*, 2670. [[CrossRef](#)] [[PubMed](#)]
5. Muro-de-la-Herran, A.; Garcia-Zapirain, B.; Mendez-Zorrilla, A. Gait Analysis Methods: An Overview of Wearable and Non-Wearable Systems, Highlighting Clinical Applications. *Sensors* **2014**, *14*, 3362–3394. [[CrossRef](#)] [[PubMed](#)]
6. Tao, W.; Liu, T.; Zheng, R.; Feng, H. Gait analysis using wearable sensors. *Sensors* **2012**, *12*, 2255–2283. [[CrossRef](#)] [[PubMed](#)]
7. Zadka, A.; Rabin, N.; Gazit, E.; Mirelman, A.; Nieuwboer, A.; Rochester, L.; Del Din, S.; Pelosin, E.; Avanzino, L.; Bloem, B.R.; et al. A wearable sensor and machine learning estimate step length in older adults and patients with neurological disorders. *npj Digit. Med.* **2024**, *7*, 142. [[CrossRef](#)] [[PubMed](#)]
8. Hannink, J.; Kautz, T.; Pasluosta, C.F.; Barth, J.; Schüle, S.; Gaßmann, K.G.; Klucken, J.; Eskofier, B.M. Mobile stride length estimation with deep convolutional neural networks. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 354–362. [[CrossRef](#)] [[PubMed](#)]
9. Abdul Razak, A.H.; Zayegh, A.; Begg, R.K.; Wahab, Y. Foot Plantar Pressure Measurement System: A Review. *Sensors* **2012**, *12*, 9884–9912. [[CrossRef](#)]
10. Allet, L.; IJzerman, H.; Meijer, K.; Willems, P.; Savelberg, H. The influence of stride-length on plantar foot-pressures and joint moments. *Gait Posture* **2011**, *34*, 300–306. [[CrossRef](#)] [[PubMed](#)]
11. McGrath, R.L.; Ziegler, M.L.; Pires-Fernandes, M.; Knarr, B.A.; Higginson, J.S.; Sergi, F. The effect of stride length on lower extremity joint kinetics at various gait speeds. *PLoS ONE* **2019**, *14*, e0200862. [[CrossRef](#)] [[PubMed](#)]
12. Franco, A.; Russo, M.; Amboni, M.; Ponsiglione, A.M.; Di Filippo, F.; Romano, M.; Amato, F.; Ricciardi, C. The Role of Deep Learning and Gait Analysis in Parkinson's Disease: A Systematic Review. *Sensors* **2024**, *24*, 5957. [[CrossRef](#)] [[PubMed](#)]
13. Shalin, G.; Pardoel, S.; Lemaire, E.D.; Nantel, J.; Kofman, J. Prediction and detection of freezing of gait in Parkinson's disease from plantar pressure data using long short-term memory neural-networks. *J. Neuroeng. Rehabil.* **2021**, *18*, 167. [[CrossRef](#)] [[PubMed](#)]
14. Park, J.M.; Moon, C.W.; Lee, B.C.; Oh, E.; Lee, J.; Jang, W.J.; Cho, K.H.; Lee, S.H. Detection of freezing of gait in Parkinson's disease from foot-pressure sensing insoles using a temporal convolutional neural network. *Front. Aging Neurosci.* **2024**, *16*, 1437707. [[CrossRef](#)] [[PubMed](#)]
15. Zhou, R.; Liu, H.; Le, H.; Xiao, J.; Ma, Y.; Yuan, J.; Shan, W.; Wang, M. CSA-WTConvLSTM: A TEC Spatiotemporal Prediction Model Focusing on Both Low-Frequency and High-Frequency Features. *Space Weather* **2025**, *23*, e2025SW004338. [[CrossRef](#)]
16. Sun, L.; Liu, H.; Le, H.; Yuan, J.; Li, S.; Yao, X.; Yuan, G. ST-TFFM-Observer: A Global Ionospheric TEC Prediction Model Based on Dynamic System Theory with Enhanced Time-Frequency Dependence Modeling. *Space Weather* **2026**, *24*, e2025SW004876. [[CrossRef](#)]
17. Yang, Y.; Zha, K.; Chen, Y.C.; Wang, H.; Katabi, D. Delving into Deep Imbalanced Regression. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual, 18–24 July 2021; pp. 11842–11851.
18. Koo, H.; Kim, J.-Y.; Shim, S.R.; Lee, Y.; Shin, J. Understanding Bias in Stride Length Estimation: Toward a Gating-Based Adaptive Modeling Approach. In Proceedings of the 17th International Conference on Future Information & Communication Engineering (ICFICE 2025), Guam, GU, USA, 1–4 July 2025; Paper BI-03.
19. Torgo, L.; Ribeiro, R.P.; Pfahringer, B.; Branco, P. SMOTE for Regression. In *Progress in Artificial Intelligence, Proceedings of EPIA 2013*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 378–389.
20. Branco, P.; Torgo, L.; Ribeiro, R.P. SMOGN: A Pre-processing Approach for Imbalanced Regression. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*; PMLR: Cambridge, MA, USA, 2017; Volume 74, pp. 36–50.
21. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive mixtures of local experts. *Neural Comput.* **1991**, *3*, 79–87. [[CrossRef](#)] [[PubMed](#)]
22. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
23. Crenna, F.; Rossi, G.B.; Berardengo, M. Filtering biomechanical signals in movement analysis. *Sensors* **2021**, *21*, 4580. [[CrossRef](#)] [[PubMed](#)]
24. Melai, T.; IJzerman, T.H.; Schaper, N.C.; de Lange, T.L.H.; Willems, P.J.B.; Meijer, K.; Lieve, A.G.; Savelberg, H.H.C.M. Calculation of plantar pressure time integral, an alternative approach. *Gait Posture* **2011**, *34*, 379–383. [[CrossRef](#)] [[PubMed](#)]
25. Jlassi, O.; Dixon, P.C. The effect of time normalization and biomechanical signal processing techniques of ground reaction force curves on deep-learning model performance. *J. Biomech.* **2024**, *168*, 112116. [[CrossRef](#)] [[PubMed](#)]

26. Um, T.T.; Pfister, F.M.J.; Pichler, D.; Endo, S.; Lang, M.; Hirche, S.; Fietzek, U.; Kulić, D. Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI), Glasgow, UK, 13–17 November 2017; pp. 216–220.
27. Iwana, B.K.; Uchida, S. An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* **2021**, *16*, e0254841. [[CrossRef](#)] [[PubMed](#)]
28. Ordóñez, F.J.; Roggen, D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. [[CrossRef](#)] [[PubMed](#)]
29. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 448–456.
30. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
31. Lin, M.; Chen, Q.; Yan, S. Network in Network. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
32. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
33. Schuster, M.; Paliwal, K.K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681. [[CrossRef](#)]
34. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)] [[PubMed](#)]
35. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning (ICML), Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
36. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
37. Pinheiro, J.C.; Bates, D.M. *Mixed-Effects Models in S and S-PLUS*; Springer: New York, NY, USA, 2000.
38. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
39. Gong, Y.; Mori, G.; Tung, F. RankSim: Ranking Similarity Regularization for Deep Imbalanced Regression. In *Proceedings of the 39th International Conference on Machine Learning*; PMLR: Baltimore, MD, USA, 2022; Volume 162, pp. 7634–7649.
40. Lau, L.K.; Wee, S.L.; Pang, W.W.; Chen, K.K.; Jabbar, K.A.; Seah, W.T.; Ng, T.P. Reference Values of Gait Speed and Gait Spatiotemporal Parameters for a South East Asian Population: The Yishun Study. *Clin. Interv. Aging* **2020**, *15*, 1753–1765. [[CrossRef](#)] [[PubMed](#)]
41. Kawai, H.; Taniguchi, Y.; Seino, S.; Sakurai, R.; Osuka, Y.; Obuchi, S.; Watanabe, Y.; Kim, H.; Inagaki, H.; Kitamura, A.; et al. Reference values of gait parameters measured with a plantar pressure platform in community-dwelling older Japanese adults. *Clin. Interv. Aging* **2019**, *14*, 1265–1276. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.