

INERTIAL HEADING ESTIMATION FOR DOOR-MOUNTED IMUs VIA CNN-RNN MODELS

Ankit Kumar¹, Yadangi Abhishek¹, and Filippo Sanfilippo¹

¹Department of Engineering Sciences, University of Agder, 4879 Grimstad, Norway, e-mails: ankitk@uia.no, abhisheky@uia.no, filippo.sanfilippo@uia.no

KEYWORDS

Inertial measurement unit, heading estimation, sensor fusion, indoor navigation, deep learning.

ABSTRACT

Cheap IMUs struggle with indoor heading estimation because magnetic interference ruins magnetometer readings, causing classical fusion methods to fail. We explored whether deep learning could solve this by testing four CNN-RNN architectures (LSTM, BiLSTM, GRU, BiGRU) across all seven possible sensor combinations of accelerometer, gyroscope, and magnetometer data. Evaluated on the DoorINet dataset (Zakharchenko et al., 2024) using five error metrics, our seven-seed trial setup confirmed that gyroscopes are essential for low error. Yet, our key takeaway is that you shouldn't give up on the magnetometer entirely. When we combined gyroscope and magnetometer data in a CNN-BiLSTM, it significantly outperformed the gyroscope-only baseline ($p=0.024$), hitting a mean RMSE of just 0.2275 and an R^2 of 0.9703.

INTRODUCTION

Accurate indoor heading is a foundational requirement for pedestrian tracking, robotics, and smart buildings. Standard approaches fuse accelerometer, gyroscope, and magnetometer data to estimate heading. However, indoor magnetic distortions from steel and electronics can skew magnetometer readings by over 30%, causing severe drift in traditional fusion algorithms.

Deep learning bypasses this issue by learning patterns directly from raw data, removing the need for perfect magnetic models. Previous work like DoorINet (Zakharchenko et al., 2024) showed this works well, but simply ignored the magnetometer to avoid the noise entirely. We asked a different question: can an advanced temporal model actually extract useful heading information from a distorted magnetometer signal?

To find out, we systematically tested four CNN-RNN architectures across all seven possible sensor subsets (28 total configurations). We evaluated them using five distinct error metrics and ran seven randomised seeds per setup to guarantee statistical significance via Wilcoxon signed-rank tests.

Our key contributions are:

1. A complete factorial evaluation of four CNN-RNN models across all sensor subsets on the DoorINet dataset, validated with rigorous statistical testing.
2. Evidence that pairing gyroscope and magnetometer data in a CNN-BiLSTM significantly outperforms the gyroscope-only baseline ($p=0.024$), proving indoor magnetometer data is salvageable.
3. A breakdown of how bidirectional processing enables magnetometer fusion, including a clear characterisation of failure modes when gyroscope data is removed.

RELATED WORK

The problem of accurate heading and localisation using inertial sensors has been widely studied, evolving from classical sensor fusion techniques to modern learning-based approaches. In early work, multi-sensor fusion frameworks demonstrated how heterogeneous sensing modalities can be integrated to improve robustness and reliability in real-world applications (Sanfilippo and Pettersen, 2015). This direction was further extended to indoor positioning, where hybrid approaches combining wireless signals with inertial sensors highlighted the importance of orientation estimation while exposing the limitations of magnetometer-based heading under indoor magnetic disturbances (Ho-Sy et al., 2019). Subsequent studies on inertial-based localisation in robotic platforms emphasised challenges such as drift accumulation and sensitivity to noise when relying primarily on IMU data (Shahid et al., 2023). The literature has increasingly shifted toward data-driven methods for inertial navigation and heading estimation, particularly in scenarios where classical models struggle due to sensor imperfections and environmental disturbances.

Recent surveys by Chen and Pan (Chen and Pan, 2024) and Cohen and Klein (Cohen and Klein, 2024) highlight a growing consensus in the field: classical methods struggle indoors, prompting a major shift toward deep learning. Yet, before abandoning traditional techniques entirely, researchers have tried various ways to salvage the magnetometer. These approaches range from bypassing the

magnetic reference entirely using double-quaternion estimators (Wondosen et al., 2021), to dynamically filtering out corrupted samples via decision trees within a Kalman filter (Hu et al., 2020). Others have even argued that, with the right adaptive motion compensation, IMU-only setups can match magnetometer-aided accuracy outdoors (Ambuj et al., 2026).

On the learning side, neural networks have proven highly effective at handling sensor quirks. For example, convolutional networks can denoise gyroscope data to the point of matching visual-inertial benchmarks (Brossard et al., 2020), while multi-scale architectures with attention mechanisms excel at pedestrian odometry (Chen et al., 2022). The secret sauce in wearable sensor fusion often comes down to how the network weights different sensor channels (Qiu et al., 2022). Interestingly, even in deep learning, the magnetometer is often treated with caution; Yan et al. (Yan et al., 2025), for instance, rely on field consistency checks before allowing the network to use magnetometer updates.

Hybrid CNN–RNN models have become the workhorse for a variety of inertial tasks, including general attitude estimation (Golroudbari and Sabour, 2023), inertial odometry (Chen et al., 2021), and compensating for GNSS outages (Meng et al., 2024). Research in this space has shown that these models are highly sensitive to motion phases and noise (Bergamini et al., 2014), driving extensions into activity recognition (Kim et al., 2021), incremental learning (Narkhede et al., 2021), and attention-augmented variants (He et al., 2025; Ullah et al., 2025). Beyond core model design, the broader field is actively refining its foundations through better reproducibility protocols (Ahmed and Lofstead, 2022; Semmelrock et al., 2024), standardised error metrics (Botchkarev, 2018; Huber, 1964), automated IMU self-calibration (Cardoni et al., 2024), and specialised smartphone heading fusion (Ye et al., 2025).

Given the success of Transformers elsewhere, attention-based architectures are gaining traction in inertial positioning (Cohen and Klein, 2024; He et al., 2025), but typically demand larger datasets to outperform recurrent baselines. Since no attention model has been evaluated for door-mounted heading—and our goal is a controlled comparison with DoorINet’s BiGRU baseline—we retain CNN–RNN architectures, leaving attention-based extensions for future work. To date, no study has comprehensively explored how different deep learning architectures interact with all possible sensor combinations for heading estimation under real indoor magnetic distortion.

DATASET AND PREPROCESSING

Data Description

Experiments use the DoorINet dataset (Zakharchenko et al., 2024), which contains tri-axial accelerometer, gyroscope, and magnetometer recordings from Xsens DOT sensors at 120 Hz, with ground-truth heading from a Mem-sense MS-IMU3025 unit at 250 Hz. Sessions 1–3 cover

controlled door cycles at fixed angles, varying mount positions (handle, centre, hinge), and different opening speeds.

Three held-out sequences serve as test data: Test_5 (handle mount), Test_12 (hinge mount), and Test_14 (centre mount), each with distinct motion and noise characteristics. The full dataset spans 390 minutes, split into 119.1 minutes for training and validation and 271.9 minutes for testing.

Preprocessing

Gyroscope bias is corrected by identifying static windows of 30–50 frames during door-closed intervals using a low angular-velocity threshold on the gyroscope norm. The mean angular rate within this window is subtracted from the entire sequence as a first-order bias correction. Zero-drift alignment is then applied by forcing the estimated heading to zero during detected static intervals.

The reference heading $\psi(t)$ is obtained using a modified Madgwick filter. For each sliding window, the regression target is defined as the heading increment

$$\Delta\psi = \psi(t + w) - \psi(t), \quad (1)$$

which is wrapped to $[-180^\circ, 180^\circ]$ and encoded as

$$y = [\cos(\Delta\psi), \sin(\Delta\psi)]. \quad (2)$$

This unit-circle representation avoids discontinuities at angular wrap boundaries.

Inputs are constructed using a sliding window of $w = 40$ samples with unit stride, forming tensors

$$\mathbf{X} \in \mathbb{R}^{w \times d}, \quad (3)$$

where $d = 3k$ for k active sensors. All channels are standardised using training-set statistics,

$$\hat{x}_i = \frac{x_i - \mu_i}{\sigma_i}, \quad (4)$$

and the same parameters are applied to test data. The sine-cosine targets are not standardised. For test sequences, a thresholded Madgwick filter suspends updates during low-motion intervals to preserve zero-rate alignment.

Model Training and Evaluation

Each model–sensor configuration was trained under seven random seeds (35, 42, 123, 256, 512, 700, 999), and all reported results represent the mean and standard deviation across runs. Wilcoxon signed-rank tests are applied to the per-seed results across test sequences to assess statistical significance of key comparisons. Predictions are decoded using

$$\Delta\hat{\psi} = \text{atan2}(\sin \Delta\hat{\psi}, \cos \Delta\hat{\psi}), \quad (5)$$

and evaluated over the full test trajectory using five metrics: RMSE, MAE, MAD, LPD (terminal error), and R^2 .

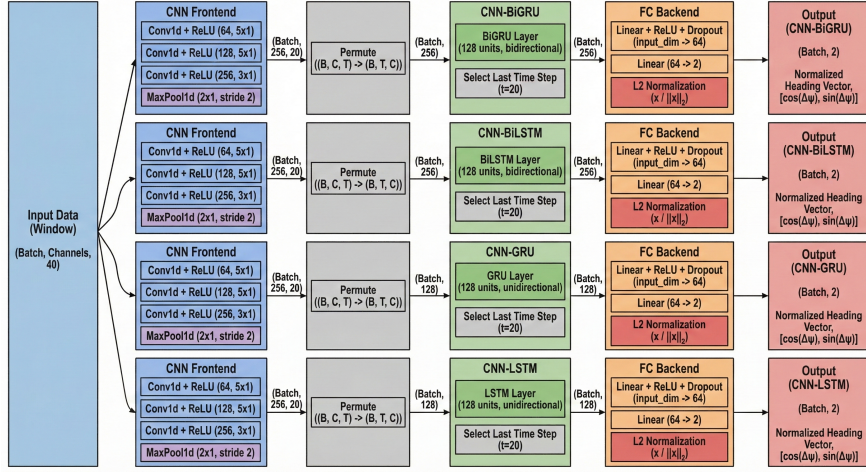


Figure 1: Architecture overview of the four CNN-RNN variants.

HEADING ESTIMATION APPROACHES

A simple way to estimate heading is to integrate the vertical angular rate from the gyroscope,

$$\psi(t) = \psi_0 + \int_0^t \omega_z(\tau) d\tau. \quad (6)$$

This method is easy to implement, but it does not correct for sensor bias or noise. As a result, small errors accumulate over time and lead to drift in the estimated heading.

The Madgwick filter (Zakharchenko et al., 2024) reduces this drift by representing orientation as a unit quaternion and applying a gradient-based correction at each time step. The correction aligns predicted gravity and magnetic field directions with accelerometer and magnetometer measurements. A gain parameter controls the balance between gyroscope integration and sensor feedback. While the method works well when the magnetic field is stable, its performance drops indoors where magnetometer readings are often disturbed.

DoorINet (Zakharchenko et al., 2024) proposed a learning-based solution using bidirectional GRU networks to predict heading increments from fixed-length IMU windows. The gyroscope-only version uses two stacked BiGRU layers followed by a pyramidal fully connected regression head with $\tanh$ activations. The accelerometer-gyroscope version processes each signal in separate BiGRU branches, combines their outputs, and applies an additional BiGRU layer before regression. Magnetometer data were not used due to magnetic disturbances. Among these variants, the accelerometer-gyroscope model achieved the best results and serves as the main baseline in this work.

Problem Formulation and Sensor Configurations

The models predict the heading increment $\Delta\psi$ from each sliding window. To handle the $\pm 180^\circ$ wraparound, the

target is encoded as $y = [\cos(\Delta\psi), \sin(\Delta\psi)]$, mapping the problem to $\mathbb{R}^2$ and avoiding angular discontinuities.

Seven configurations are formed from all non-empty subsets of accelerometer (A), gyroscope (G), and magnetometer (M), each contributing three channels, giving input dimensions of 3, 6, or 9. Among single-sensor inputs, the gyroscope performs best with the lowest RMSE and highest R^2 . The accelerometer performs poorly alone as it does not measure heading rotation directly, and the magnetometer gives the worst mean RMSE due to indoor distortions.

For two-sensor pairs, AG is consistently reliable across all architectures. AM is less stable due to indoor magnetic disturbance. GM works best under CNN-BiLSTM, which reaches its lowest test errors with this pairing. The three-sensor case (AGM) does not always outperform two-sensor setups as magnetometer noise can affect training, though it remains competitive under CNN-LSTM and CNN-BiLSTM.

All combinations use the same training settings: window size $w = 40$, MSE loss, Adam optimiser, and 50 epochs, so result differences reflect only architecture and sensor choice.

Architecture Overview

As shown in Fig. 1, all four models take windowed inputs of shape (Batch, Channels, 40). Each variant shares a common CNN frontend, a recurrent layer, and a fully connected regression head.

The CNN frontend is identical across all models and consists of three 1D convolutional layers with 64, 128, and 256 filters (kernel sizes 5, 5, and 3), each followed by ReLU activation. A max pooling layer (kernel size 2, stride 2) reduces the temporal dimension from 40 to 20 steps. Convolutions are applied across all input channels from the first layer, enabling early cross-sensor interaction. The CNN hyperparameters follow established practice in inertial deep learning (Brossard et al., 2020; Chen et al., 2022, 2021) and were held constant across all archi-

tures to isolate the effect of the recurrent module. The output tensor of shape (Batch, 256, 20) is then permuted to (Batch, 20, 256) for the recurrent layer.

The recurrent module is the only component that differs across variants. CNN–LSTM and CNN–GRU use unidirectional layers with 128 hidden units, retaining the final hidden state to give a 128-dimensional representation. CNN–BiLSTM and CNN–BiGRU use bidirectional layers with 128 units per direction; concatenating forward and backward states gives a 256-dimensional representation.

The regression head maps the recurrent output to 64 units via a linear layer with ReLU and dropout, then to two outputs (sine and cosine of $\Delta\psi$). L2 normalisation is applied to keep the prediction on the unit circle,

$$\hat{y}_{\text{norm}} = \frac{\hat{y}}{\|\hat{y}\|_2}, \quad (7)$$

giving the final representation $[\cos(\Delta\psi), \sin(\Delta\psi)]$ of shape (Batch, 2). Since the CNN frontend and regression head are fixed across models, result differences are due solely to the recurrent unit type and its directionality.

7. RESULTS SECTION (condensed from 5 subsections to 2 paragraphs)

RESULTS

Tables 1–4 report all 28 architecture–sensor configurations across four datasets (mean $\pm$ std over seven seeds). Gyroscope-inclusive inputs produce the lowest errors in every architecture and test sequence. On Test_14, CNN–BiLSTM (GM) reaches 0.1770° RMSE with $R^2 = 0.9822$, and CNN–LSTM (AGM) reaches 0.1852° . Under G and AG inputs, all four architectures stay below 0.31° RMSE on every test sequence. Accelerometer-only and magnetometer-only inputs exceed 1° RMSE under the more demanding conditions, with negative R^2 in several cases indicating prediction worse than a constant baseline.

Among two-sensor pairs, AG is the most reliable across all architectures. AM is unstable due to indoor magnetic disturbance, producing negative R^2 on Test_14 under CNN–LSTM (-6.79) and CNN–GRU (-5.39). CNN–BiLSTM (GM) achieves the best aggregate performance across all five metrics and all three test sequences, with mean test RMSE of 0.2275° and mean test R^2 of 0.9703 . Bidirectional architectures under gyroscope-inclusive inputs also record the smallest peak errors (MAD), with CNN–BiLSTM (GM) staying below 12° across all sequences. LPD values confirm that cumulative drift remains controlled over full test trajectories for all gyroscope-inclusive configurations.

DISCUSSION

Gyroscope data are essential. Removing gyroscope input raises RMSE by a factor of three to ten across all architectures. Wilcoxon signed-rank tests confirm this is statistically significant: CNN–BiLSTM (G) versus CNN–BiLSTM (A) yields $p < 0.001$, and CNN–BiLSTM (G)

versus CNN–BiLSTM (M) likewise yields $p < 0.001$. Accelerometer and magnetometer readings alone do not carry sufficient heading-rate information to generalise across mounting positions and motion profiles.

Indoor magnetometer data improve accuracy under bidirectional models. CNN–BiLSTM (GM) significantly outperforms CNN–BiLSTM (G) on RMSE (Wilcoxon $p = 0.024$) and R^2 ($p = 0.032$) across all test sequences. This advantage does not hold for unidirectional architectures: CNN–LSTM (GM) does not significantly differ from CNN–LSTM (G), suggesting that temporal smoothing of anomalous magnetometer readings specifically requires bidirectional context.

Why bidirectionality enables magnetometer fusion. A bidirectional recurrent layer processes each time step with access to both past and future context within the sliding window. When a transient magnetic spike occurs, the forward pass alone cannot distinguish it from a genuine heading change until subsequent samples arrive. The backward pass provides this future context, allowing the model to recognise short-lived anomalies and down-weight them implicitly. This mechanism acts as a learned soft anomaly filter. CNN–BiLSTM under GM benefits most because the LSTM’s gating structure retains finer-grained memory of magnetic field consistency across the window, while GRU’s simpler gate may discard this context prematurely. Indeed, CNN–BiLSTM (GM) significantly outperforms both CNN–LSTM (GM) ($p = 0.006$) and CNN–BiGRU (GM) ($p = 0.046$) on RMSE.

Failure cases. Two systematic failure modes emerge. First, accelerometer-only (A) inputs under CNN–BiLSTM collapse to near-constant predictions across several seeds, producing $R^2 \approx 0$ and identical metrics across test sequences (Table 2). This occurs because the accelerometer signal contains no direct heading-rate information, and the bidirectional architecture overfits to the training distribution rather than learning a generalisable mapping. Second, AM configurations exhibit extreme instability on Test_14, with R^2 as low as -10.59 under CNN–BiLSTM and -6.79 under CNN–LSTM, driven by the centre-mount position’s distinct magnetic signature interacting with accelerometer tilt coupling.

Broader applicability. Although evaluated on door-mounted sensors, the CNN–RNN framework is sensor-agnostic. The architecture takes fixed-length windows of triaxial IMU channels as input and predicts heading increments, a formulation that transfers directly to wearable IMUs, mobile robots, or handheld devices. The primary adaptation required is retraining on domain-specific data to account for different motion dynamics, sampling rates, and mounting orientations. The finding that bidirectional models can exploit noisy magnetometer data is particularly relevant for resource-constrained platforms where discarding an available sensor channel wastes information.

Limitations and generalisability. All experiments use a single dataset recorded with Xsens DOT sensors at 120 Hz in one building. Whether the GM advantage and the bidirectional benefit persist under different sen-

Table 1: Test performance of the CNN–LSTM architecture across sensor configurations

Mode	Case	RMSE (°)	MAE (°)	R^2	LPD (°)	MAD (°)
A	Test_12	1.0773 ± 0.3009	0.4771 ± 0.3943	0.3310 ± 0.3560	0.3668 ± 0.4091	35.2861 ± 8.0894
	Test_14	2.1673 ± 0.9886	1.5314 ± 1.2970	-1.9905 ± 2.4751	1.3228 ± 1.3174	31.3531 ± 6.0677
	Test_5	1.0971 ± 0.2146	0.3527 ± 0.2458	0.3283 ± 0.2625	0.2554 ± 0.2911	31.4412 ± 5.5635
G	Test_12	0.2402 ± 0.0346	0.1349 ± 0.0647	0.9683 ± 0.0097	0.1122 ± 0.0672	8.6771 ± 0.6143
	Test_14	0.1923 ± 0.0363	0.1223 ± 0.0592	0.9794 ± 0.0084	0.1062 ± 0.0616	7.2558 ± 0.6693
	Test_5	0.2870 ± 0.0270	0.1376 ± 0.0608	0.9552 ± 0.0088	0.1086 ± 0.0634	10.9647 ± 0.5698
M	Test_12	0.8266 ± 0.4055	0.4519 ± 0.2341	0.5546 ± 0.3999	0.3872 ± 0.2844	15.9128 ± 5.8607
	Test_14	1.0232 ± 0.4436	0.6812 ± 0.4739	0.3432 ± 0.4165	0.6227 ± 0.5211	23.9282 ± 2.2337
	Test_5	0.9307 ± 0.3963	0.5689 ± 0.3586	0.4593 ± 0.3805	0.5002 ± 0.4107	16.2369 ± 5.7342
AG	Test_12	0.2216 ± 0.0238	0.0899 ± 0.0525	0.9732 ± 0.0059	0.0700 ± 0.0535	9.0545 ± 0.7107
	Test_14	0.1734 ± 0.0284	0.0827 ± 0.0530	0.9834 ± 0.0058	0.0696 ± 0.0541	7.7051 ± 0.6756
	Test_5	0.2740 ± 0.0207	0.0960 ± 0.0526	0.9592 ± 0.0063	0.0699 ± 0.0539	11.3288 ± 0.6996
AM	Test_12	0.6112 ± 0.4048	0.4600 ± 0.4755	0.7223 ± 0.4160	0.4209 ± 0.4917	10.9318 ± 2.5353
	Test_14	2.8404 ± 2.7228	2.5268 ± 2.8030	-6.7924 ± 12.9487	2.4461 ± 2.8487	22.8676 ± 7.6959
	Test_5	0.4295 ± 0.0669	0.2533 ± 0.0782	0.8983 ± 0.0347	0.2050 ± 0.0956	12.6951 ± 2.1280
GM	Test_12	0.2597 ± 0.0761	0.1674 ± 0.1082	0.9609 ± 0.0264	0.1487 ± 0.1086	9.2185 ± 0.5445
	Test_14	0.2166 ± 0.0760	0.1550 ± 0.1048	0.9720 ± 0.0218	0.1422 ± 0.1048	7.8849 ± 0.6506
	Test_5	0.3008 ± 0.0705	0.1525 ± 0.1236	0.9488 ± 0.0263	0.1256 ± 0.1250	11.3337 ± 0.4947
AGM	Test_12	0.2263 ± 0.0364	0.1135 ± 0.0749	0.9717 ± 0.0092	0.0918 ± 0.0777	8.4103 ± 0.6944
	Test_14	0.1852 ± 0.0465	0.1114 ± 0.0752	0.9805 ± 0.0107	0.0972 ± 0.0785	6.9560 ± 0.6833
	Test_5	0.2724 ± 0.0191	0.1248 ± 0.0458	0.9597 ± 0.0056	0.0985 ± 0.0460	10.6040 ± 0.7070

Table 2: Test performance of the CNN–BiLSTM architecture across sensor configurations

Mode	Case	RMSE (°)	MAE (°)	R^2	LPD (°)	MAD (°)
A	Test_12	1.1955 ± 0.3780	0.6674 ± 0.5506	0.1617 ± 0.5369	0.5739 ± 0.5684	32.9785 ± 6.6694
	Test_14	3.7690 ± 2.9089	3.2927 ± 3.2726	-10.5937 ± 16.4114	3.1331 ± 3.3514	29.0895 ± 5.5005
	Test_5	1.0919 ± 0.2070	0.4264 ± 0.2077	0.3360 ± 0.2558	0.3099 ± 0.2343	29.1392 ± 5.2779
G	Test_12	0.2549 ± 0.0352	0.1480 ± 0.0694	0.9643 ± 0.0098	0.1262 ± 0.0739	9.0169 ± 0.4615
	Test_14	0.2175 ± 0.0424	0.1515 ± 0.0618	0.9736 ± 0.0100	0.1366 ± 0.0654	7.7239 ± 0.5617
	Test_5	0.3031 ± 0.0293	0.1603 ± 0.0635	0.9500 ± 0.0097	0.1329 ± 0.0675	11.2428 ± 0.5916
M	Test_12	0.9628 ± 0.4988	0.2347 ± 0.0651	0.3839 ± 0.4819	0.1174 ± 0.0948	19.9274 ± 5.7840
	Test_14	1.3666 ± 0.5390	0.6645 ± 0.8013	-0.1436 ± 0.7515	0.5478 ± 0.8501	26.1758 ± 3.9979
	Test_5	1.0393 ± 0.4278	0.3365 ± 0.2427	0.3316 ± 0.4277	0.2241 ± 0.2963	20.3273 ± 5.2855
AG	Test_12	0.2978 ± 0.1048	0.2108 ± 0.1467	0.9470 ± 0.0391	0.1889 ± 0.1502	8.9108 ± 0.7501
	Test_14	0.2591 ± 0.1099	0.2024 ± 0.1424	0.9581 ± 0.0355	0.1868 ± 0.1458	7.4833 ± 0.7279
	Test_5	0.3397 ± 0.0900	0.2156 ± 0.1436	0.9339 ± 0.0367	0.1874 ± 0.1473	11.1597 ± 0.7531
AM	Test_12	1.2133 ± 0.3882	0.3859 ± 0.4034	0.1348 ± 0.3554	0.2576 ± 0.4553	20.4490 ± 7.0058
	Test_14	1.2064 ± 0.2729	0.3207 ± 0.1266	0.1791 ± 0.3175	0.1805 ± 0.1765	26.4729 ± 3.4303
	Test_5	1.1152 ± 0.4272	0.2548 ± 0.0789	0.2435 ± 0.4256	0.1161 ± 0.1054	21.2869 ± 5.5446
GM	Test_12	0.2301 ± 0.0488	0.1081 ± 0.0818	0.9703 ± 0.0144	0.0876 ± 0.0838	9.4748 ± 1.0221
	Test_14	0.1770 ± 0.0438	0.0879 ± 0.0735	0.9822 ± 0.0096	0.0726 ± 0.0750	8.0142 ± 0.9950
	Test_5	0.2753 ± 0.0389	0.0957 ± 0.0771	0.9583 ± 0.0127	0.0667 ± 0.0792	11.7286 ± 1.0102
AGM	Test_12	0.2191 ± 0.0293	0.1070 ± 0.0702	0.9737 ± 0.0073	0.0866 ± 0.0745	8.0854 ± 0.8629
	Test_14	0.1831 ± 0.0472	0.1134 ± 0.0782	0.9809 ± 0.0103	0.1008 ± 0.0810	6.6479 ± 0.7494
	Test_5	0.2791 ± 0.0296	0.1360 ± 0.0773	0.9575 ± 0.0093	0.1100 ± 0.0814	10.3371 ± 0.9991

Table 3: Test performance of the CNN–GRU architecture across sensor configurations

Mode	Case	RMSE (°)	MAE (°)	R^2	LPD (°)	MAD (°)
A	Test_12	1.0885 ± 0.2882	0.5018 ± 0.4167	0.3213 ± 0.3380	0.3924 ± 0.4588	33.9202 ± 6.7313
	Test_14	1.7929 ± 0.6364	1.1245 ± 0.8695	-0.9244 ± 1.3858	0.8065 ± 0.9689	31.5615 ± 5.0460
	Test_5	1.2691 ± 0.4229	0.6091 ± 0.6691	0.0469 ± 0.6432	0.5064 ± 0.7121	31.1206 ± 5.1940
G	Test_12	0.2609 ± 0.0534	0.1454 ± 0.0898	0.9619 ± 0.0162	0.1232 ± 0.0899	12.5753 ± 1.6397
	Test_14	0.2191 ± 0.0598	0.1375 ± 0.0881	0.9724 ± 0.0157	0.1218 ± 0.0877	11.4716 ± 1.7541
	Test_5	0.3063 ± 0.0464	0.1504 ± 0.0882	0.9483 ± 0.0163	0.1222 ± 0.0882	14.4423 ± 1.5930
M	Test_12	0.9358 ± 0.5536	0.5268 ± 0.5416	0.3849 ± 0.5670	0.4611 ± 0.5787	15.2555 ± 6.4671
	Test_14	1.3717 ± 1.3126	0.9565 ± 1.4083	-0.8146 ± 3.3939	0.8770 ± 1.4328	24.8691 ± 6.3328
	Test_5	0.8703 ± 0.4373	0.4553 ± 0.3957	0.5022 ± 0.4019	0.3892 ± 0.4303	16.3339 ± 5.9171
AG	Test_12	0.2635 ± 0.0322	0.1597 ± 0.0653	0.9620 ± 0.0097	0.1396 ± 0.0670	13.4161 ± 1.2108
	Test_14	0.2290 ± 0.0379	0.1566 ± 0.0717	0.9710 ± 0.0094	0.1417 ± 0.0760	12.3694 ± 1.1564
	Test_5	0.3101 ± 0.0263	0.1672 ± 0.0683	0.9477 ± 0.0090	0.1406 ± 0.0709	15.1543 ± 1.2500
AM	Test_12	0.8069 ± 0.4418	0.4432 ± 0.2665	0.5578 ± 0.4150	0.3721 ± 0.3031	14.6536 ± 7.2609
	Test_14	2.1764 ± 2.8757	1.6864 ± 2.9937	-5.3889 ± 15.0352	1.5674 ± 3.0429	25.2714 ± 5.0042
	Test_5	0.8620 ± 0.4520	0.4604 ± 0.3654	0.5039 ± 0.4278	0.3916 ± 0.4023	17.0188 ± 5.4552
GM	Test_12	0.2959 ± 0.0579	0.2032 ± 0.0881	0.9511 ± 0.0194	0.1824 ± 0.0863	14.7235 ± 1.2083
	Test_14	0.2598 ± 0.0668	0.1979 ± 0.0874	0.9615 ± 0.0212	0.1837 ± 0.0863	13.6536 ± 1.1758
	Test_5	0.3382 ± 0.0469	0.2097 ± 0.0834	0.9372 ± 0.0182	0.1830 ± 0.0821	16.3755 ± 1.2357
AGM	Test_12	0.2464 ± 0.0241	0.1124 ± 0.0529	0.9669 ± 0.0067	0.0914 ± 0.0559	15.7305 ± 0.9281
	Test_14	0.2080 ± 0.0320	0.1175 ± 0.0558	0.9762 ± 0.0078	0.1037 ± 0.0587	14.7216 ± 1.0622
	Test_5	0.3052 ± 0.0247	0.1440 ± 0.0540	0.9494 ± 0.0084	0.1181 ± 0.0572	17.2882 ± 0.9038

Table 4: Test performance of the CNN–BiGRU architecture across sensor configurations

Mode	Case	RMSE (°)	MAE (°)	R^2	LPD (°)	MAD (°)
A	Test_12	1.3215 ± 0.1254	0.3522 ± 0.1969	0.0492 ± 0.1635	0.1967 ± 0.2385	26.3306 ± 4.7986
	Test_14	1.9041 ± 1.4167	0.9597 ± 1.7823	−1.8883 ± 4.9639	0.7976 ± 1.8104	25.7553 ± 3.2775
	Test_5	1.3299 ± 0.1034	0.3416 ± 0.1728	0.0395 ± 0.1380	0.1855 ± 0.2119	25.7917 ± 3.3738
G	Test_12	0.2959 ± 0.0503	0.2048 ± 0.0868	0.9515 ± 0.0156	0.1847 ± 0.0892	13.5098 ± 2.2980
	Test_14	0.2516 ± 0.0530	0.1862 ± 0.0843	0.9645 ± 0.0139	0.1723 ± 0.0869	12.4801 ± 2.4336
	Test_5	0.3335 ± 0.0435	0.2032 ± 0.0859	0.9390 ± 0.0154	0.1765 ± 0.0891	15.3170 ± 2.1525
M	Test_12	1.0739 ± 0.3977	0.5621 ± 0.4240	0.3037 ± 0.4103	0.4626 ± 0.4928	16.8529 ± 7.4099
	Test_14	0.9864 ± 0.4486	0.4148 ± 0.3644	0.3811 ± 0.4441	0.3193 ± 0.4055	29.0729 ± 4.8069
	Test_5	1.0014 ± 0.5710	0.4394 ± 0.5371	0.3072 ± 0.5925	0.3370 ± 0.5761	17.9728 ± 6.3419
AG	Test_12	0.2415 ± 0.0272	0.1272 ± 0.0439	0.9681 ± 0.0075	0.1058 ± 0.0419	12.4162 ± 1.9004
	Test_14	0.1978 ± 0.0315	0.1165 ± 0.0494	0.9784 ± 0.0075	0.1018 ± 0.0483	11.3422 ± 1.8539
	Test_5	0.2894 ± 0.0246	0.1310 ± 0.0462	0.9545 ± 0.0080	0.1034 ± 0.0447	14.2794 ± 1.8395
AM	Test_12	0.8986 ± 0.4784	0.8353 ± 0.5156	0.4577 ± 0.4826	0.8221 ± 0.5209	10.7994 ± 1.0226
	Test_14	1.4492 ± 0.6926	1.2726 ± 0.7661	−0.3568 ± 1.2489	1.1807 ± 0.7596	23.7015 ± 9.1282
	Test_5	0.6645 ± 0.2507	0.5415 ± 0.3213	0.7323 ± 0.2008	0.5192 ± 0.3302	13.6940 ± 1.6425
GM	Test_12	0.2438 ± 0.0234	0.1299 ± 0.0446	0.9676 ± 0.0062	0.1095 ± 0.0447	11.8437 ± 1.2955
	Test_14	0.2058 ± 0.0302	0.1248 ± 0.0543	0.9767 ± 0.0067	0.1104 ± 0.0556	10.7828 ± 1.3111
	Test_5	0.3014 ± 0.0369	0.1490 ± 0.0735	0.9503 ± 0.0125	0.1227 ± 0.0747	13.5854 ± 1.2231
AGM	Test_12	0.3113 ± 0.1646	0.1868 ± 0.2164	0.9351 ± 0.0789	0.1672 ± 0.2185	14.8543 ± 3.2096
	Test_14	0.2735 ± 0.1811	0.1880 ± 0.2164	0.9444 ± 0.0821	0.1739 ± 0.2189	13.6960 ± 3.1921
	Test_5	0.3613 ± 0.1536	0.2111 ± 0.2095	0.9185 ± 0.0815	0.1861 ± 0.2115	16.5625 ± 3.1283

sensor platforms (e.g. smartphone IMUs with higher noise floors), different sampling rates, or different magnetic environments (e.g. industrial facilities with stronger distortion) remains open. Cross-dataset validation and transfer learning experiments are natural next steps. Additionally, the current study compares recurrent architectures only; extending the comparison to include Transformer-based and temporal convolutional models would clarify whether the bidirectional magnetometer fusion effect is specific to LSTM gating or a more general property of models with future-context access.

CONCLUSION

Four CNN–RNN architectures were evaluated under seven sensor-input configurations using five metrics on three held-out door-motion sequences, with seven training seeds and Wilcoxon signed-rank tests for statistical validation.

Gyroscope data are necessary. Removing angular rate information raises RMSE by a factor of three to ten regardless of architecture ($p < 0.001$); no other sensor subset recovers comparable accuracy.

CNN–BiLSTM with the GM input produces the best overall results: mean test RMSE of 0.2275° , mean R^2 of 0.9703, and on Test_14 specifically 0.1770° RMSE with $R^2 = 0.9822$. This configuration significantly outperforms the gyroscope-only baseline ($p = 0.024$) and the unidirectional CNN–LSTM under the same GM input ($p = 0.006$), establishing bidirectional temporal modelling as the deciding factor.

Indoor magnetometer channels need not be discarded. Gyroscope and magnetometer fusion under a bidirectional LSTM lowers prediction error below the gyroscope-only baseline with statistical significance. The common practice of ignoring magnetometer readings in indoor inertial estimation is unnecessarily conservative when a temporally expressive model is available.

ACKNOWLEDGEMENT

This research is supported by the Artificial Intelligence, Biomechatronics, and Collaborative Robotics research group, University of Agder (UiA), Norway.

References

- H. Ahmed and J. Lofstead. Managing randomness to enable reproducible machine learning. In *Proceedings of P-RECS*, 2022.
- Ambuj, R. Machavaram, and A. Majhi. Enhancement of mobile robot performance on inclined agricultural terrains through imu-based yaw and heading angle estimation. *Measurement*, 258, 2026.
- E. Bergamini et al. Estimating orientation using magnetic and inertial sensors. *Sensors*, 14, 2014.
- A. Botchkarev. Performance metrics in machine learning regression and forecasting. arXiv:1809.03006, 2018.
- M. Brossard, S. Bonnabel, and A. Barrau. Denoising imu gyroscopes with deep learning for open-loop attitude estimation. *IEEE Robotics and Automation Letters*, 5 (3), 2020.
- M. Cardoni et al. Imu self-calibration via quantization-aware neural networks. *Electronics*, 13, 2024.
- B. Chen et al. Deep-learning-based inertial odometry for pedestrian tracking using attention mechanism and res2net module. *IEEE Sensors Letters*, 6(11), 2022.
- C. Chen and X. Pan. Deep learning for inertial positioning: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 25(9):10506–10523, 2024.
- C. Chen et al. Deep neural network-based inertial odometry using low-cost imus. *IEEE Transactions on Mobile Computing*, 20(4), 2021.

N. Cohen and I. Klein. Inertial navigation meets deep learning: A survey of current trends and future directions. *Results in Engineering*, 24, 2024.

A. A. Golroudbari and M. H. Sabour. Generalizable end-to-end deep learning frameworks for real-time attitude estimation. *Measurement*, 217, 2023.

J.-L. He et al. Human activity recognition via attention-augmented tcn-bigru fusion. *Sensors*, 25, 2025.

Thong Ho-Sy, Filippo Sanfilippo, and Vinh Truong-Quang. A hybrid algorithm based on wifi for robust and effective indoor positioning. In *Proc. of the 19th IEEE International Symposium on Communications and Information Technologies (ISCIT)*, pages 416–421, 2019.

G. Hu, W. Zhang, H. Wan, and X. Li. Improving the heading accuracy in indoor pedestrian navigation based on a decision tree and kalman filter. *Sensors*, 20, 2020.

P. J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101, 1964.

Y. W. Kim et al. Wearable imu-based activity recognition using 1d-cnn and gru ensemble. *Sensors*, 21, 2021.

X. Meng et al. A gnss/ins integrated navigation compensation method based on cnn-gru + irakf. *IEEE Transactions on Instrumentation and Measurement*, 73, 2024.

P. Narkhede et al. Incremental learning of lstm framework for sensor fusion in attitude estimation. *PeerJ Computer Science*, 7, 2021.

S. Qiu et al. Multi-sensor information fusion based on machine learning for human activity recognition. *Information Fusion*, 80, 2022.

Filippo Sanfilippo and Kristin Y Pettersen. A sensor fusion wearable health-monitoring system with haptic feedback. In *Proc. of the 11th IEEE International conference on innovations in information technology (IIT)*, pages 262–266, 2015.

H. Semmelrock et al. Reproducibility in machine learning-based research: Barriers and drivers. arXiv:2406.14325, 2024.

Ramsha Shahid, Anam Baloch, Humayun Tahir, and Anayat Ullah. Odometry and inertial sensor-based localization of a snake robot. In *Proc. of the IEEE International Conference on Robotics and Automation in Industry (ICRAI)*, pages 1–6, 2023.

K. Ullah et al. Hybrid cnn-bigru model with attention for short-term load forecasting. *Energy Reports*, 14, 2025.

A. Wondosen et al. Improved attitude and heading accuracy with double quaternion parameters estimation and magnetic disturbance rejection. *Sensors*, 21(16), 2021.

S. Yan et al. A heading angle optimization algorithm based on inertial sensor and magnetometer. In *Proceedings of IPIN*, 2025.

J. Ye et al. Enhancing real-time heading estimation via deep learning and smartphone sensors. *Scientific Reports*, 15, 2025.

A. Zakharchenko, S. Farber, and I. Klein. Doorinet: Door heading prediction through inertial deep learning. *IEEE Sensors Journal*, 24(19):31456–31465, October 2024. doi: 10.1109/JSEN.2024.3442812.

AUTHOR BIOGRAPHIES



ANKIT KUMAR received the Master's degree in Smart Manufacturing from the Indian Institute of Information Technology, Design and Manufacturing (IIITDM), Kancheepuram, India, in 2025. He is currently a Visiting Researcher in the Department of Mechatronics at the University of Agder (UiA), Norway. His research interests include inertial sensing, machine learning, and learning-based state estimation.



YADANGI ABHISHEK received the B.Tech. degree in Mechanical Engineering from Vardhaman College of Engineering, Hyderabad, India, in 2024. He is currently pursuing the M.Tech. degree in Smart Manufacturing at the Indian Institute of Information Technology Design and Manufacturing (IIITDM), Kancheepuram, India. He is participating in a one-year academic exchange program at the University of Agder, Grimstad, Norway.



FILIPPO SANFILIPPO holds a PhD in Engineering Cybernetics from the Norwegian University of Science and Technology (NTNU), Norway. He is a Professor at the University of Agder (UiA), Norway, and heads the Artificial Intelligence, Biomechanics, and Collaborative Robotics research group. His research focuses on Human-Robot Teaming (HRT). He is an IEEE Senior Member and current Chair of the IEEE Robotics and Automation, Control Systems, and Intelligent Transportation Systems Joint Chapter.