Neuro-Symbolic Fall Risk Detection from Wearable IMU Sensors: A Concept Bottleneck and Rule-Based Reasoning Approach

Louis Mono[1,*], Anna-Katharina Herke[1] and Redon Jashari[1]

[1]Department of Applied Artificial Intelligence, Alma Mater Europaea University, Maribor, Slovenia

Abstract

Falls are a major source of injury and mortality in older adults, yet many AI-based fall-risk systems remain black-box classifiers that provide no clinical justification for their predictions. We present a neuro-symbolic pipeline for interpretable classification of historical faller status from wearable inertial measurement unit (IMU) gait signals. A 1D convolutional neural network maps raw accelerometer and gyroscope signals to six clinically interpretable concepts: Timed Up-and-Go (TUG), gait speed, Berg Balance, Four Square Step Test, Activities-specific Balance Confidence, and Dynamic Gait Index. These concepts feed both a neural classifier and a symbolic rule engine based on clinical cut-points. Evaluated with Leave-One-Subject-Out cross-validation on 69 LTMM participants, the final 6-concept ensemble achieves 71.0% accuracy and AUC 0.763, compared with 73.9% accuracy and AUC 0.880 for a black-box CNN. The observed 11.7 AUC-point difference quantifies a performance–interpretability trade-off while the concept model provides predicted clinical quantities, rule firings, and auditable explanation traces.

Keywords

neuro-symbolic AI, concept bottleneck models, fall risk detection, wearable sensors, explainable AI, clinical decision support

1. Introduction

Falls among older adults are a major public-health burden, with the World Health Organization estimating hundreds of thousands of fatal falls annually [1]. Wearable inertial measurement units (IMUs) provide a scalable way to assess gait and mobility, and deep neural networks can learn discriminative patterns directly from raw sensor signals [2, 3]. However, high-performing fall-risk models often remain clinically opaque: they output a probability without explaining which mobility or balance deficit contributed to the decision. This limits trust, auditability, and clinical uptake [4, 5]. Recent IMU work has also used attention weights to expose model focus [6]; such explanations are useful but do not directly provide predicted clinical measurements and executable rule traces.

Clinicians assess interpretable constructs such as TUG time, gait speed, balance, stepping ability, and confidence, then combine these observations with clinical cut-points. This work asks whether a neural model can retain useful classification performance while exposing its reasoning through clinically meaningful intermediate concepts. We investigate a concept bottleneck model (CBM) [7] integrated with symbolic fall-risk rules, framing the task as a neuro-symbolic bridge between neural gait perception and clinical reasoning [8].

AIxIA 2026 Discussion Papers, 24th International Conference of the Italian Association for Artificial Intelligence, Perugia, Italy, 2026

*Corresponding author.
louis.mono@almamater.si (L. Mono); Anna-katharina.herke@almamater.si (A. Herke); redon.jashari@almamater.si (R. Jashari)
0009-0005-8110-0787 (L. Mono)

2. Method

2.1. Data and Evaluation

The PhysioNet LTMM cohort contains lower-back IMU gait recordings and clinical annotations from 71 community-dwelling older adults [9, 10]. The present analysis uses 69 participant-level LabWalk recordings after study-specific quality filtering: 31 participants reporting at least two falls in the preceding year and 38 reporting fewer than two. The evaluated outcome is therefore retrospective faller-status classification, not prospective fall prediction. Each input is a six-channel IMU signal (three accelerometer and three gyroscope axes, 100 Hz). Transfer learning used SisFall [11] for fall/ADL pre-training and a 2025 clinical gait dataset [12] for gait-quality pre-training.

Classification is evaluated using Leave-One-Subject-Out cross-validation. We report accuracy, AUC, sensitivity, and specificity. AUC is the primary comparison metric because it is threshold-independent. To reduce training stochasticity in the small-sample setting, the final model averages predictions across three random seeds. AUC uncertainty was estimated using percentile bootstrap 95% confidence intervals over the out-of-fold predictions (2,000 resamples).

2.2. Neuro-Symbolic Pipeline

The architecture has four stages:

1. a 1D-CNN backbone with convolutional blocks of 32, 64, and 128 filters extracts a gait representation from raw IMU signals;
2. a concept bottleneck predicts six clinical concepts from this representation;
3. a neural classifier maps the predicted concepts to fall-risk probability;
4. a symbolic rule engine evaluates the same concepts against clinical cut-points and produces a symbolic risk score and rule trace.

The fused score is

$$r = \alpha P_{\text{neural}} + (1 - \alpha) S_{\text{symbolic}}, \tag{1}$$

where P_{neural} is the neural probability and S_{symbolic} is the rule-derived score. The reported optimized fused configuration uses $\alpha = 0.75$. Because this weight was selected post hoc on the same evaluation data, the associated performance is exploratory and may be optimistic. The six-concept ensemble is emphasized for stability across three seeds, not as an unbiased estimate of deployment performance.

2.3. Clinical Concepts and Rules

The six concepts were selected from the available clinical annotations with guidance from the clinical author. They include TUG, gait speed, Berg Balance, Four Square Step Test (FSST), Activities-specific Balance Confidence (ABC), and Dynamic Gait Index (DGI). The present implementation applies TUG > 13.5 s [13], gait speed < 0.8 m/s [14], Berg Balance < 50 and DGI < 19 [15], and ABC $< 67\%$ [16]. Combination rules encode multi-factorial reasoning, for example jointly impaired TUG and Berg Balance.

3. Results

Table 1 summarizes the central trade-off. The black-box CNN reaches the highest AUC (0.880; 95% CI [0.791, 0.947]), but provides no concept-level rationale. The interpretable 6-concept ensemble achieves 71.0% accuracy and AUC 0.763 (95% CI [0.646, 0.863]), an observed 11.7 AUC-point difference from the CNN. The width of these intervals reflects the uncertainty associated with the small cohort. Its sensitivity and specificity are balanced (70.9% and 71.1%), whereas the CNN is sensitivity-heavy (90.3% sensitivity, 60.5% specificity).

Table 1
LOSO classification results on the LTMM fall-risk task. Values are percentages except AUC.

Model	Interp.	Acc.	AUC	Sens.	Spec.
SVM-RBF baseline	–	65.2	0.620	64.5	65.8
CNN black-box	No	73.9	**0.880**	**90.3**	60.5
CBM 8c fused	Yes	**73.9**	0.794	67.7	**78.9**
CBM 6c ensemble + rules	Yes	71.0	0.763	70.9	71.1

The concept bottleneck is clinically meaningful: TUG and gait speed are predicted from raw IMU signals with Pearson correlations of $r = 0.822$ and $r = 0.806$, respectively. DGI ($r = 0.761$) and Berg Balance ($r = 0.715$) are also learnable from motion data. The more subjective ABC measure is weaker ($r = 0.590$). This hierarchy is clinically plausible: biomechanical concepts are more directly visible in gait dynamics than self-reported confidence.

Transfer learning partially closes the interpretability gap. Training the CBM from scratch gives AUC 0.742, while pre-training on the 2025 clinical gait dataset raises AUC to 0.800, a gain of 5.8 AUC points. In contrast, progressive transfer through SisFall gives AUC 0.763, consistent with possible negative transfer from fall/ADL events to subtle gait-quality patterns. These point estimates suggest that source-task similarity may matter more than source-dataset size.

Clinician-guided ablation reduced the original 8 concepts to 6 by removing SF-36 Physical and GDS. Dropping SF-36 reduced AUC from 0.765 to 0.734, but subsequently removing GDS improved AUC to 0.768; removing DGI then reduced AUC to 0.742. These point estimates support DGI as a classification-relevant dynamic gait concept.

4. Interpretability

Every prediction can be accompanied by a trace grounded in concepts and rules. A representative high-risk participant produced the following condensed explanation:

```
Prediction: FALLER (High Risk)
Fused: 0.962 | Neural: 0.949 | Symbolic: 1.000
Concepts: TUG=27.00s, Speed=0.25m/s,
Berg=38.44/56, FSST=27.86s, ABC=37.44%,
DGI=12.56/24
Rules: TUG>13.5, Berg<50, Speed<0.8,
DGI<19, ABC<67, TUG+Berg jointly impaired,
Speed+FSST jointly impaired
```

This trace changes the role of the model output. Rather than asking clinicians to trust an opaque score, the system exposes predicted clinical quantities and the rules that fired. A clinician can inspect, contest, or override the prediction by examining the concept values. This is a practical advantage over post-hoc saliency methods, which typically explain input regions rather than an explicit clinical reasoning chain.

5. Discussion and Conclusion

The contribution is methodological rather than a claim of state-of-the-art accuracy. The study shows how a faller-status classifier can be constrained through clinically meaningful concepts and symbolic rules, making the observed performance–interpretability trade-off explicit. On this cohort, the six-concept model trails the CNN by 11.7 AUC points while providing auditable concept predictions and rule-based explanations. The transfer experiments further suggest that pre-training on a clinically similar gait task may be more useful than transfer from a larger but less aligned fall/ADL task.

The evidence remains preliminary. Limitations include the small retrospective cohort, a single-dataset evaluation, wide uncertainty intervals due to the limited sample size, and post-hoc selection of the fusion weight. We also did not reproduce a recent explainable or neuro-symbolic fall-risk baseline on LTMM, so the comparison is restricted to the SVM, CNN, and internal CBM configurations reported here. External validation, nested parameter selection, more comprehensive uncertainty analysis, and clinician-facing studies are required before any deployment claim. The present results therefore support a research direction for interpretable wearable AI rather than a clinically validated screening system.

Acknowledgments

This work was prepared as part of the PhD program in Applied Artificial Intelligence at Alma Mater Europaea. The authors thank Prof. Uroš Marušič for constructive methodological and statistical feedback on this work.

Declaration on Generative AI

During preparation of this paper, the authors used OpenAI ChatGPT for the following contributions from the CEUR-WS GenAI Usage Taxonomy: citation management, grammar and spelling check and formatting assistance. All AI-assisted suggestions were reviewed and, where appropriate, rewritten by the authors. No AI-generated experimental results or unverified citations were added. The authors take full responsibility for the content of the publication.

References

[1] World Health Organization, Falls, https://www.who.int/news-room/fact-sheets/detail/falls, 2021. Accessed 2026-06-27.

[2] A. Weiss, T. Herman, M. Plotnik, M. Brozgol, N. Giladi, J. M. Hausdorff, An instrumented timed up and go: the added value of an accelerometer for identifying fall risk in idiopathic fallers, Physiological Measurement 32 (2011) 2003–2018. doi:10.1088/0967-3334/32/12/009.

[3] J. L. Gonzalez-Castro, R. Leiros-Rodriguez, C. Prada-Garcia, J. A. Benitez-Andrades, The applications of artificial intelligence for assessing fall risk: Systematic review, Journal of Medical Internet Research 26 (2024) e54934. doi:10.2196/54934.

[4] L. Xiang, Z. Gao, P. Yu, J. Fernandez, Y. Gu, R. Wang, E. M. Gutierrez-Farewik, Explainable artificial intelligence for gait analysis: Advances, pitfalls, and challenges – a systematic review, Frontiers in Bioengineering and Biotechnology 13 (2025) 1671344. doi:10.3389/fbioe.2025.1671344.

[5] R. Rosenbacke, A. Melhus, M. McKee, D. Stuckler, How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: Systematic review, JMIR AI 3 (2024) e53207. doi:10.2196/53207.

[6] Y. El Marhraoui, S. Bouilland, M. Boukallel, M. Anastassova, M. Ammi, CNN-based self-attention weight extraction for fall event prediction using balance test score, Sensors 23 (2023) 9194. doi:10.3390/s23229194.

[7] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, 2020, pp. 5338–5348. URL: https://proceedings.mlr.press/v119/koh20a.html.

[8] A. d. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, S. N. Tran, Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning, Journal of Applied Logics 6 (2019) 611–632. URL: https://arxiv.org/abs/1905.06088.

[9] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: Components

of a new research resource for complex physiologic signals, Circulation 101 (2000) e215–e220. doi:10.1161/01.CIR.101.23.e215.

[10] A. Weiss, M. Brozgol, M. Dorfman, T. Herman, S. Shema, N. Giladi, J. M. Hausdorff, Does the evaluation of gait quality during daily life provide insight into fall risk? a novel approach using 3-day accelerometer recordings, Neurorehabilitation and Neural Repair 27 (2013) 742–753. doi:10.1177/1545968313491004.

[11] A. Sucerquia, J. D. Lopez, J. F. Vargas-Bonilla, Sisfall: A fall and movement dataset, Sensors 17 (2017) 198. doi:10.3390/s17010198.

[12] C. Voisard, R. Barrois, N. de l'Escalopier, N. Vayatis, P.-P. Vidal, A. Yelnik, D. Ricard, L. Oudre, A dataset of clinical gait signals with wearable sensors from healthy, neurological, and orthopedic cohorts, Scientific Data 12 (2025) 1674. doi:10.1038/s41597-025-05959-w.

[13] A. Shumway-Cook, S. Brauer, M. Woollacott, Predicting the probability for falls in community-dwelling older adults using the timed up & go test, Physical Therapy 80 (2000) 896–903. doi:10.1093/ptj/80.9.896.

[14] M. Montero-Odasso, et al., World guidelines for falls prevention and management for older adults: A global initiative, Age and Ageing 51 (2022) afac205. doi:10.1093/ageing/afac205.

[15] M. M. Lusardi, S. Fritz, A. Middleton, L. Allison, M. Wingood, E. Phillips, M. Criss, E. Verma, J. Osborne, K. K. Chui, Determining risk of falls in community-dwelling older adults: A systematic review and meta-analysis using posttest probability, Journal of Geriatric Physical Therapy 40 (2017) 1–36. doi:10.1519/JPT.0000000000000099.

[16] Y. Lajoie, S. P. Gallagher, Predicting falls within the elderly community: Comparison of postural sway, reaction time, the berg balance scale and the activities-specific balance confidence (ABC) scale for comparing fallers and non-fallers, Archives of Gerontology and Geriatrics 38 (2004) 11–26. doi:10.1016/S0167-4943(03)00082-7.