Bursting the Bubble: Data Leakage and Inflated Deep Learning Accuracy in Multivariate Time-Series Frailty Classification

Charmayne M.L. Hughes*, *Member, IEEE*, and Yan Zhang, *Student Member, IEEE*

Abstract*— *Objective: **To evaluate the reliability of deep learning models for multi-class frailty prediction and to identify methodological issues that may lead to inflated performance.** ***Methods:*** **We reimplemented a recently published InceptionTime-based approach that reported >98% accuracy for classifying frail, pre-frail, and non-frail states using the GSTRIDE dataset. The original pipeline was examined for potential data leakage. We then corrected these issues by applying subject-wise data partitioning, restricting feature scaling to training sets, and ensuring independence of sliding time windows.** ***Results*****: The original implementation achieved near-perfect accuracy, consistent with prior claims. However, we identified multiple sources of data leakage, including pre-split scaling, overlapping windows across training and test sets, and record-wise rather than subject-wise partitioning. When corrected, the model's accuracy decreased to ~42%, reflecting a more realistic estimate of generalization performance.** ***Conclusion*****: High accuracy in multi-class frailty prediction reported in prior work can be largely attributed to methodological flaws. Our corrected pipeline demonstrates that frailty classification from wearable sensor data is substantially more challenging than previously suggested.** ***Significance*****: This study highlights the risks of overoptimistic claims in clinical AI research arising from data leakage and improper evaluation. By providing a reproducible and rigorously validated pipeline, our work establishes a reliable baseline for future studies, supporting the development of clinically meaningful and generalizable AI models for frailty assessment.**

***Index Terms*—frailty, multivariate time-series classification, deep learning, data leakage**

I. Introduction

Over the past few years, advances in wearable sensor technology [1], combined with deep learning algorithmic advances and large labeled datasets [2-3], have significantly bolstered our ability to assess and predict age-related health states in older adults. Within this context, predicting frailty has emerged as a particularly important application, as frailty is a strong predictor of adverse outcomes such as hospitalization, disability, and mortality, yet remains difficult to diagnose reliably in clinical practice.

Against this backdrop, a recent paper [4] reported results that are unusually high compared to prior studies. Applying deep learning models (i.e., convolutional neural network [CNN], convolutional LSTM [ConvLSTM], InceptionTime) to raw acceleration and gyroscope signals from foot-worn IMUs in the GSTRIDE database [5-6], the authors achieved exceptionally strong performance, with the best performing model, InceptionTime, reaching validation and test accuracies of 98%. While high classification performance can be easily achieved in binary classification tasks (e.g., frail vs. non-frail, robust vs. prefrail and frail) using sociodemographic survey [7], voice [8], and sensor data [9], such accuracy is unusually high for multi-class frailty prediction, where distinguishing between frail, pre-frail, and non-frail states is considerably more challenging and typically yields substantially lower performance, generally in the range of 65–85% [10-12].

Collectively, these findings suggest that the exceptionally high accuracy reported by [4] for multi-class frailty prediction warrants closer scrutiny. Examination of the methods reveals several concerns that undermine the reported results. First, data were scaled prior to splitting into training (DT) and out-of-training (DE) subsets using RobustScaler. RobustScaler transforms each feature xi according to $x'_i = \frac{x_i - Q_2(x)}{Q_3(x) - Q_1(x)}$, where $Q2$ is the median and $Q1, Q3$ are the first and third quartiles, respectively. Because these scaling parameters Q were estimated per feature across the entire dataset D, the resulting scaled training set DT implicitly incorporated information about the distribution of the out-of-training data DE. Although DE was not directly included in model fitting, leakage occurred through the use of its statistical properties (i.e., through $Q1, Q3$), which biased the evaluation process [13].

Second, before partitioning D into DT, DV, and test (DT') subsets, the raw IMU signals were segmented into overlapping windows (window size = 200, 50% overlap; i.e., step size = 100 [although incorrectly stated as 50]). Randomly distributing these windows across subsets violates the assumption that samples are independently and identically distributed (*i.i.d.*), as consecutive and overlapping windows are not independent [14].

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. National Science Foundation under Grant BS123456". Charmayne Mary Lee Hughes is with the Department of Age-Appropriate Human-Machine Systems at the Technische Universität Berlin, Berlin, Germany (correspondence e-mail: hughes@tu-berlin.de). Yan Zhang is with the Department of Age-Appropriate Human-Machine Systems at the Technische Universität Berlin.

Consequently, overlapping windows derived from the same temporal segment of a participant's gait can appear in multiple subsets, meaning the sample set S is not separated into disjoint subsets DT and DE (i.e., $DE \cap DT \neq \emptyset$). This allows information from DV and DT' to influence training, artificially inflating performance estimates and violating the principle of evaluating on truly unseen data.

Third, the authors used a record-wise split, randomly assigning individual measurement to DT, DV, and DT'. Consequently, records from the same participant appear in multiple partitions, allowing classifiers to learn subject-specific signatures and effectively perform subject identification instead of the intended frailty class recognition. Record-wise splitting has been shown, by simulation [15-16] and by analyses of clinical datasets [17-18], to substantially underestimate the true prediction error and dramatically overestimate real-world performance.

Taken together, these sources of preprocessing, segmentation, and partitioning leakage strongly suggest that the high classification accuracies reported by [4] are inflated and do not reflect true model generalizability. Amjad et al. [4] implemented their deep learning models using the Python-based McFly framework, which prevents exact replication of the original model architectures without access to their code. To address this limitation, we reimplemented the best-performing model, InceptionTime, in PyTorch using the publicly available GSTRIDE dataset. To ensure methodological rigor, we applied a corrected pipeline with subject-wise partitioning, scaling restricted to the training data, and moving windows applied separately to each subset to prevent leakage. Under these conditions, classification accuracy decreased substantially (~42%), providing a more realistic assessment of model performance and highlighting the significant impact that data leakage can have on deep learning approaches for frailty prediction.

II. METHODS

A. The GSTRIDE Database

The publicly available GSTRIDE dataset [5-6] was used to classify the current frailty status of older adults. Frailty status was defined according to the standardized Fried's phenotype criteria [19], with 65 participants classified as non-frail, 73 as pre-frail, and 20 as frail (see Table 1 for participant characteristics). Participants performed a 4-meter walk test while wearing the G-STRIDE device, which comprises an IMU attached to the top of the foot. Two types of IMU modules were used: the iNEMO inertial module (LSM6DSRX, STMicroelectronics, CH) sampled at 104 Hz with a measurement range of ±2000°/s and ±16 g, or the Physilog 6 S (GaitUp, CH), sampled at 128 Hz with a range of ±1000°/s and ±8 g. Raw acceleration and gyroscope signals from the IMUs were used as input for all analyses without additional filtering or adjustments, consistent with the original GSTRIDE dataset protocol.

TABLE I
CHARACTERISTICS OF PARTICIPANTS BY FRAILTY STATUS CATEGORY.

	Non-Frail	Pre-Frail	Frail
Biological women (%)	70.8	74.0	75.1
Mean age (year)	80.0	84.1	85.4
Mean body mass index (kg/m^2)	25.9	26.1	25.5
Residing in nursing home (%)	21.5	39.7	50
Falls in past year (%)	27.7	58.9	100
Mean global deterioration scale	1.5	2.4	2.5

B. Pipeline

To replicate and critically evaluate the results reported by [4], the InceptionTime model was reimplemented using the GSTRIDE dataset (Figure 1, upper panel). Amjad et al [4] implemented their deep learning models using the Python-based McFly framework, which, without access to the original code, prevents exact replication of their model architectures. To address this, we reimplemented the best-performing model, InceptionTime, in PyTorch, applying the same preprocessing and subject-wise partitioning used in the corrected pipeline. Full details of the original pipeline are available in [4].

Inputs consisted of tri-axial accelerometer and gyroscope signals, normalized using RobustScaler. In the original pipeline, signals for each participant were segmented into overlapping windows (window size = 200 samples, 50% overlap), resulting in input arrays of size 200 × 6 (200 time points × 6 features). The dataset was randomly divided into

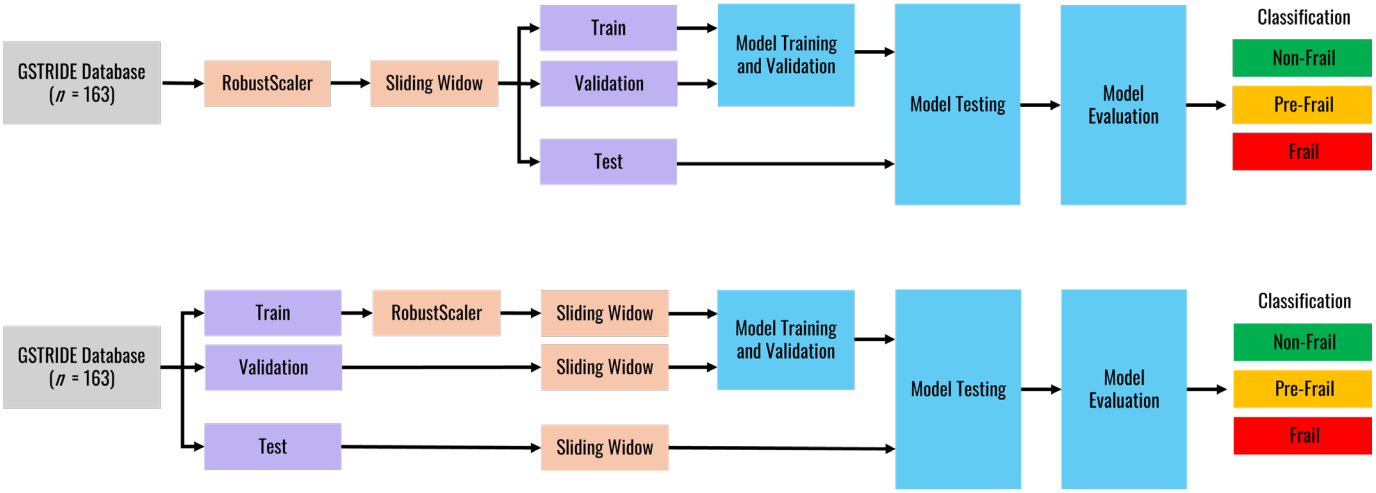

Fig. 1. Complete pipeline from raw IMU signal input to frailty classification output. The top panel shows the workflow outlined in [4], while the bottom panel presents the leakage-corrected pipeline used in the current study.

training (70%), validation (15%), and testing (15%) subsets using a fixed random seed of 42. The model was trained for 25 epochs with a batch size of 64 and early stopping with a patience of 3 epochs. Reported hyperparameter-optimized values were used: learning rate 0.005890, regularization rate 0.025038, network depth 6, 70 filters per convolutional layer, and maximum kernel size 23. Performance was evaluated on the validation and test sets using macro-averaged accuracy, precision, recall, and F1-score.

To obtain a realistic assessment and mitigate data leakage, a corrected pipeline was implemented (Figure 1, lower panel). Five-fold cross-validation was performed with subject-wise splitting, ensuring that no participant contributed data to multiple folds. Each fold maintained disjoint training, validation, and test subsets, comprising 111, 24, and 23 participants respectively, ensuring that no participant contributed data to more than one subset. Feature scaling using RobustScaler was applied exclusively to the training set, with resulting parameters applied to validation and test subsets. Overlapping windows were generated independently within each subset, preserving temporal structure while maintaining independence. The InceptionTime model was trained on this leakage-corrected pipeline using the same hyperparameters as the original implementation to isolate the effect of methodological corrections from hyperparameter tuning. Performance was evaluated on the validation and test sets using macro-averaged accuracy, precision, recall, and F1-score. Experiments were conducted on a workstation with dual NVIDIA GeForce RTX 4090 GPUs and an Intel Core i9-14900K CPU, using Python 3.10 and PyTorch 2.7.1.

TABLE II
CLASSIFICATION PERFORMANCE FOR MULTI-CLASS FRAILTY PREDICTION (FRAIL, PRE-FRAIL, NON-FRAIL) USING RAW IMU SIGNALS FROM THE GSTRIDE DATASET.

Pipeline	Original	Leakage-Corrected
Training Accuracy	96.8%	95.7%
Validation Accuracy	95.9%	49.2%
Test Accuracy	96.4%	55.7%
Test Precision	96.4%	54.9%
Test Recall	96.4%	55.7%
Test F1-Score	96.4%	55.2%

III. RESULTS

Table 2 and Figure 2 compare the performance of the pipeline reported by [4] with the leakage-corrected pipeline implemented in the present study. The original pipeline produced extremely high values across all metrics, with a training accuracy of 96.8%, validation accuracy of 95.9%, and a test accuracy of 96.4%. Precision, recall, and F1-score were also consistently 96.4%. The corresponding confusion matrix (Figure 2, top panel) shows near-perfect classification across the three frailty categories, with very few misclassifications.

When the leakage was corrected, performance dropped substantially. The leakage-corrected pipeline achieved 95.7% accuracy on the training set but only 49.2% on validation and 55.7% on the test set. Precision (54.9%), recall (55.7%), and F1-score (55.2%) on the test set reflect this more realistic performance. The confusion matrix (Figure 2, lower panel) highlights frequent misclassifications between the Pre-Frail and Non-Frail classes, as well as partial overlap between Frail and the other categories.

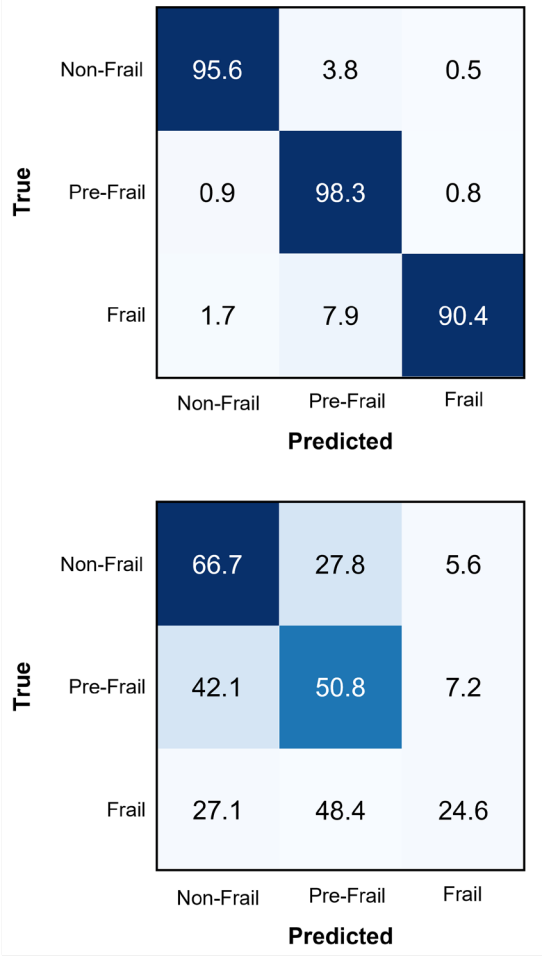

Fig. 1. Confusion matrices showing the classification performance arising from the workflow outlined in [4] (top) and the current study (bottom).

The dramatic performance drop from 95.7% to 55.7% test accuracy demonstrates the substantial impact of data leakage on model evaluation. The corrected results align more closely with typical performance ranges reported in the literature for multi-class frailty prediction (65-85%), suggesting that the original high performance was primarily due to methodological artifacts rather than genuine predictive capability. Notably, the corrected pipeline shows substantial overfitting (95.7% training vs 55.7% test accuracy), whereas the original pipeline appeared to have minimal overfitting, further evidence that data leakage masked the model's true generalization capability.

Figure 3 shows the evolution of training and validation loss over epochs for both the pipeline outlined in [4] and the leakage-corrected pipeline. In both cases, training loss decreased steadily with additional epochs, indicating successful optimization. However, while the Amjad pipeline maintained a close alignment between training and validation curves, the

corrected pipeline exhibited a larger gap, with validation loss plateauing at a higher value. This divergence reflects reduced generalization and highlights the greater difficulty of the task once data leakage is removed.

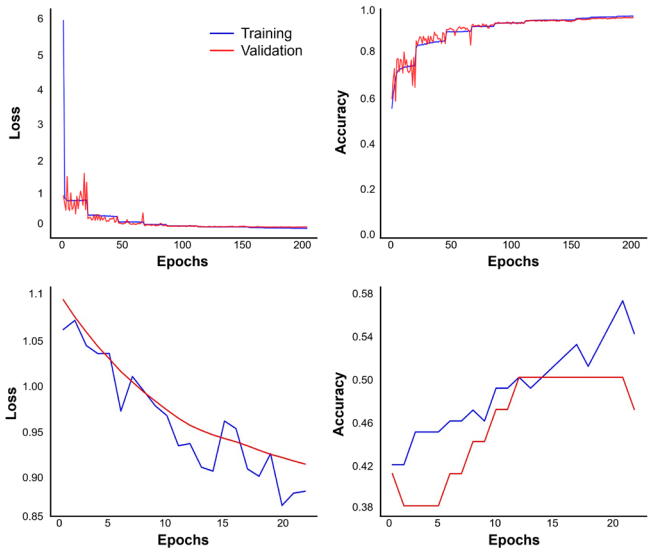

Fig. 3. Training and validation loss curves of the five-fold cross validation for the workflows outlined in [4] (top) and the current study (bottom).

IV. Discussion

This study reimplemented the best-performing model reported by [4], InceptionTime, using the publicly available GSTRIDE dataset. Replicating their original pipeline yielded exceptionally high performance. However, after adjusting the workflow to eliminate potential sources of overfitting, validation and test set accuracy dropped markedly. These results suggest that the original performance was likely inflated due to data leakage, whereby information from the test set inadvertently influenced model training and validation. By applying the same hyperparameters to both the original and leakage-corrected pipelines, this approach isolates the impact of methodological corrections, demonstrating the true effect of data leakage on model performance without conflating it with differences arising from hyperparameter optimization.

Data leakage is a pervasive risk in machine learning pipelines, particularly in high-stakes applications such as healthcare. It can occur in subtle ways, including feature engineering performed before data splitting, the inadvertent inclusion of future outcomes as predictors, improper cross-validation, or overlap between training and test cohorts. In clinical contexts, overly optimistic performance is not harmless: it can mislead clinicians, funders, and policymakers, and may result in false reassurance, inappropriate treatment decisions, or unnecessary interventions. Mitigating these risks requires careful methodology, including subject-level splitting for time-series data, pipeline-aware preprocessing, rigorous cross-validation, external validation on independent cohorts, and transparent reporting of limitations.

User-friendly toolkits and software frameworks (e.g., McFly, scikit-learn, TensorFlow, and PyTorch) provide accessible methods that enable a broad range of practitioners to implement advanced models. While these tools can accelerate research and democratize access to complex algorithms, they also lower the barrier to accidental misuse, allowing sophisticated methods to be applied without full awareness of potential pitfalls such as data leakage. Although the toolkit itself did not contribute to leakage in this study, these software frameworks themselves do not inherently cause data leakage, these considerations underscore the need for careful pipeline design, critical evaluation of methodological choices, and transparency when developing ML workflows.

While our leakage-corrected results are substantially lower than the near-perfect values reported by Amjad et al., they are consistent with other studies using single-modality IMU data. For instance, [10] applied an LSTM network to upper-limb IMU signals and achieved 74% accuracy, illustrating the typical performance range when using raw sensor data alone. Multimodal approaches can further improve performance: [11] combined lower-limb IMU signals with clinical features processed via principal component analysis, achieving 85.2% accuracy using a ResNet-50 and feedforward network. These examples highlight that even carefully engineered, multimodal systems rarely approach the inflated performance seen in leakage-prone pipelines. Despite being more modest, our corrected performance provides a reliable baseline that can support early identification of individuals at risk for frailty, complementing clinical judgment, and guiding future methodological improvements.

In healthcare applications, methodological rigor is not only a scientific concern but also an ethical imperative. Inflated or overfit models can lead to premature adoption, false reassurance, or inappropriate clinical decisions, with direct implications for patient safety and trust. Transparent reporting, reproducible pipelines, and realistic performance baselines are essential to ensure that ML models provide clinically meaningful and actionable results. Our findings underscore that rigorous safeguards against data leakage are critical for trustworthy machine learning, where inflated accuracy does not merely mislead the research community but can also have direct consequences for patient care.

V. Conclusion

This study reimplemented the InceptionTime model reported by [4] using the GSTRIDE dataset and demonstrated that its near-perfect performance was largely the result of data leakage. By correcting methodological flaws such as pre-split feature scaling, overlapping time windows, and record-wise rather than subject-wise partitioning, accuracy dropped to a more realistic level. These results isolate the impact of leakage and underscore its profound influence on model performance.

Data leakage remains one of the most pervasive risks in clinical machine learning. Although often subtle, such errors can substantially inflate results and create misleading impressions of generalizability. In healthcare, these risks extend beyond methodological validity, as overly optimistic models may lead to premature adoption, false reassurance, or misguided clinical decision-making. Methodological rigor, therefore, is both a scientific and ethical requirement.

Our findings also highlight the double-edged nature of accessible machine learning toolkits. While frameworks such as scikit-learn, TensorFlow, and PyTorch accelerate innovation, they can also obscure potential pitfalls. Researchers must remain vigilant, employing subject-level splitting, pipeline-aware preprocessing, robust cross-validation, external validation, and transparent reporting.

Despite the lower performance observed in our corrected pipeline, the results align with prior studies using single-modality sensor data and provide a credible baseline for future work. Improvements are more likely to come from multimodal integration and refined clinical features than from methodological shortcuts. Ultimately, ensuring reproducibility and preventing leakage are essential for building trustworthy AI systems. Our work provides both a cautionary example and a methodological framework, reinforcing that reliable evaluation is indispensable if machine learning models are to make safe and meaningful contributions to frailty prediction and broader healthcare applications.

REFERENCES

[1] S. Malwade, S. S. Abdul, M. Uddin, A. A. Nursetyo, L. Fernandez-Luque, X. K. Zhu, et al., "Mobile and wearable technologies in healthcare for the ageing population," Computer Methods and Programs in Biomedicine, vol. 161, pp. 233–237, 2018.

[2] A. Das and P. Dhillon, "Application of machine learning in measurement of ageing and geriatric diseases: A systematic review," BMC Geriatrics, vol. 23, no. 1, p. 841, 2023.

[3] J. T. Wassan, H. Zheng, and H. Wang, "Role of deep learning in predicting aging-related diseases: a scoping review," Cells, vol. 10, no. 11, p. 2924, 2021.

[4] A. Amjad, A. Szczęsna, M. Błaszczyszyn, and A. Anwar, "Inertial measurement unit signal-based machine learning methods for frailty assessment in geriatric health," Signal, Image and Video Processing, vol. 19, no. 2, p. 105, 2025.

[5] G. García-Villamil Neira, M. Neira Álvarez, E. Huertas Hoyas, L. Ruiz Ruiz, S. García-de-Villa, A. J. del-Ama, M. C. Rodríguez Sánchez, and A. Jiménez Ruiz, "GSTRIDE: A database of frailty and functional assessments with inertial gait data from elderly fallers and non-fallers populations (Version v1.0)," Zenodo, 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6883292

[6] S. García-de-Villa, G. G. V. Neira, M. N. Álvarez, E. Huertas-Hoyas, L. R. Ruiz, A. J. Del-Ama, et al., "A database with frailty, functional and inertial gait metrics for the research of fall causes in older adults," Scientific Data, vol. 10, no. 1, p. 566, 2023.

[7] C. M. L. Hughes, Y. Zhang, A. Pourhossein, and T. Jurasova, "A comparative analysis of binary and multi-class classification machine learning algorithms to detect current frailty status using the English longitudinal study of ageing (ELSA)," Frontiers in Aging, vol. 6, p. 1501168, 2025.

[8] T. Kim, J. Y. Choi, M. J. Ko, and K. I. Kim, "Development and validation of a machine learning method using vocal biomarkers for identifying frailty in community-dwelling older adults: cross-sectional study," JMIR Medical Informatics, vol. 13, no. 1, p. e57298, 2025.

[9] A. Apsega, L. Petrauskas, V. Alekna, K. Daunoraviciene, V. Sevcenko, A. Mastaviciute, et al., "Wearable sensors technology as a tool for discriminating frailty levels during instrumented gait analysis," Applied Sciences, vol. 10, no. 23, p. 8451, 2020.

[10] M. Asghari, H. Ehsani, and N. Toosizadeh, "Frailty identification using a sensor-based upper-extremity function test: a deep learning approach," Scientific Reports, vol. 15, no. 1, p. 13891, 2025.

[11] J. Griškevičius, K. Daunoravičienė, L. Petrauskas, A. Apšega, and V. Alekna, "Retrospective frailty assessment in older adults using inertial measurement unit-based deep learning on gait spectrograms," Sensors, vol. 25, no. 11, p. 3351, 2025.

[12] C. Y. Yang, N. Premakumara, H. L. Chiu, Y. H. Feng, T. Y. Chen, and C. Shiranthika, "Development of gravitationally aligned pendant IMU frailty identifier," Computers and Electrical Engineering, vol. 118, p. 109466, 2024.

[13] S. Whalen, J. Schreiber, W. S. Noble, and K. S. Pollard, "Navigating the pitfalls of applying machine learning in genomics," Nature Reviews Genetics, vol. 23, no. 3, pp. 169–181, 2022.

[14] T. Plötz, "Applying machine learning for sensor data analysis in interactive systems: common pitfalls of pragmatic use and ways to avoid them," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1–25, 2021.

[15] E. C. Neto, A. Pratap, T. M. Perumal, M. Tummalacherla, B. M. Bot, A. D. Trister, et al., "Learning disease vs participant signatures: a permutation test approach to detect identity confounding in machine learning diagnostic applications," arXiv preprint arXiv:1712.03120, 2017.

[16] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," Gigascience, vol. 6, no. 5, p. gix019, 2017.

[17] G. Brookshire, J. Kasper, N. M. Blauch, Y. C. Wu, R. Glatt, D. A. Merrill, et al., "Data leakage in deep learning studies of translational EEG," Frontiers in Neuroscience, vol. 18, p. 1373515, 2024.

[18] E. C. Neto, A. Pratap, T. M. Perumal, M. Tummalacherla, P. Snyder, B. M. Bot, A. D. Trister, S. H. Friend, L. Mangravite, and L. Omberg, "Detecting the impact of subject characteristics on machine learning-based diagnostic applications," NPJ Digital Medicine, vol. 2, p. 99, 2019. [Online]. Available: https://doi.org/10.1038/s41746-019-0178-x.

[19] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, et al., "Frailty in older adults: evidence for a phenotype," The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, vol. 56, no. 3, pp. M146–M157, 2001.