

DMD2026-1095

EFFICIENCY OVER COMPLEXITY: OPTIMAL NEURAL ARCHITECTURES FOR NEAR-TERM FORECASTING OF PARKINSONIAN FREEZING OF GAIT

Yichen Wu
Storm King School

Junyu Chen
Storm King School

Hao Su*
Tandon School of Engineering
New York University

ABSTRACT

Freezing of gait (FOG) is a disabling Parkinson's symptom. Wearable IMUs provide continuous monitoring and require careful model selection. We frame FOG prediction as near-term forecasting, predicting onsets over a short horizon from brief IMU windows. We use a public freezing locomotion detection dataset and compare a frequency domain baseline (band-power ratio baseline) with eight deep learning models, including CNN, LSTM, Transformer, and hybrids, under a unified pre-processing and two evaluation protocols. We evaluated performance on both our prior held-out and random split. Hybrids with CNN front-ends and lightweight temporal layers performed best, achieving 90% accuracy and F1 score of 0.94 for the impending-FOG class (precision 0.91, recall 0.97) while remaining compact ($\approx 5 \times 10^4$ parameters). Additionally, re-evaluating the same models with a 16-fold leave-one-subject-out cross-validation (LOSO-CV) yields more realistic and lower scores that reflect generalization to unseen users. Under LOSO-CV, the best model (Deeper CNN) attains accuracy = 0.80 ± 0.09 , FOG precision = 0.81 ± 0.11 , recall = 0.89 ± 0.10 , and F1 = 0.84 ± 0.09 (mean $\pm$ SD across 16 subjects). These results show that the old protocol overestimates generalization, whereas LOSO-CV reveals true cross-subject performance, providing a more reliable benchmark for subject robustness FOG forecasting and threshold calibration.

1 Introduction

Freezing of gait (FOG) is a debilitating Parkinson's disease symptom linked to falls and loss of independence. Wearable in-

ertial measurement units (IMUs) enable continuous gait monitoring, providing reliable data for anticipating FOG episodes. Recent studies in wearable robots and on-body inertial sensing have improved data accuracy and deployment feasibility for subject independence forecasting [1–9]. Recent deep learning (DL) methods learn discriminative temporal spatial patterns directly from raw signals [10]. However, there remains limited systematic comparative research on near-term FOG forecasting, where the goal is to predict onsets within short IMU windows.

We present a unified comparison that uses a frequency domain baseline [11] and seven DL architectures, including CNN, LSTM, Transformer, and hybrid architectures, for IMU-based FOG on both feet forecasting on a public dataset [12]. Our preprocessing follows established human activity recognition methodology [10, 11]. Beyond overall accuracy, a FOG predictor must balance sensitivity to impending events with a low false alarm rate during normal gait. We therefore benchmark multiple models with respect to precision–recall trade-offs that directly impact the practical usability of forecasting systems.

A conventional window-level random split of this dataset can result in windows from the same subject appearing in both the training and test sets. As a result, the evaluation results mainly reflect how well the model recognizes additional windows from the same individuals, rather than its ability to generalize to new patients. In contrast, cross-subject generalization, the extent to which a model performs well on individuals not seen during training, is important for real-world robustness.

Therefore, in this work, we evaluate each architecture using two complementary protocols: (i) a conventional window-level random split, and (ii) a subject-robust 16-fold Leave-One-Subject-Out Cross-Validation (LOSO-CV). Using the same

*Corresponding Author.

dataset, preprocessing pipeline, and training setup for both protocols allows us to isolate the result of evaluation design on model performance.

This paper makes two main contributions. First, we compare a frequency domain baseline with seven deep learning architectures for near-term FOG forecasting from foot-mounted IMUs, using a unified preprocessing and training pipeline. Second, we evaluate all architectures using both a window-level random split and a subject robustness 16-fold LOSO-CV protocol. These evaluations quantify the generalization gap and provide practical recommendations for improving subject robustness evaluation, threshold selection, and model capacity in this context.

While traditional Frequency Domain (FD) approaches offer a straightforward mechanism for FOG detection, they rely on identifying the spectral footprint of an ongoing freeze. Consequently, they are ill-equipped for near-term forecasting, which requires anticipating an episode before the distinct freezing oscillation manifests. As our benchmark reveals, FD methods yield negligible recall for impending events. This profound limitation necessitates the adoption of Deep Learning (DL) models, which can extract highly non-linear, subtle spatiotemporal gait degradation patterns that precede an FOG episode.

2 Methods

We established a systematic framework to investigate the optimal neural architecture for near-term FOG prediction. The methodology is designed to rigorously compare frequency domain baselines against deep learning models of varying complexity, specifically focusing on their ability to generalize to unseen subjects.

2.1 Data and Preprocessing Pipeline

We utilized the public dataset from Dvorani et al. [12], which captures high-fidelity gait dynamics from 16 individuals with Parkinson's disease (PD) using synchronized 3D acceleration and angular velocity sensors recorded at 200 Hz, as shown in Figure 1. Unlike detection tasks that simply classify the current state, we framed the problem as near-term forecasting to facilitate early intervention. To achieve this, we implement a sliding-window pipeline designed to capture "pre-FOG" kinematic signatures.

Specifically, we utilized a 4-second input window to provide sufficient historical context for the model to identify gait degradation patterns, targeting a prediction horizon of 2 seconds immediately following the input. Crucially, to prioritize clinical safety and sensitivity, we employed an "any-event" labeling strategy: if any time step within the 2-second future horizon contains a FOG event, the current 4-second input is labeled as "Impending FOG". This safety-critical approach trains the model to anticipate the onset of freezing rather than just the continuation

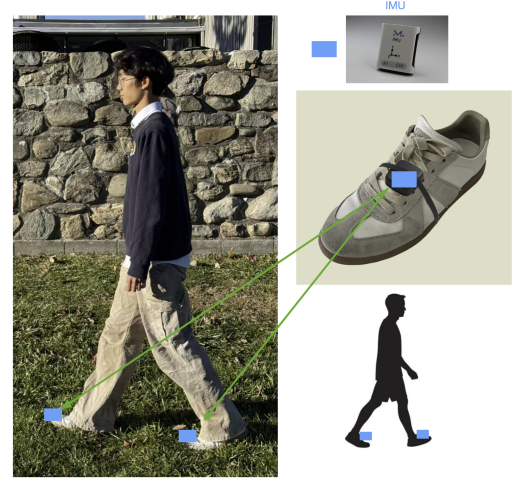


FIGURE 1. A visual representation of the sensor placement and a sample of the raw time-series data can be seen

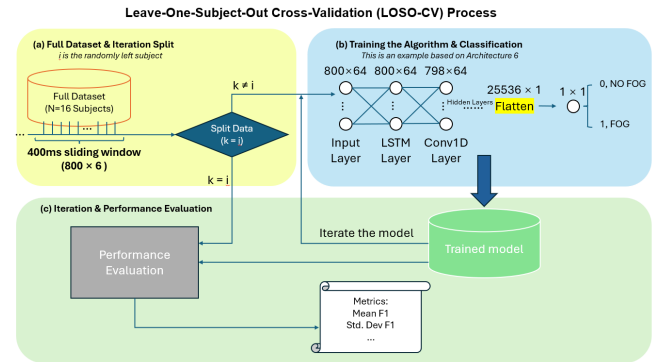


FIGURE 2. The generalization of the pipeline. All data are preprocessed as an 800×6 matrix and then processed with the algorithm. The figure used Architecture 2, the CNN baseline algorithm, as an example.

of an ongoing episode, as shown in Figure 2

2.2 Architectures of Deep Learning

To systematically isolate the specific contribution of spatial features versus temporal memory and model capacity, we organized the eight evaluated architectures into three distinct comparative blocks.

The first category evaluates fundamental and lightweight architectures to determine the efficacy of single-modality processing. Architecture 1 acts as a pure baseline. It utilizes the Frequency Domain (FD) Freeze Index algorithm directly from Bachlin et al as a non-learning reference [11]. The model outputs a binary prediction by applying thresholds to the Freeze Index and band power, using values tuned on training data within

each evaluation fold.

From Architecture 2 and 3, we employ a single LSTM layer with 64 units and a stacked 1D Convolutional layers with 64 filters, respectively, to extract local kinematic motifs. This group acts as the simple group in our examination.

The second category investigates spatiotemporal hybrids to test the synergy between feature extraction and temporal modeling. Architecture 4 (CNN-LSTM) and Architecture 5 (LSTM-CNN) integrate both convolutional and recurrent layers but vary the module order to investigate information flow.

The third category explores whether increasing model complexity and introducing advanced mechanisms improve generalization. Architecture 6, the Deeper CNN, optimizes the layer capacity for this specific dataset by scaling up the standard CNN baseline. We increased the convolutional filters from 64 to 128 per layer. Architecture 7 deploys a state-of-the-art Transformer Encoder with Multi-Head Attention. Finally, Architecture 8 (LSTM-TCN) builds upon the insights from our hybrid models by replacing the standard convolutional layer with a Temporal Convolutional Network.

2.3 Evaluation Protocol

To address the issue of subject-specific overfitting common in prior studies, we adhered to a strict Leave-One-Subject-Out Cross-Validation (LOSO-CV) protocol. The evaluation consisted of 16 separate folds, where in each fold, one subject was completely isolated as the test set while the model was trained on the remaining 15 subjects. This rigorous method simulates the "new patient" scenario, addressing a critical gap in the literature where findings based on randomized splits do not necessarily validate a model's true clinical utility. Within each training fold, we further withheld a 15% validation split for early stopping to prevent optimization leakage into the test set. We also performed a standard randomized split evaluation to explicitly quantify the "generalization gap"—the performance drop observed when moving from optimistic random splits to strict subject independence evaluation.

In the clinical context of FOG forecasting, prioritizing Recall is essential to maximize sensitivity to impending episodes and minimize the risk of unmitigated falls. Consequently, we utilize the F1-score to rigorously balance this critical need for patient safety against the risk of excessive false alarms, ensuring practical usability without inducing alarm fatigue. Thus, all model benchmarking will be based on the recall rate and F1 score in the following evaluation of algorithms.

3 Results

The forecasting performance of the eight architectures was evaluated using two distinct protocols to assess both theoretical capacity and clinical generalization. To comprehensively

quantify model reliability, we calculated Recall, Precision, and the F1-score for the "Impending FOG" class. Recall measures the model's sensitivity in correctly identifying upcoming freeze events, while Precision reflects the proportion of true alarms among all positive predictions. Consequently, the F1-score, representing the harmonic mean of these two metrics, serves as the primary indicator of overall performance by balancing the critical trade-off between missing a freeze event and triggering excessive false alarms.

First, under the randomized train-test split shown in Table 2, all deep learning models demonstrated high predictive capability. The deep learning architectures consistently achieved F1-scores exceeding 0.90, with the LSTM-CNN hybrid recording the highest score of 0.94. In contrast, the Frequency Domain baseline (Architecture 1) proved ineffective for the forecasting task. While it maintained high precision, it recorded a negligible recall of approximately 0.06, resulting in an F1-score of 0.11.

Second, applying the subject independence LOSO-CV protocol in Table 1 resulted in a performance decrease across all learning models compared to the randomized split, reflecting the challenge of generalizing to unseen gait patterns. In this rigorous evaluation, the Convolutional Neural Networks emerged as the most robust architectures. Architecture 3 (CNN-64) achieved the highest performance balance, recording an F1-score of 0.84 and the highest recall of 0.92 among all evaluated models.

Third, regarding model complexity, increasing the parameter count did not yield proportional improvements in accuracy. Architecture 6 (Deeper CNN-128), despite doubling the number of convolutional filters, achieved an identical F1-score (0.84) to the lighter Architecture 3 and recorded a slightly lower recall of 0.89. The spatiotemporal hybrids (Architecture 4 and 5) followed closely behind the pure CNNs with F1-scores of approximately 0.82. Finally, the architectures designed for long-range dependencies, specifically the Transformer (Architecture 7) and LSTM-TCN (Architecture 8), did not outperform the baseline CNNs, achieving F1-scores of 0.79 and 0.80, respectively.

4 Discussion

4.1 Overall comparison

The Frequency Domain (FD) baseline performed poorly under the forecasting evaluation [11]. Under both the randomized split and LOSO-CV protocols, it maintained high precision but had extremely low recall ($F1 \approx 0.11$). This indicates thresholds tuned for detection do not transfer well to predicting near-term onsets from short IMU windows. In contrast, all deep learning architectures substantially improved forecasting performance, achieving F1-scores above 0.90 across randomized splits and LOSO-CV.

By comparing the two evaluation protocols. The random split yields uniformly high F1 Scores for all deep learning models, with at least 0.90. However, this protocol allows windows

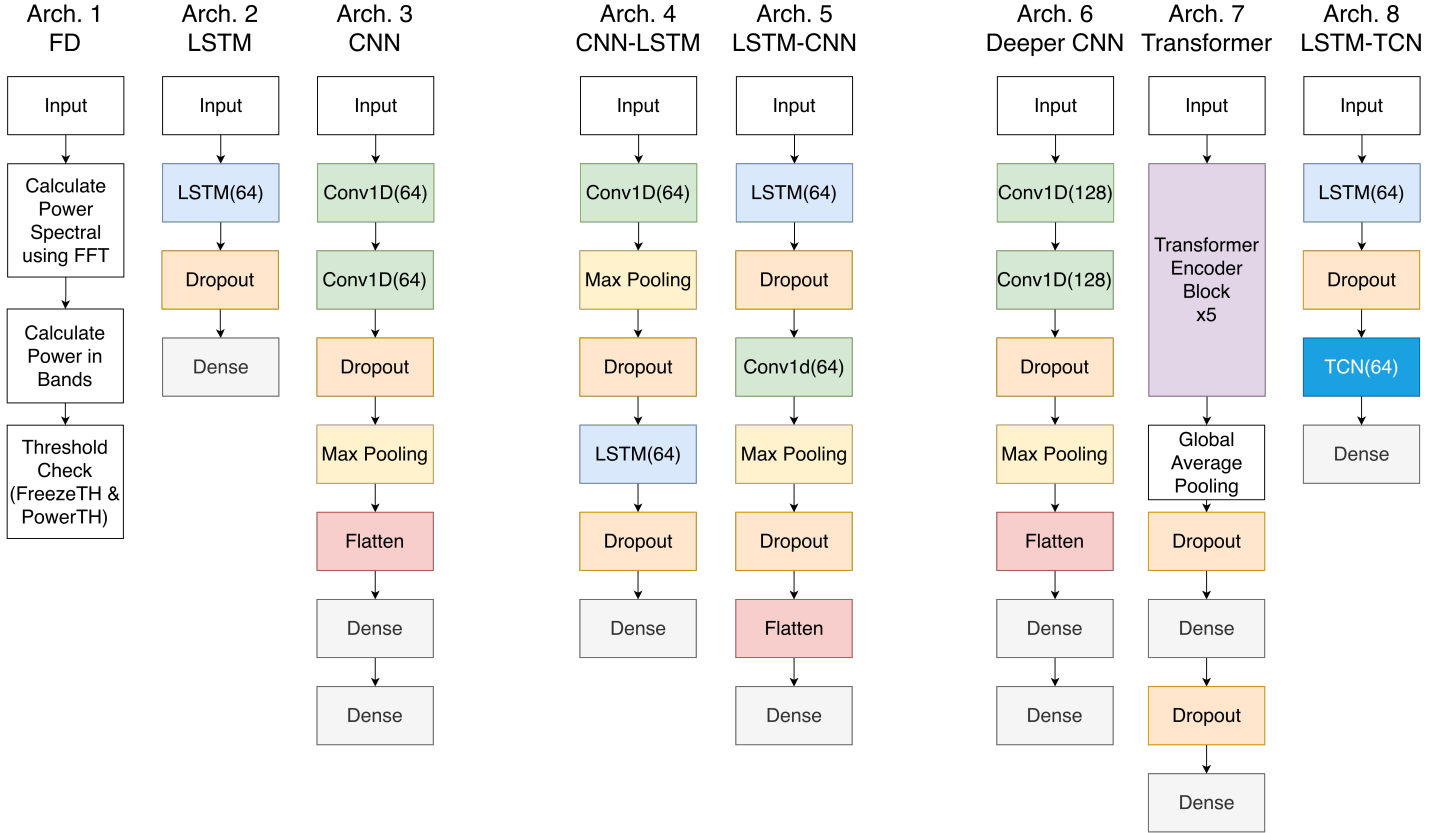


FIGURE 3. Schematic overview of the eight proposed algorithmic architectures. We categorized them into three comparative blocks to sequentially evaluate fundamental single-modality architectures, investigate the synergy of spatiotemporal fusion, and assess the limits of high-capacity advanced models. The number in parentheses stands for how many neurons are involved in the training process.

TABLE 1. LOSO-CV PERFORMANCE COMPARISON (MEAN VALUES). FD BASELINE SHOWN FOR REFERENCE.

Architecture	Recall	Precision	F1-Score	Avg. Accuracy
FD (Baseline)	0.0588	0.9812	0.1110	0.1730
LSTM (64)	0.8807	0.7530	0.7967	0.7288
CNN (64)	0.9159	0.7883	0.8415	0.7916
CNN (128)	0.8920	0.8075	0.8415	0.7981
CNN-LSTM	0.8883	0.7829	0.8240	0.7718
LSTM-CNN	0.8433	0.8097	0.8183	0.7744
Transformer	0.8176	0.7852	0.7870	0.7369
LSTM-TCN	0.9053	0.7419	0.7978	0.7218

TABLE 2. ACCURACY COMPARISONS OF ALGORITHMS

Architecture	Recall	Precision	F1-Score	Avg. Accuracy
FD	0.0588	0.9812	0.1110	0.1730
LSTM (64)	0.9256	0.8732	0.8986	0.8324
CNN (64)	0.9420	0.9008	0.9210	0.8677
CNN (128)	0.9426	0.8938	0.9176	0.8641
CNN-LSTM	0.9464	0.9142	0.9300	0.8825
LSTM-CNN	0.9691	0.9093	0.9382	0.8968
Transformer	0.9616	0.9126	0.9365	0.8917
LSTM-TCN	0.9559	0.8711	0.9115	0.8498

from the same subject to appear in both training and test sets, so the results mainly reflect how well models recognize additional windows from the same individuals rather than how well

they perform on unseen patients. Using the subject independence LOSO-CV protocol results in a consistent drop in performance across architectures, with the best F1 Score decreasing to 0.84. This gap highlights the challenge of cross-subject generalization and underscores the importance of subject awareness evaluation

for forecasting models intended for real-world use.

Under LOSO-CV, convolutional neural networks proved to be the most robust class of models. The CNN-64 architecture achieved the best overall balance, with the highest recall (0.92) and the highest mean F1-score (0.84) among all evaluated models. Hybrid and long-range architectures (CNN-LSTM, LSTM-CNN, LSTM-TCN, and the Transformer) remained competitive but did not surpass the simpler CNNs in this LOSO-CV evaluation.

4.2 Model Complexity and Size of CNNs

Our core finding underscores the principle of efficiency over complexity within the selection of deep learning architectures for near-term FOG forecasting. Evaluating the models under the LOSO-CV protocol reveals that increasing structural complexity and parameter count does not directly lead to improved generalization. As detailed in Table 1, the lightweight CNN (64) architecture achieves the optimal balance, delivering the highest F1-score of 0.8415 and a leading recall rate of 0.9159 among the top-performing configurations.

Scaling up the model capacity yields diminishing returns, particularly concerning the model's sensitivity to impending events. For instance, the deeper CNN (128), which doubles the number of convolutional filters, maintains an identical F1-score of 0.8415 but shows a reduction in recall to 0.8920. This indicates that the added capacity does not assist in identifying subtle pre-freeze signatures, but rather marginally shifts the decision boundary.

Similarly, architectures designed for long-range temporal dependencies fundamentally underperform in this specific forecasting context. The Transformer model yields the lowest deep learning performance across both key metrics of recalling rate and F1 score. For the spatiotemporal hybrids (CNN-LSTM and LSTM-CNN), they fail to surpass the baseline CNN, hovering around F1-scores of 0.81 to 0.82. These findings confirm that for predicting FOG within short 4-second windows from foot-mounted IMUs, high-capacity recurrent and attention-based models are prone to overfitting to subject-specific gait patterns or background noise. Lighter 1D-CNNs successfully avoid this over-parameterization, offering the best trade-off between predictive sensitivity (recall) and overall reliability (F1-score), which is particularly attractive for wearable systems constrained by embedded compute capabilities [1, 2, 5, 6, 9].

4.3 Attributes and limitations

The superior performance of lightweight CNNs stems from the highly localized nature of pre-FOG transitions. 1D convolutions efficiently capture these short-duration, rapid kinematic shifts due to their restricted receptive fields, a distinct advantage in extracting local morphological motifs from IMU data [10, 13]. In contrast, heavy sequential models (LSTMs, Transformers) at-

tempt to construct historical contexts over the entire 4-second window. This excessive temporal integration dilutes critical pre-freeze signatures by scattering the model's attention across preceding normal walking cycles. Ultimately, the inductive bias of CNNs—prioritizing local spatiotemporal patterns over long-term dependencies—aligns perfectly with the abrupt physiological onset of an impending FOG episode.

This study has several limitations. First, the analysis is based on window-level metrics. Future work should include event-level measures such as event-F1 and onset latency to better reflect clinical usefulness. Second, the 4-second/2-second look-ahead labeling scheme may introduce boundary noise, and the small cohort of 16 subjects limits the characterization of population-level variability. Finally, experiments were conducted on a single public locomotion dataset [12], evaluating cross-dataset generalization, robustness to sensor placement variation, and end-to-end online latency on target hardware. These will be important next steps toward deployment in daily living conditions.

5 Conclusion

We conducted a unified comparison for IMU-based FOG forecasting under subject robustness evaluation. Under LOSO-CV, CNN-based and compact hybrid architectures provided the best overall balance, with top models achieving impending-FOG $F1 \approx 0.84$ and recall ≈ 0.91 across 16 subjects.

A key observation is that increasing model complexity does not improve subject robustness performance beyond CNNs. The deeper CNN, hybrid CNN-LSTM, and LSTM-TCN models, and the compact Transformer all achieved similar or slightly lower LOSO-CV F1 scores despite having more parameters and more expressive temporal modules. This likely reflects the structure of the forecasting task, which is predicting FOG within short 4-s windows from foot-mounted IMUs, relies mainly on local pre-freeze kinematics and short-range temporal patterns, which are already well captured by 1D convolutions. This is consistent with a short-horizon forecasting setup in 4s windows where local pre-freeze kinematics dominate, so 1D convolutions capture most useful structure while higher-capacity recurrent and attention models are more prone to fitting subject-specific variability.

Beyond model choice, practical deployment benefits from onset oriented labeling, threshold calibration to a predefined false-cue budget, and subject-aware evaluation with event-level reporting alongside window-level scores. These choices align performance reporting with how a wearable cueing system would be used. Future work will expand the evaluation to daily living conditions, examine robustness to differences in sensor placement, and quantify end-to-end online latency on target hardware. Overall, the results suggest FOG forecasting can be integrated into wearable robotic systems with modest computational demands. [14–19]

REFERENCES

- [1] Chen, Y., Li, X., Su, H., Yu, H., and Zhang, D., 2022. "Design of a bio-inspired gait phase decoder based on temporal convolution network architecture with contralateral surface electromyography sensory feedback towards hip prosthesis control". *Frontiers in Neurorobotics*, **16**, p. 791169.
- [2] Zhu, J., Jiao, C., Dominguez, I., Yu, S., and Su, H., 2022. "Design and backdrivability modeling of a portable high torque robotic knee prosthesis with intrinsic compliance for agile activities". *IEEE/ASME Transactions on Mechatronics*, **27**(4), pp. 1837–1845.
- [3] Zhang, S., Zhu, J., Huang, T.-H., Yu, S., Huang, J. S., Lopez-Sanchez, I., Devine, T., Abdelhady, M., Zheng, M., Bulea, T. C., and Su, H., 2024. "Actuator optimization and deep learning-based control of pediatric knee exoskeleton for community-based mobility assistance". *Mechatronics*, **97**, p. 103109.
- [4] Luo, S., Jiang, M., Zhang, S., Zhu, J., Yu, S., Silva, I. D., Wang, T., Rouse, E., Zhou, B., Yuk, H., Zhou, X., and Su, H., 2024. "Experiment-free exoskeleton assistance via learning in simulation". *Nature*, **630**, pp. 353–359.
- [5] Huang, T.-H., Zhang, S., Yu, S., MacLean, M., Zhu, J., Di Lallo, A., Jiao, C., Bulea, T. C., Zheng, M., and Su, H., 2022. "Modeling and stiffness-based continuous torque control of lightweight quasi-direct-drive knee exoskeletons for versatile walking assistance". *IEEE Transactions on Robotics*, **38**(3), pp. 1442–1459.
- [6] Di Lallo, A., Yu, S., Huang, T.-H., Bulea, T. C., and Su, H., 2021. "High-performance soft wearable robots for human augmentation and gait rehabilitation". In *Soft Robotics in Rehabilitation*. Elsevier, pp. 1–38.
- [7] Yu, S., Liu, L., Zhang, S., Di Lallo, A., Zhu, J., Wu, Q., Zuo, G., Zhou, X., and Su, H., 2025. "Controlling negative and positive power for efficiency enhancement and muscle strain mitigation during squatting with a portable knee exoskeleton". *Annals of Biomedical Engineering*.
- [8] Rodriguez-Jorge, D., Zhang, S., Huang, J. S., Lopez-Sanchez, I., Srinivasan, N., Zhang, Q., Zhou, X., and Su, H., 2025. "Biomechanics-informed mechatronics design of comfort-centered portable hip exoskeleton: Actuator, wearable interface, controller". *IEEE Transactions on Medical Robotics and Bionics*.
- [9] Di Lallo, A., Yu, S., Slightam, J., Gu, G. X., Yin, J., and Su, H., 2024. "Untethered fluidic engine for high-force soft wearable robots". *Advanced Intelligent Systems*.
- [10] Trujillo-Guerrero, M., Niemes, S., Jaén-Vargas, M., Cadiz, A., Fonseca, R., and Serrano Olmedo, J., 2023. "Accuracy comparison of cnn, lstm, and transformer for activity recognition using imu and visual markers". *IEEE Access*, **PP**, 09, pp. 1–1.
- [11] Bächlin, M., Plotnik, M., Roggen, D., Maidan, I., Hausdorff, J. M., Giladi, N., and Tröster, G., 2010. "Wearable assistant for Parkinson's disease patients with the freezing of gait symptom". *IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society*, **14**(2), Mar., pp. 436–446.
- [12] Dvorani, A., Waldheim, V., Magdalena, Salchow-Hömmen, C., Meyer-Ohle, J., Kühn, A. A., Wenger, N., and Schauer, T., 2021. "Real-time detection of freezing motions in parkinson's patients for adaptive gait phase synchronous cueing". *Frontiers in Neurology*, **12**, Dec.
- [13] Hammerla, N. Y., Halloran, S., and Plötz, T., 2016. "Deep, convolutional, and recurrent models for human activity recognition using wearables". *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1533–1540.
- [14] Yan, Y., Huang, J. S., Zhu, J., Hou, Z., Gao, W., Lopez-Sanchez, I., Srinivasan, N., and Su, H., 2025. "Compact and foldable hip exoskeleton with high torque density actuator for walking and stair-climbing assistance in young and older adults". *IEEE/ASME Transactions on Mechatronics*.
- [15] Yu, S., Huang, T.-H., Wang, D., Lynn, B., Sayd, D., Sili-vanov, V., Park, Y. S., Tian, Y., and Su, H., 2019. "Design and control of a high-torque and highly-backdrivable hybrid soft exoskeleton for knee injury prevention during squatting". *IEEE Robotics and Automation Letters*, **4**(4), pp. 4579–4586.
- [16] Yang, X., Huang, T.-H., Hu, H., Yu, S., Zhang, S., Zhou, X., Carriero, A., Yue, G., and Su, H., 2019. "Spine-inspired continuum soft exoskeleton for stoop lifting assistance". *IEEE Robotics and Automation Letters*, **4**(4), Aug., pp. 4547–4554.
- [17] Wang, J., Li, X., Huang, T.-H., Yu, S., Li, Y., Chen, T., Carriero, A., Oh-Park, M., and Su, H., 2018. "Comfort-centered design of a lightweight and backdrivable knee exoskeleton". *IEEE Robotics and Automation Letters*, **3**(4), Oct., pp. 4265–4272.
- [18] Yu, S., Huang, T.-H., Yang, X., Jiao, C., Yang, J., Chen, Y., Yi, J., and Su, H., 2020. "Quasi-direct drive actuation for a lightweight hip exoskeleton with high backdrivability and high bandwidth". *IEEE/ASME Transactions on Mechatronics*, **25**(4), pp. 1794–1802.
- [19] Luo, S., Androwis, G., Adamovich, S., Nunez, E., Su, H., and Zhou, X., 2023. "Robust walking control of a lower limb rehabilitation exoskeleton coupled with a musculoskeletal model via deep reinforcement learning". *Journal of Neuroengineering and Rehabilitation*, **20**(1), Mar., p. 34.