Adaptive Multi-Agent Feature Selection for Personalized Fall Risk Prevention

Chang Liu CHANG.LIU@UCF.EDU
School of Data, Mathematical, and Statistical Sciences
University of Central Florida
Orlando, FL, USA

Ladda Thiamwong LADDA.THIAMWONG@UCF.EDU
College of Nursing
University of Central Florida
Orlando, FL, USA

Yanjie Fu YANJIE.FU@ASU.EDU
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ, USA

Rui Xie† RUI.XIE@UCF.EDU
School of Data, Mathematical, and Statistical Sciences, College of Nursing
University of Central Florida
Orlando, FL, USA

Abstract

Falls among older adults represent a major public health challenge driven by complex, time-varying interactions across multiple risk domains. Effective fall risk factor identification requires learning from heterogeneous longitudinal data while accounting for sparse and delayed fall-related outcome events. However, existing approaches are largely static and fail to adaptively model evolving, individualized risk factors across modalities and time. We propose **PAFIR**, a **P**ersonalized and **A**daptive **F**eature selection framework for fall risk **I**dentification and p**R**evention, which formulates adaptive feature selection as a reinforcement learning problem over longitudinal multimodal health data. PAFIR jointly models structural dependencies among correlated assessment variables and temporal dynamics in wearable-derived physical activity data, and learns adaptive selection policies across repeated study visits using reward signals derived from sparse fall incidence outcomes. We apply PAFIR to data from the Physio fEedback Exercise pRogram (PEER) cluster-randomized trial. Experimental results demonstrate that PAFIR more effectively captures longitudinal and structural patterns of feature relevance than state-of-the-art baselines, and enables dynamic, subject-specific feature selection. By adapting selected features over time, PAFIR supports more timely and personalized fall prevention strategies.

1. Introduction

Fall prevention among older adults remains an urgent and complex public health challenge. In the United States, falls are the leading cause of fatal and non-fatal injuries among adults

†. Corresponding author.

aged 65 and older, resulting in over 38,000 deaths annually and substantial healthcare costs (Centers for Disease Control and Prevention, 2026; Florence et al., 2018). Beyond acute injury, falls often trigger long-term functional decline, loss of independence, and diminished quality of life (Paliwal et al., 2017). As populations age and care resources become increasingly constrained, there is a critical need for timely, scalable, and personalized strategies to identify fall risk and intervene before adverse events occur.

Recent advances in mobile and smart health technologies enable continuous monitoring of gait, activity, balance, and contextual behaviors, creating new opportunities for early fall risk detection (Pfortmueller et al., 2014; Giovannini et al., 2022; Mortazavi et al., 2023). However, translating these heterogeneous, high-dimensional data sources, which include continuous sensor streams, multi-time-scale signals, and structured survey-based measurements, into effective prevention remains challenging.

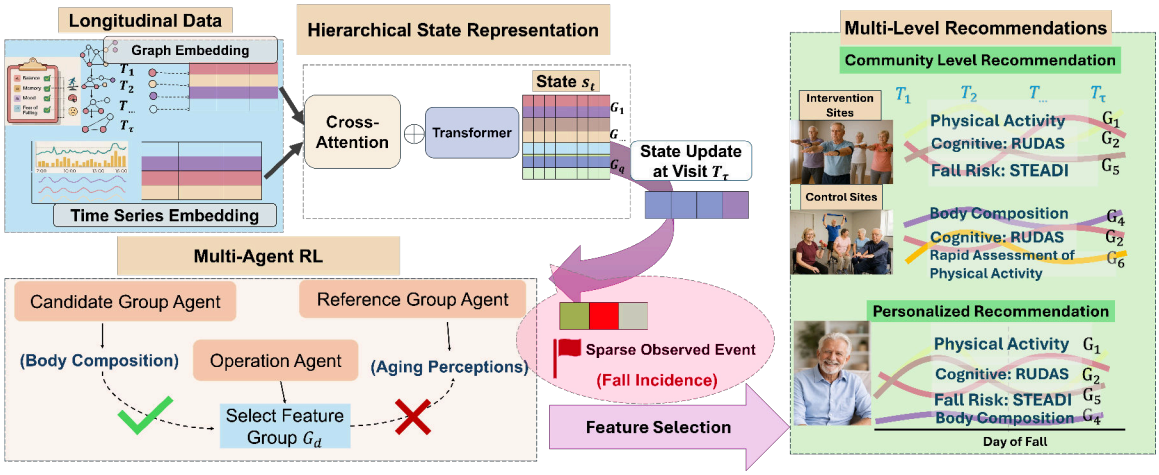


Figure 1: Overview of the **PAFIR** framework. Heterogeneous **longitudinal data** from multiple modalities are integrated to construct a **hierarchical state representation** of the individual state. This state is iteratively updated across clinical visits and in response to sparsely observed events, i.e., fall incidence. A **multi-agent reinforcement learning** module then operates on the evolving state to perform adaptive **feature selection** at both group and individual levels. The resulting learned policies support **personalized and community-level recommendations** for fall risk identification and prevention.

A critical insight, however, is that fall risk is inherently dynamic and highly individualized, shaped by both objective physical capacity and subjective risk perception. Older adults may underestimate their physiological vulnerability or, conversely, restrict activity due to fear of falling despite preserved function (Duncan et al., 1993; Delbaere et al., 2010; Cuevas-Trisan, 2017; Liu et al., 2025b). Such misalignment between the *Body* (mobility, balance) and the *Mind* (confidence, cognition, and risk perception) can lead to maladaptive behaviors, delayed risk recognition, and reduced engagement with preventive care (He et al., 2022; Wang et al., 2024; Maruszewska et al., 2025). These complexities make effective prevention not only a problem of detecting risk signals, but also of adaptively determining *which* features are most informative and *when* to intervene for each individual (Dautzenberg et al., 2021; Pillay et al., 2024).

Although many individual fall risk factors have been studied, how physiological, psychological, and behavioral factors interact and co-evolve over time remains poorly understood. Most existing approaches model these factors in isolation or assume static relationships across visits (Gong et al., 2024; Yassine et al., 2021; Theng and Bhoyar, 2024), limiting their ability to capture dynamic longitudinal interactions between physical capacity, risk perception, and behavior. The Physio fEedback Exercise pRogram (PEER) study (Thiamwong et al., 2023a) is a technology-based intervention, specifically, designed to address this mismatch, integrating real-time physio-feedback, cognitive reframing, and peer-led exercise to jointly engage physical and psychological fall risk domains. This study offers rich, longitudinal multimodal data spanning structured survey-based assessments, clinical measurements, fall incidence, and wearable sensor-derived time-series data, as visualized in Fig. 2, providing a comprehensive data foundation for data-driven fall risk identification and prevention. This gap motivates data-driven methods that integrate multimodal signals and adaptively model evolving risk factor interactions, framing personalized fall risk identification as an adaptive feature selection problem over heterogeneous physiological, behavioral, and contextual data, as illustrated in the overview of the proposed PAFIR framework in Fig. 1.

Emerging developments in reinforcement learning (RL) for healthcare have shown strong promise for adaptive feature selection as a sequential decision process, enabling models to update feature relevance based on observations and clinical feedback (Yu et al., 2021; Komorowski et al., 2018). Despite these advances, existing feature selection systems continue to face fundamental limitations when deployed in real-world healthcare settings, particularly for fall prevention.

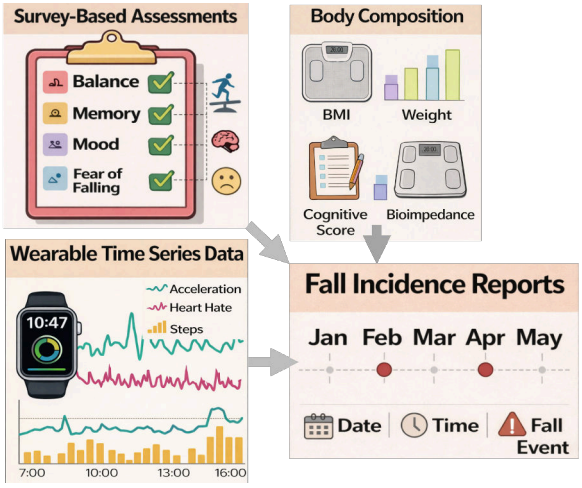


Figure 2: Illustration of the multimodal longitudinal PEER dataset across visits (T1-T4), including (1) survey-based fall risk assessments, (2) high-frequency wearable time-series data capturing physical activity, and (3) instrument-based measurements such as body composition. **Fall incidence**, the primary outcome, is recorded as sparse, time-stamped events.

A primary challenge in fall risk factor identification arises from the integration of highly diverse data sources, including self-reported surveys (Ritchey et al., 2022), body composition measurements (Nguyen et al., 2024), muscle strength and balance evaluations (McManus et al., 2022), daily physical activity captured via wearable devices (Howcroft et al., 2017; Liu et al., 2025a), and gait and posture assessments (Lim et al., 2024). Most existing studies focus on only a subset of these data types or rely on simplified fusion strategies (González-Castro et al., 2024), limiting their ability to capture cross-modal interactions and evolving risk patterns.

Additionally, fall risk modeling is further complicated by the dynamic and heterogeneous nature of activity and risk signals in older adults. Wearable sensors produce continuous, minute-level physical activity data (Alsadoon et al., 2024), while fall outcomes are sparse,

delayed, and irregular (Tonchoy et al., 2024; Li and Surineni, 2025), complicating temporal alignment and longitudinal analysis. Gradual changes in mobility, cognition, and psychological state may develop over extended periods, whereas acute events can induce abrupt shifts in risk. However, learning from sparse outcome signals for fall incidence remains a major challenge. For example, in the PEER study (Thiamwong et al., 2023a), fall events are inherently rare, with only 69 events observed across 1364 visit-level observations and many participants experiencing no falls throughout the entire trial. When the reward is defined solely based on the binary occurrence of a fall, the RL agent receives informative feedback in only a small fraction of training steps (Wang et al., 2020; Ecoffet et al., 2021), corresponding to a setting with limited and delayed feedback where informative signals are difficult to propagate across time (Wang et al., 2023).

Fall-risk research has employed a broad range of statistical and machine-learning methods for identifying relevant risk factors. Marginal filtering and importance-ranking approaches typically evaluate variables separately, whereas multivariable methods, including LASSO (Tibshirani, 1996) and Group LASSO (Yuan and Lin, 2006), perform joint selection under sparsity constraints. Established longitudinal and survival-analysis approaches can also accommodate repeated measurements and event outcomes. For example, joint models (Wang and Zhong, 2025; Elashoff et al., 2016; Crowther et al., 2016) link longitudinal biomarker trajectories with time-to-event outcomes, while frailty models (Hougaard, 1995) account for unobserved heterogeneity. These methods are particularly appropriate when the primary objective is to estimate time to fall, covariate effects, or the probability of falling within a prespecified prediction horizon (Pepe, 2003; Suresh et al., 2022).

To address the challenges above mentioned, we introduce **PAFIR**, a **P**ersonalized and **A**daptive **F**eature Selection Fall Risk **I**dentification and p**R**evention framework (Fig. 1) that formulates fall risk factor identification as a multi-agent reinforcement learning feature selection problem over longitudinal multimodal health data. PAFIR models heterogeneous clinical assessments and high-resolution activity data while accounting for temporal dynamics and individual variability. By jointly capturing structural relationships among risk factors and their evolution over time, PAFIR learns adaptive feature relevance from longitudinal data. An RL policy guided by sparse fall outcomes and proxy fall risk appraisal measures enables personalized and group-aware identification of evolving fall risk factors to support timely prevention.

Contributions Overview

- **Hierarchical structure for fall risk feature selection in the clinical trial**: We embedded a hierarchical structure in a topology-aware reinforcement learning framework to effectively capture and represent multi-level dependencies among features, which is crucial for fall risk screening, where factors are naturally organized hierarchically. For example, within fear-of-falling assessments, questionnaire items measuring activity avoidance and balance confidence belong to the same higher-level domain while reflecting distinct behavioral and perceptual contributors to fall risk.
- **Adaptive comparison-driven feature selection mechanism**: We propose an adaptive comparison-driven feature selection mechanism that decomposes the selection process into candidate generation, reference-based comparison, and iterative refinement. Instead of evaluating features independently, the method assesses their incremental contribution

by comparing candidate feature groups against reference groups, enabling more reliable fall risk factor identification under complex healthcare settings.

- **Sparse-aware rewards design with clinical proxies**: We introduce a sparse-aware reward design that augments sparse fall-event signals with clinically informed proxy outcomes (i.e., FES-I and BBS), allowing the PAFIR framework to capture broader fall-risk patterns and improve the robustness of feature selection.

Generalizable Insights about Machine Learning in the Context of Healthcare The proposed framework can generalize to settings that require identifying clinically relevant risk factors from heterogeneous and multimodal data, including structured clinical assessments, behavioral measures, and high-frequency sensor signals. Such settings are common in healthcare, where patient information is collected across diverse modalities and scales, and effective modeling requires integrating these sources while preserving their structural relationships. Moreover, the approach accommodates both discrete and continuous outcomes, ranging from binary events (e.g., disease onset or adverse events) to continuous measures (e.g., functional decline or physiological indicators), enabling a unified and flexible framework for modeling feature relevance across diverse clinical settings. In addition, the framework naturally extends to sequential decision-making settings, where feature relevance evolves over time and depends on prior observations and selections. This is particularly important in longitudinal healthcare data, where risk factors may change gradually or abruptly, and adaptive modeling is required to capture temporally evolving patterns. Finally, the use of informative feedback mechanisms, such as incorporating proxy or auxiliary signals, enables effective learning in scenarios with sparse or delayed outcomes. Such conditions are prevalent in healthcare applications, where clinically meaningful events are often rare, and leveraging additional signals can improve stability and guide learning toward meaningful trajectories. Together, these insights suggest that integrating multimodal data, supporting diverse outcome types, and enabling adaptive sequential learning are key to developing generalizable machine learning methods for healthcare.

2. Related Work

Fall-risk assessment traditionally relies on clinically interpretable screening and functional measures, including the Timed Up and Go test, balance assessments, and STEADI-related indicators. With the increasing availability of wearable sensors and longitudinal cohorts (Picerno et al., 2021), machine-learning studies have incorporated physical activity, gait, cognitive, psychological, and functional features for broader risk characterization (Fan et al., 2024). However, many existing applications still use fixed feature sets or collapse longitudinal observations into visit-level summaries (Chantanachai et al., 2021; Rykov et al., 2021).

Statistical methods provide an established framework for longitudinal fall-outcome analysis. Time-varying and recurrent-event survival models accommodate changing covariates and repeated falls, frailty models account for unobserved heterogeneity (Hougaard, 1995), joint models link longitudinal biomarker trajectories with event outcomes (Wang and Zhong, 2025; Elashoff et al., 2016; Crowther et al., 2016), and competing-risk models address terminal events such as death (Li et al., 2022). Feature selection can be incorporated through LASSO, Group LASSO, longitudinal penalization, or boosting (Tibshirani, 1996; Yuan and Lin, 2006; Xu et al., 2015; Dong et al., 2022). These approaches are well suited to es-

timating covariate effects, event risk, or survival probabilities under a specified outcome model (Kalbfleisch and Schaubel, 2023; Suresh et al., 2022; Smith et al., 2022). PAFIR instead focuses on adaptively updating feature sets from visit-level assessments and wearable sequences as new fall or proxy feedback becomes available.

Healthcare data are inherently multimodal, irregular, and hierarchically structured, combining sparse events, wearable sensor streams, and structured assessments (Ghat, 2023; Franklin et al., 2024). Many clinical surveys exhibit hierarchical organization, such as cognitive domains composed of multiple subscales (Rosellini and Brown, 2021), yet most models ignore these dependencies, leading to information loss (Mukherjee et al., 2023), particularly in longitudinal settings. Recent work on hierarchical structural encoding (Luo et al., 2024) motivates the need for personalized, temporally adaptive modeling, which our framework aims to address.

To address these limitations, recent work has explored dynamic and adaptive feature selection strategies. Reinforcement learning has emerged as a promising approach for dynamic feature selection and sequential decision-making in healthcare, with applications such as adaptive screening (Yu et al., 2021) and treatment optimization in critical care (Komorowski et al., 2018). For feature selection, TTG (Khurana et al., 2018) formulates feature transformation as a graph construction problem and applies RL to search for optimal subsets, but lacks temporal modeling and assumes static relationships. Structure-Aware Transformer (SAT) (Chen et al., 2022) improves graph-based modeling by extracting node-centric subgraphs before attention computation, yet it does not capture evolving longitudinal or personalized dynamics. Topology-Aware Reinforcement Learning (TAR) (Ying et al., 2024) introduces dynamic graph adaptation through reinforcement-based feature space reconstruction. However, existing adaptive feature-selection methods provide important foundations for graph-based or sequential feature-space optimization, but they are not designed specifically for the combination considered here: clinically defined hierarchical feature groups, visit-level structured assessments, minute-level wearable sequences, asynchronous information updates, and sparse fall outcomes supplemented by proxy feedback. PAFIR integrates these elements within a unified adaptive feature-selection framework for longitudinal multimodal health data.

3. Clinical Trial for Fall Prevention: Physio-fEedback Exercise pRogram

We are motivated by the heterogeneous and longitudinal data structure (Fig.2) from the **P**hysio-f**E**edback **E**xercise p**R**ogram (PEER) study (Thiamwong et al., 2023a), a two-arm clustered randomized controlled clinical trial designed to evaluate a technology-assisted, body and mind intervention for fall prevention and physical activity promotion among older adults in a free-living environment (clinicaltrials.gov ID: NCT05778604). The analyzed cohort includes 341 participants, aged 61 to 89 years, from 15 senior living sites in Central Florida. In this two-arm cluster randomized trial, senior living centers were randomly assigned to either the PEER intervention arm or the control arm, with all participants within each center receiving the same allocation.

Each participant completed four clinical assessments: baseline at week 0 (T1), post-intervention at week 9 (T2), 2 months (60 days) post-intervention (T3), and 6 months post-intervention (T4). All assessments and wearable-monitoring periods followed the scheduled

visit protocol and were not triggered by fall events. Following each scheduled visit, participants were asked to wear the accelerometer for seven consecutive days, regardless of whether a fall had occurred. At each visit, participants contributed three complementary categories of data. First, **structured survey-based assessments** captured physical, cognitive, and functional fall risk factors using validated instruments (e.g., the 7-item Short Fall Efficacy Scale), resulting in low-frequency, visit-level tabular measurements. Second, **clinical measurements**, including body composition, grip strength, and standardized evaluations of muscle strength and balance, provided objective physiological measures at each visit. Third, to characterize physical activity in naturalistic settings, participants wore ActiGraph triaxial accelerometers for seven consecutive days following each visit, generating high-frequency, minute-level **time-series physical activity data** such as step counts, vector magnitude (VM), and posture states (e.g., sitting, standing, and lying). ***Fall incidence*** served as the primary outcome and was recorded as self-reported, time-stamped events, resulting in sparse and irregular event sequences for most participants. No participant in the analyzed cohort died before the end of follow-up; therefore, death-related competing-risk censoring did not arise in the current experiments. Together, these multimodal data spanning survey-based, clinical, and wearable-derived measurements across multiple temporal resolutions enable the identification of personalized and dynamically evolving fall risk factors, supporting adaptive fall risk identification and prevention.

4. PAFIR Framework

The **P**ersonalized and **A**daptive **F**eature selection framework for fall risk **I**dentification and p**R**evention (PAFIR; Fig. 1) is designed to tackle the challenges of selecting effective fall risk factors from longitudinal, multimodal health data and developing personalized prevention strategies for healthcare applications. PAFIR integrates three key innovations: (1) hierarchical structure linking feature groups to individual features; (2) a comparison-driven feature selection mechanism that iteratively contrasts candidate and reference feature groups to refine the identification of clinically relevant features; and (3) a sparse-aware reward design with clinical proxies. PAFIR optimizes RL policies for outcome-relevant feature selection, enhancing the interpretability and effectiveness of longitudinal health modeling in real-world, dynamic care environments. Formally, let i index participants and let $\mathcal{X}$ denote the complete feature space, including structured clinical variables and wearable-derived temporal features. PAFIR learns a feature-selection policy shared across the training population and applies it to each participant-specific state to produce an individualized selected feature set. The index t denotes an asynchronous information update. Let $\mathbf{s}_{i,t}$ denote the multimodal state available for participant i at update t, and let $F_{i,t-1} \subseteq \mathcal{X}$ denote the active feature set carried into that update. The three coordinated agent policies define a distribution over the composite feature-selection action, $\pi_\theta(a_{i,t} \mid \mathbf{s}_{i,t}, F_{i,t-1})$, where $a_{i,t}$ consists of candidate selection, reference selection, and an add, remove, or retain operation. Applying $a_{i,t}$ produces an updated feature set $F_{i,t}$, which may contain more, fewer, or the same number of features as $F_{i,t-1}$. At each update, $y_{i,t}$ denotes the available outcome signal: the observed fall outcome for a fall-related update, or a visit-level clinical proxy, such as FES-I or BBS, for a scheduled assessment update.

4.1. Hierarchical Multimodal Structure for Fall Risk Feature Selection

To effectively model both structural clinical assessments and temporal dynamics in time-series wearable sensor data, we construct a state representation $\mathbf{s}_{i,t}$ that integrates structured clinical features and time-series physical activity signals. This design captures both the organization of clinical variables and the evolution of patient status across visits.

Hierarchical Structural Representation Longitudinal clinical data and time-series physical activity signals form a structured and multimodal representation of patient health. Features are organized into clinically meaningful feature groups based on their functional roles. For example, the body composition group includes variables such as BMI and weight, the physical functional group includes the measurements of grip strength, and the psychological group includes assessments such as the FES, which captures fear of falling. Features within each group are often correlated and reflect related aspects of participant health, while relationships across groups capture interactions among broader fall-risk domains. Deterioration in physical performance often co-occurs with elevated fear of falling, and cognitive decline frequently accompanies functional limitation. Capturing both levels is essential for identifying which risk factors are clinically relevant and how they jointly evolve over time. We partition the feature space $\mathcal{X}$ into M clinically meaningful feature sets, $\{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_M\}$, where each $\mathcal{V}_m \subseteq \mathcal{X}$ consists of a subset of related features, and $\bigcup_{m=1}^{M} \mathcal{V}_m = \mathcal{X}$.

Within each clinically defined group, features are represented as nodes in an intra-group graph $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m)$. The feature groups are predefined according to the PEER assessment domains, and every pair of features within the same domain is connected, forming a fully connected intra-group graph. This construction encodes shared clinical-domain membership and enables related features to be evaluated jointly.

Across groups, we further define an inter-group topology $\mathcal{T} = (\mathcal{U}, \mathcal{W})$, where each node $u_m \in \mathcal{U}$ represents a feature group $\mathcal{V}_m$. For participant i at update t, let $\Phi_{i,t}^{(m)} = \sum_{v \in \mathcal{V}_m} \phi_{i,t,v}$ denote the initial importance score of group m, where $\phi_{i,t,v}$ denotes the importance of feature v. These initial importance scores are used to construct the initial structural representation and to initialize the reward reference and feature-selection process. Each weighted edge $w_{mn} \in \mathcal{W}$ reflects the co-variation in initial importance between groups m and n, that is, the extent to which their aggregated initial importance scores increase or decrease together across participants and updates, $w_{mn} = \mathrm{Corr}_{(i,t)\in\mathcal{D}_{\text{train}}}\left(\Phi_{i,t}^{(m)}, \Phi_{i,t}^{(n)}\right)$, where the correlation is computed across available participant-update observations in the training data.

For each feature $v \in \mathcal{V}_m$, we construct a structural representation $\mathbf{h}_{i,t,v}$ that captures its relationships with other features in the same clinical domain for participant i at update t. Let $\mathbf{H}_{i,t} = \{\mathbf{h}_{i,t,v} : v \in \mathcal{V}_m,\ m = 1, \ldots, M\}$ denote the collection of structural feature representations at update t. These representations summarize within-group dependencies and group-level organization, providing a compact encoding of the structured feature space. Detailed descriptions of the encoding procedure are provided in Appendix C.1.

Temporal Representation In addition to structured clinical features, physical activity data are inherently time-series observations, capturing fine-grained behavioral dynamics across time. While clinical variables are typically recorded at discrete visits, physical activity signals provide continuous measurements between visits, offering complementary information about patient behavior. Let $\mathbf{X}_{i,:,b}^{(\tau)} \in \mathbb{R}^{L\times 1}$ denote the time-series data collected

for participant i following visit τ, where L is the sequence length and b denotes the corresponding temporal feature. At the corresponding update t, each time series is mapped to a latent representation $\mathbf{z}_{i,t,b}$ that captures temporal patterns such as step counts and physical activity vector magnitude. A detailed description of the construction of the temporal representation $\mathbf{z}_{i,t,b}$ is provided in Appendix C.2.

State Construction Let $\mathbf{Z}_{i,t} = \{\mathbf{z}_{i,t,b} : b \in \mathcal{B}\}$ denote the set of temporal embeddings available for participant i at update t, where $\mathcal{B}$ denotes the set of time-series physical activity features. The final state representation is constructed by integrating structural and temporal information,

$$\mathbf{s}_{i,t} = f\left(\mathbf{H}_{i,t}, \mathbf{Z}_{i,t}\right), \tag{1}$$

where $f(\cdot)$ represents a fusion mechanism that aligns structured features with temporal dynamics. This formulation allows $\mathbf{s}_{i,t}$ to capture both the topology of the feature space and the temporal evolution of patient behavior across visits, supporting more stable and context-aware feature selection.

4.2. Adaptive Comparison-Driven Feature Selection Mechanism

We design an adaptive comparison-driven feature selection mechanism to iteratively identify outcome-relevant features under complex and correlated clinical settings. The key idea is to decompose feature selection into a sequence of structured decisions that balance candidate discovery, comparative evaluation, and refinement (Zhong et al., 2012). To implement this process, we adopt a three-agent feature selection framework (Ying et al., 2024), where each agent is responsible for a distinct role in the stepwise selection procedure. Specifically, the candidate agent proposes a feature group for evaluation, the reference agent selects a competing group for comparison, and the operation agent determines whether the active feature set should be updated based on their relative outcome relevance (in Fig.1). A single agent would need to select the candidate group, reference group, and operation jointly. The three-agent formulation instead factorizes this action into smaller conditional decisions. The agents are executed sequentially, share the same participant-specific state and reward, and condition each downstream decision on the preceding actions. They therefore form a coordinated feature-selection policy rather than independent decision makers.

Let $F_{i,t-1} \subseteq \mathcal{X}$ denote the active feature set for participant i before update t, and let $\mathcal{C} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_M\}$ denote the collection of feature groups derived from the hierarchical structure. At each update, the candidate agent selects a feature group based on its potential relevance to the outcome, $C_{i,t}^{\text{cand}} \sim \pi_{\theta_1}\left(C \mid \mathbf{s}_{i,t}, F_{i,t-1}\right), \quad C_{i,t}^{\text{cand}} \in \mathcal{C}$. Rather than evaluating the candidate group in isolation, the reference agent selects a comparison group from the remaining feature space, $C_{i,t}^{\text{ref}} \sim \pi_{\theta_2}\left(C \mid \mathbf{s}_{i,t}, F_{i,t-1}, C_{i,t}^{\text{cand}}\right), \; C_{i,t}^{\text{ref}} \in \mathcal{C} \setminus \{C_{i,t}^{\text{cand}}\}$. The operation agent then selects an add, remove, or retain operation, $o_{i,t} \sim \pi_{\theta_3}\left(o \mid \mathbf{s}_{i,t}, F_{i,t-1}, C_{i,t}^{\text{cand}}, C_{i,t}^{\text{ref}}\right)$, where $o_{i,t} \in \{\text{add}, \text{remove}, \text{retain}\}$. The resulting composite action is $a_{i,t} = \left(C_{i,t}^{\text{cand}}, C_{i,t}^{\text{ref}}, o_{i,t}\right)$.

The candidate and reference groups are evaluated relative to the current active feature set using the available outcome feedback. The same comparison procedure is subsequently

applied within each retained group to refine the selection at the individual-feature level while preserving the hierarchical structure.

4.3. Sparse-Aware Reward Design With Clinical Proxies

Learning an effective feature selection policy from fall incidence is challenging due to the sparsity and temporal irregularity of observed events. In many cases, participants do not experience a fall across multiple visits, resulting in limited and delayed feedback if rewards are defined solely based on fall occurrence. To address this, we design a dynamic reward mechanism that provides continuous and adaptive feedback throughout the selection process. At the beginning of training, baseline feature-importance scores are used to provide an informed initial reward reference and selection prior. During training, the reward is computed using the outcome information available at each update. In addition to fall incidence, the reward signal incorporates clinically validated proxy measures that are available at every visit and are known to be associated with fall risk, such as the Falls Efficacy Scale-International (FES-I, Yardley et al. (2005)) and the BTrackS Balance Tracking System score (BBS, Levy et al. (2018)). These proxy signals, as fall risk appraisal, provide continuous supervision even in the absence of observed falls, allowing the model to receive meaningful feedback at every step (Thiamwong et al., 2020a).

Specifically, when a fall incident occurs between visits (e.g., between Visit 2 and Visit 3), the reward is immediately adjusted to reflect the contribution of the selected features. Feature groups that are more strongly associated with fall risk receive higher rewards, while those that do not contribute are penalized. At scheduled assessment updates without fall feedback, the reward is evaluated using an available clinical proxy, such as FES-I or BBS. This design ensures that rare but clinically important fall events provide strong learning signals when they occur, while proxy-based signals maintain stable feedback throughout the remaining visits. The reward function is defined as

$$r_{i,t} = \mathrm{Perf}\left(F_{i,t}, y_{i,t}\right) - p_{i,t-1}, \tag{2}$$

where $\mathrm{Perf}(F_{i,t}, y_{i,t})$ denotes the evaluation measure used for the fall outcome when fall feedback is available and for the clinical proxy outcome otherwise. The term $p_{i,t-1}$ denotes the best previously observed performance for the corresponding type of outcome feedback, following Zhong et al. (2012).

Each agent is trained using a separate deep Q-network. For agent $j \in \{\mathrm{cand}, \mathrm{ref}, \mathrm{op}\}$, the temporal-difference loss at update t is,

$$\mathcal{L}_{i,t}^{(j)} = \left(Q_j\left(\mathbf{s}_{i,t}, a_{i,t}^{(j)}\right) - \left[r_{i,t} + \gamma \max_{a'} Q_j\left(\mathbf{s}_{i,t+1}, a'\right) \right] \right)^2, \tag{3}$$

where Q_j denotes the action-value function of agent j, $a_{i,t}^{(j)}$ denotes the corresponding agent action, $r_{i,t}$ is the shared reward, and γ is the discount factor. All Q-networks are randomly initialized and trained using the temporal-difference loss. The baseline feature importance scores are used to initialize the reward and the initial feature selection.

This reward design allows the model to learn from both sparse fall events and continuous clinical signals, making the learning process more stable while maintaining alignment with clinically meaningful patterns.

5. In-Field Experiment Evaluations

PEER Fall Prevention Data. We evaluate PAFIR using longitudinal, multimodal data from the PEER study collected in real-world settings, emphasizing adaptive feature selection across repeated clinical visits. The dataset includes 341 community-dwelling participants assessed at four time points, yielding 1,364 participant–visit observations. Following each scheduled visit, participants wore sensors continuously for seven consecutive days, producing minute-level wearable data and totaling over 2.5 million time-stamped observations across all participants and visits.

At each scheduled visit, participants contribute high-dimensional structured clinical assessments, spanning physical, cognitive, psychological, and functional domains. After preprocessing, the structured modality comprises 587 fall-related features organized into 35 clinically meaningful feature groups, including body composition, physical activity, cognitive function, balance confidence, and psychological status. Each participant–visit instance is modeled as a graph-structured input, which is encoded by a graph encoder to capture hierarchical and correlational dependencies among risk factors.

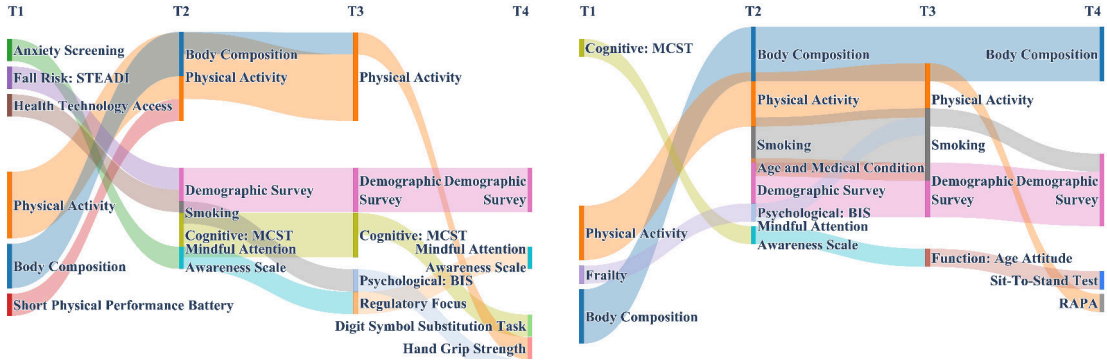
In parallel, minute-level wearable sensor data, including step counts, vector magnitude (VM), and posture states (sitting, standing, and lying), are processed by a time-series encoder to model temporal dynamics within each seven-day monitoring period. The resulting temporal representations are aligned with visit-level structured embeddings and fused via a cross-attention Transformer module, yielding a hierarchical latent state that integrates structural and temporal information and serves as the input to the reinforcement learning component.

Fall incidence events are recorded as self-reported, time-stamped outcomes, providing sparse and delayed supervision signals that guide adaptive learning of feature relevance across visits. The proxy outcomes, namely the FES-I and BBS scores, are assessed at each clinical visit.

PAFIR Implementation. Reinforcement learning feature selection operates at the visit level or when a fall incidence occurs, enabling the model to update feature importance dynamically as new temporal and structural information becomes available. The structured feature space is encoded using a six-layer graph encoder with a hidden dimension of 128, residual connections, and a dropout rate of 0.5. Mean pooling and Laplacian positional encoding with up to ten eigenvectors are applied to aggregate node-level representations. The temporal encoder uses four stacked Transformer encoder blocks, each with four attention heads and a feed-forward dimension of 256, with wearable inputs projected through a shared linear embedding layer.

The model is trained for 100 epochs using the Adam optimizer with a learning rate of 0.001. The Q-network of each agent is optimized using the temporal-difference loss defined in the previous section. Binary cross-entropy loss is used for the fall-outcome classification model employed in the fall-based reward calculation. All experiments are conducted on NVIDIA Quadro RTX 6000 GPUs using CUDA 13.0. Each random-seed run requires approximately 101 seconds, and the complete 20-seed longitudinal stability analysis requires approximately 34 minutes. The peak allocated GPU memory is approximately 240 MB. Computational resource requirements are reported in Appendix B.5 and Table 6.

5.1. Results: Recommendations for Fall Prevention in PEER Study



(a) PEER intervention site: *Kinneret* (b) Control site: *LCA*

Figure 3: Top fall risk factors identified across four visits at the community level.

Community-Level Recommendations for Fall Prevention. We first describe how PAFIR produces the longitudinal community-level results shown in Fig. 3. PAFIR is applied to multimodal data collected at four visits (T1–T4), including clinical assessments, survey-based measures, and wearable-derived physical activity features. The model is trained using data from all participants at the community level. At each visit, PAFIR selects important fall risk factors by considering both changes over time and relationships between feature groups. This procedure produces visit-level feature selection results for each site, allowing us to examine how selected fall risk factors evolve over time in both the intervention and control communities.

Across both sites, PAFIR consistently identifies *physical activity*, *body composition*, and *demographic information* across visits, which aligns with established evidence that mobility-related measures and body composition are among the most persistent predictors of falls in community-dwelling older adults (Thiamwong et al., 2023b; Kohler-Voinov et al., 2025). At the Kinneret intervention site (Fig. 3(a)), PAFIR identifies a broader set of fall risk features beyond physical performance, including psychological and behavioral measures such as *anxiety screening*, *mindful attention awareness*, *regulatory focus*, and *psychological inhibition sensitivity* (BIS), which are selected across multiple visits. This pattern is consistent with prior evidence that psychological factors, including fear of falling, attentional control, and behavioral regulation, are independently associated with fall risk in community-dwelling older adults (Yi et al., 2022; Sturnieks et al., 2025). The consistent selection of psychological and behavioral features at Kinneret across all four visits suggests that PEER's cognitive and psychological components actively engage these domains as modifiable risk factors (Thiamwong et al., 2023a, 2020b), making them detectable and trackable by PAFIR throughout the intervention period. This is in line with evidence that multicomponent interventions combining physical exercise with cognitive-behavioral elements produce greater reductions in fall risk than single-component physical programs alone (Liu et al., 2025b). In contrast, the LCA control site (Fig. 3(b)) predominantly selects traditional physical and functional risk factors, including *frailty*, *physical activity*, *body composition*, *age-related medical conditions*, and *functional performance measures* (Sit-to-Stand, RAPA), across visits. This

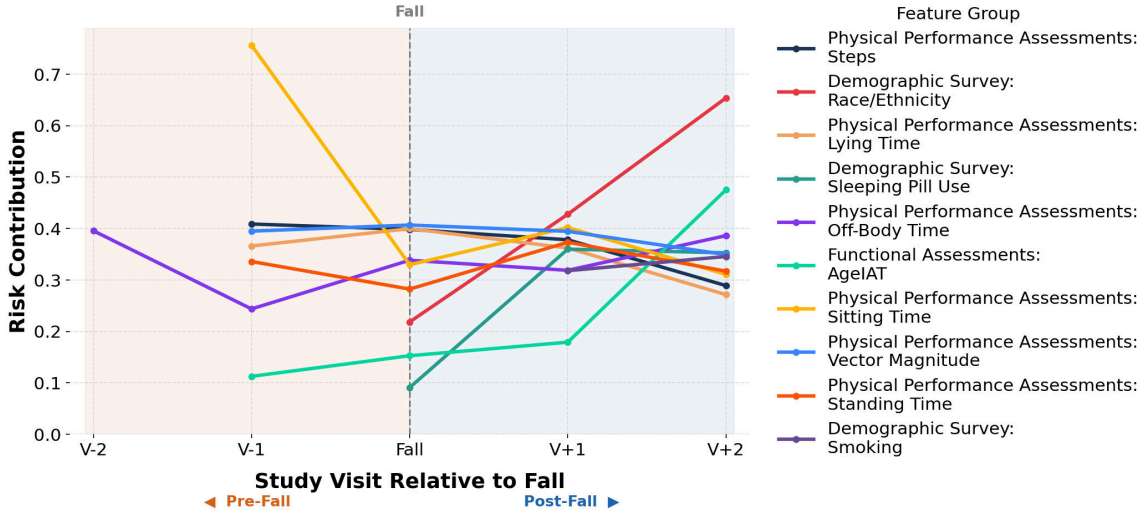


Figure 4: Population-level fall risk contributions across study visits relative to a fall event. Visits are aligned by their order before and after the fall, and the vertical dashed line indicates the fall event. Markers at the dashed line denote the fall-related update, whereas the remaining markers denote scheduled visit-level contributions.

pattern reflects the well-documented role of physical frailty, muscle function, and chronic health conditions as core fall risk determinants in older adult populations not receiving structured behavioral interventions (Chittrakul et al., 2020; Saunders et al., 2025). The absence of psychological and behavioral features at LCA reflects the profile expected in a population not receiving structured multidimensional intervention, where conventional physical and functional factors dominate the fall risk landscape (Liu et al., 2025a; Nguyen et al., 2024). The divergence in feature profiles between the two sites provides indirect evidence of PEER's effectiveness, demonstrating that an effective intervention not only addresses physical fall risk but also renders psychological and behavioral risk factors consistently detectable over time (Liu et al., 2025b). These findings reinforce the clinical value of integrating psychological engagement and cognitive reframing into community-level fall prevention, complementing physical activity and body composition monitoring as a more comprehensive risk management strategy.

Temporal Evolving of Fall Risk Factors We next examine how model-derived feature contributions vary across visits relative to a recorded fall event. Participant trajectories are aligned according to the visit order before and after the fall, with V−2 and V−1 representing pre-fall visits and V+1 and V+2 representing post-fall visits. As shown in Fig. 4, several features exhibit distinct temporal patterns. Activity-related features, including *Steps*, *Vector Magnitude*, *Lying Time*, *Off-Body Time*, and *Sitting Time*, receive relatively high contributions before or around the fall, consistent with the established importance of mobility and activity patterns in fall-risk characterization (Thiamwong et al., 2023a; Taheri et al., 2025). In contrast, the contributions of *Race/Ethnicity*, *Sleeping Pill Use*, *AgeIAT*, and *Smoking* increase across later visits, suggesting that demographic, behavioral, and functional factors remain relevant during post-fall follow-up (Xu et al., 2019; Colón-Emeric et al., 2024; Thiamwong et al., 2023b). The increasing contribution of *AgeIAT* may also reflect the broader

role of cognitive and psychological factors in fall-related outcomes (Sturnieks et al., 2025). These patterns represent changes in the relevance assigned to each feature by PAFIR rather than changes in the underlying raw feature values, illustrating how the framework updates feature priorities across longitudinal assessments.

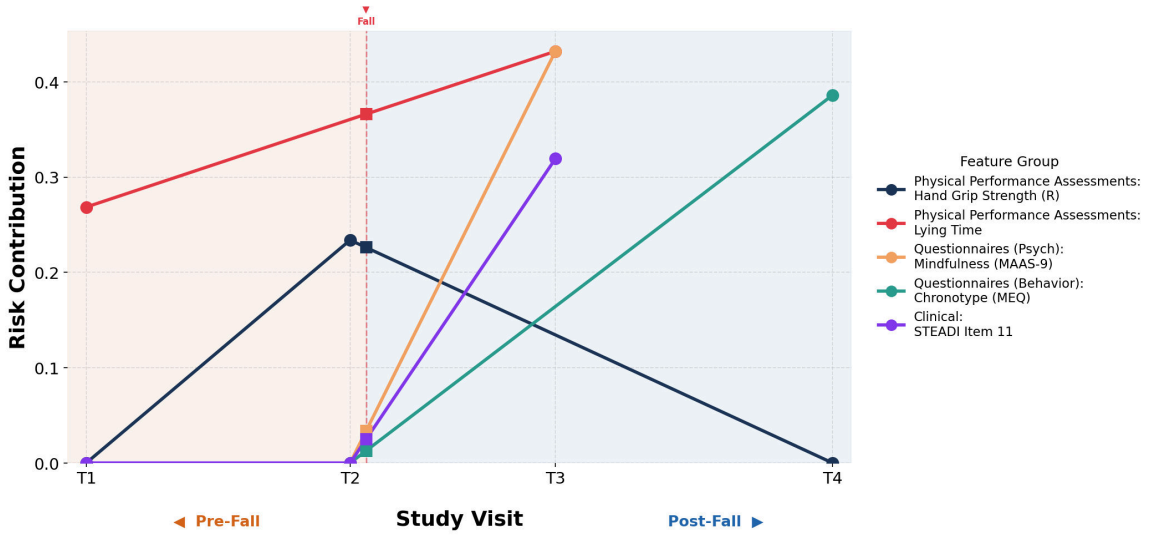


Figure 5: Personalized Temporal Evolution of Feature Importance Before and After Fall for *Subject 1017*. The vertical dashed line indicates the recorded fall event. Circles denote scheduled visit-level contributions, whereas squares denote contributions at the fall-related update.

Personalized Recommendation for Fall Prevention We next present a personalized analysis to illustrate how PAFIR supports individual-level longitudinal monitoring through temporally evolving and adaptive feature selection. Figure 5 visualizes the trajectories of selected fall-related features for Subject 1017 across four study visits, with a recorded fall occurring shortly after the T2 visit. Figure 5 reveals a clear transition in feature contributions across the pre- and post-fall periods. At T1, *Lying Time* (Physical Performance Assessments) exhibits the highest contribution, indicating that sedentary behavior is a prominent fall-related feature at baseline (Xu et al., 2019; Thiamwong et al., 2023a). At T2, shortly before the fall, *Hand Grip Strength (R)* (Physical Performance Assessments) receives the highest visit-level contribution, whereas *Lying Time* is not selected at the scheduled visit. This pre-fall pattern, centered on physical activity and muscle function, is consistent with established fall-risk factors among older adults (Thiamwong et al., 2023b). Following the fall, the selected feature profile broadens across multiple domains. At the fall-related update shortly after T2, *Lying Time* receives the largest contribution, while *Hand Grip Strength* remains prominent. At T3, *Lying Time* reaches its highest observed contribution, alongside *Mindfulness (MAAS-9)*(Questionnaires–Psychological), while *STEADI Item 11* (Clinical) is also selected. At T4, *Chronotype (MEQ)* (Questionnaires–Behavior) becomes the most prominent selected feature (Yi et al., 2022; Taheri et al., 2025). The increased contribution of *Lying Time* after the fall may be consistent with reduced mobility or fear-related activity restriction, while the emergence of psychological, behavioral, and clinical screening features illustrates the multidimensional nature of post-fall follow-up. From a personalized

Table 1: Performance results for fall risk factor selection on the PEER study.

Method	FRR ↑	FDR ↓	Selection Stability ↑	Correlation Recovery ↑
PAFIR (Ours)	**0.874 ± 0.007**	**0.207 ± 0.009**	**0.990 ± 0.003**	**0.974 ± 0.014**
TAR (2024)	0.795 ± 0.009	0.209 ± 0.008	0.989 ± 0.001	0.973 ± 0.014
SAT (2022)	0.731 ± 0.010	0.311 ± 0.011	0.985 ± 0.003	0.973 ± 0.013
TTG (2018)	0.664 ± 0.009	0.379 ± 0.012	0.985 ± 0.004	0.972 ± 0.013
PCA (1993)	0.624 ± 0.015	0.434 ± 0.012	0.978 ± 0.002	0.971 ± 0.016
Group LASSO (2006)	0.573 ± 0.154	0.667 ± 0.061	0.879 ± 0.134	0.970 ± 0.005
LASSO (1996)	0.278 ± 0.005	0.500 ± 0.000	0.990 ± 0.009	0.902 ± 0.011

monitoring perspective, these temporal patterns demonstrate PAFIR's capacity to generate adaptive, individual-level feature priorities. The pre-fall selections emphasize physical activity and muscle function, whereas the post-fall profile expands to include physical, psychological, behavioral, and clinical screening domains, suggesting that these areas may warrant broader review during subsequent follow-up (Thiamwong et al., 2023b).

5.2. Benchmark and Ablation Analysis

Evaluation Metrics. We compare PAFIR with several state-of-the-art feature selection baselines, including TAR (Ying et al., 2024), SAT (Chen et al., 2022), TTG (Khurana et al., 2018), PCA (Maćkiewicz and Ratajczak, 1993), Group LASSO (Yuan and Lin, 2006), and LASSO (Tibshirani, 1996). Participants are randomly divided into 80% training and 20% testing sets, and all longitudinal observations from the same participant are retained in the same partition. All data-dependent quantities, including feature-importance scores and inter-group correlations, are estimated using the training partition only. All experiments are repeated 20 times with different random seeds, and the results are reported as mean ± standard deviation. We assess the quality of feature selection directly using four complementary metrics. **Feature Recovery Rate (FRR)** measures the proportion of fall-relevant features identified. **False Discovery Rate (FDR)** quantifies the proportion of irrelevant features selected (Benjamini and Hochberg, 1995). **Selection Stability** evaluates consistency across runs (Kuncheva, 2007). **Correlation Recovery** assesses how well feature dependencies are preserved (Meinshausen and Bühlmann, 2010, 2006). The complete feature- and group-level initial reference sets used for evaluation are provided in Appendix B.2.

Performance and Ablation Study Table 1 summarizes the overall feature selection performance on the PEER study. PAFIR achieves the highest Feature Recovery Rate (FRR) and Selection Stability, while maintaining a competitive False Discovery Rate (FDR). This indicates that PAFIR is able to consistently identify clinically relevant fall-risk features. Its advantage is especially clear in Correlation Recovery, where PAFIR better preserves the relationships among features compared to other methods. This is important in practice, as it allows clinicians to interpret how different risk factors interact, rather than treating them as independent variables.

Table 2 shows that removing any component leads to consistent performance degradation. Without the hierarchical structure, FRR and selection stability decrease. Without the comparison-driven selection mechanism, FRR decreases and FDR increases. Without the sparse-aware reward, FRR and correlation recovery are the lowest. These results

Table 2: Ablation Study for Fall Risk Factors Selection on the PEER Study.

Method	FRR ↑	FDR ↓	Selection Stability ↑	Correlation Recovery ↑
PAFIR (Full)	**0.874 ± 0.007**	**0.207 ± 0.009**	**0.990 ± 0.003**	**0.974 ± 0.014**
w/ Random Feature-Importance Prior	0.870 ± 0.015	0.211 ± 0.008	0.980 ± 0.011	0.970 ± 0.014
w/o Hierarchical Structure	0.768 ± 0.010	0.229 ± 0.009	0.769 ± 0.010	0.752 ± 0.008
w/o Three-Agent Comparison-Driven Mechanism	0.678 ± 0.005	0.327 ± 0.013	0.675 ± 0.007	0.765 ± 0.033
w/o Sparse-Aware Rewards	0.655 ± 0.011	0.325 ± 0.008	0.665 ± 0.007	0.701 ± 0.013

highlight the complementary roles of structural representation, comparison-driven selection, and reward design. Replacing the baseline feature-importance prior with a random prior yields only modest performance changes, indicating that the prior provides useful guidance without determining the learned policy. Additional results are reported in Appendix B.4, Table 5, and Appendices D–F. The public benchmark experiments support methodological applicability but do not constitute external clinical validation.

Longitudinal Stability in the No-Fall Subgroup Among participants with no recorded falls, PAFIR shows high group-level stability across visits, with an overall Jaccard similarity of 0.890 ± 0.009. In contrast, individual-feature selections vary more substantially. Detailed group- and feature-level results are provided in Appendix B.3 and Table 4.

6. Conclusion, Future Work, and Limitations

This study presents PAFIR, a multi-level RL-based feature selection framework designed to support personalized and adaptive fall prevention for older adults. By modeling the hierarchical structural assessments and aligning evolving physical activities from wearable sensors, PAFIR adaptively identifies temporally dynamic and fall risk factors, which can inform personalized recommendations. Particularly, PAFIR demonstrates its potential to identify the actionable early warning signal from the in-field PEER study and to inform personalized monitoring and fall-prevention planning as risk factors evolve.

Future Work and Limitations Future directions for PAFIR include integration with digital health and mHealth platforms, such as Ecological Momentary Assessment (EMA) systems, to enable real-time monitoring and adaptive feedback. By incorporating high-frequency self-reported data on symptoms and behaviors, PAFIR may support just-in-time adaptive interventions (JITAIs) that deliver personalized prompts when relevant changes in mobility or functional capacity are identified. This integration could bridge passive sensing with active behavioral support. Future studies will evaluate PAFIR-powered digital interventions in community-dwelling older adults. External validation is currently constrained by the lack of comparable longitudinal multimodal fall-prevention cohorts.

Acknowledgments

We thank the reviewers and the Area Chair for their constructive feedback. This research was supported by NIH the National Institute on Minority Health and Health Disparities (R01MD018025) and the NIH Office of the Director, Chief Officer for Scientific Workforce Diversity (COSWD) (3R01MD018025-02S1), and the Learning Institute for Elders at University of Central Florida Richard Tucker Gerontology Applied Research Grant.

References

Abeer Alsadoon, Ghazi Al-Naymat, and Oday D Jerew. An architectural framework of elderly healthcare monitoring and tracking through wearable sensor technologies. *Multimedia Tools and Applications*, 83(26):67825–67870, 2024.

Murray R Barrick and Ryan D Zimmerman. Reducing voluntary, avoidable turnover through selection. *Journal of applied psychology*, 90(1):159, 2005.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Centers for Disease Control and Prevention. About older adult fall prevention. https://www.cdc.gov/falls/about/index.html, 2026. Accessed: 2026-01-27.

Thanwarat Chantanachai, Daina L Sturnieks, Stephen R Lord, Narelle Payne, Lyndell Webster, and Morag E Taylor. Risk factors for falls in older people with cognitive impairment living in the community: systematic review and meta-analysis. *Ageing research reviews*, 71:101452, 2021.

Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. Structure-aware transformer for graph representation learning. In *International conference on machine learning*, pages 3469–3489. PMLR, 2022.

Jiraporn Chittrakul, Penprapa Siviroj, Somporn Sungkarat, and Ratana Sapbamrer. Physical frailty and fall risk in community-dwelling older adults: A cross-sectional study. *Journal of aging research*, 2020(1):3964973, 2020.

Cathleen S Colón-Emeric, Cara L McDermott, Deborah S Lee, and Sarah D Berry. Risk assessment and prevention of falls in older community-dwelling adults: a review. *Jama*, 331(16):1397–1406, 2024.

Michael J Crowther, Therese M-L Andersson, Paul C Lambert, Keith R Abrams, and Keith Humphreys. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in medicine*, 35(7):1193–1209, 2016.

Ramon Cuevas-Trisan. Balance problems and fall risks in the elderly. *Physical Medicine and Rehabilitation Clinics*, 28(4):727–737, 2017.

Lauren Dautzenberg, Shanthi Beglinger, Sofia Tsokani, Stella Zevgiti, Renee CMA Raijmann, Nicolas Rodondi, Rob JPM Scholten, Anne WS Rutjes, Marcello Di Nisio, Marielle Emmelot-Vonk, et al. Interventions for preventing falls and fall-related fractures in community-dwelling older adults: a systematic review and network meta-analysis. *Journal of the American Geriatrics Society*, 69(10):2973–2984, 2021.

Kim Delbaere, Jacqueline CT Close, Henry Brodaty, Perminder Sachdev, and Stephen R Lord. Determinants of disparities between perceived and physiological risk of falling among elderly people: cohort study. *Bmj*, 341, 2010.

Paul D Dobson and Andrew J Doig. Distinguishing enzyme structures from non-enzymes without alignments. *Journal of molecular biology*, 330(4):771–783, 2003.

Jianwei Dong, Yumin Chen, Bingyu Yao, Xiao Zhang, and Nianfeng Zeng. A neural network boosting regression model based on xgboost. *Applied Soft Computing*, 125:109067, 2022.

Pamela W Duncan, Julie Chandler, Stephanie Studenski, Michael Hughes, and Barbara Prescott. How do physiological components of balance affect mobility in elderly men? *Archives of physical medicine and rehabilitation*, 74(12):1343–1349, 1993.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

Robert Elashoff, Ning Li, et al. *Joint modeling of longitudinal and time-to-event data*. Chapman and Hall/CRC, 2016.

Lingjie Fan, Junhan Zhao, Yao Hu, Junjie Zhang, Xiyue Wang, Fengyi Wang, Mengyi Wu, and Tao Lin. Predicting physical functioning status in older adults: insights from wrist accelerometer sensors and derived digital biomarkers of physical activity. *Journal of the American Medical Informatics Association*, 31(11):2571–2582, 2024.

Curtis S Florence, Gwen Bergen, Adam Atherly, Elizabeth Burns, Judy Stevens, and Cynthia Drake. Medical costs of fatal and nonfatal falls in older adults. *Journal of the American Geriatrics Society*, 66(4):693–698, 2018.

Joseph B Franklin, Caroline Marra, Kaleab Z Abebe, Atul J Butte, Deborah J Cook, Laura Esserman, Lee A Fleisher, Cynthia I Grossman, Nancy E Kass, Harlan M Krumholz, et al. Modernizing the data infrastructure for clinical research to meet evolving demands for evidence. *JAMA*, 332(16):1378–1385, 2024.

Arpitha Bhaskara Rao Ghat. *Future Evolution of Telemedicine: Enhancing Healthcare Accessibility and Reliability through the Integration of Machine Learning Techniques*. PhD thesis, Dublin, National College of Ireland, 2023.

Silvia Giovannini, Fabrizio Brau, Vincenzo Galluzzo, Domenico Alessandro Santagada, Claudia Loreti, Lorenzo Biscotti, Alice Laudisio, Giuseppe Zuccala, and Roberto Bernabei. Falls among older adults: screening, identification, rehabilitation, and management. *Applied Sciences*, 12(15):7934, 2022.

Kaiqi Gong, Xiao Song, Wenxin Li, and Senzhang Wang. Hn-gccf: High-order neighbor-enhanced graph convolutional collaborative filtering. *Knowledge-Based Systems*, 283: 111122, 2024.

Ana González-Castro, Raquel Leirós-Rodríguez, Camino Prada-García, and José Alberto Benítez-Andrades. The applications of artificial intelligence for assessing fall risk: systematic review. *Journal of medical internet research*, 26:e54934, 2024.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

Yang He, Bas Brouwers, Hesong Liu, Hailan Liu, Katherine Lawler, Edson Mendes de Oliveira, Dong-Kee Lee, Yongjie Yang, Aaron R Cox, Julia M Keogh, et al. Human loss-of-function variants in the serotonin 2c receptor associated with obesity and maladaptive behavior. *Nature Medicine*, 28(12):2537–2546, 2022.

Philip Hougaard. Frailty models for survival data. *Lifetime data analysis*, 1(3):255–273, 1995.

Jennifer Howcroft, Jonathan Kofman, and Edward D Lemaire. Prospective fall-risk prediction models for older adults based on wearable sensors. *IEEE transactions on neural systems and rehabilitation engineering*, 25(10):1812–1820, 2017.

John D Kalbfleisch and Douglas E Schaubel. Fifty years of the cox model. *Annual Review of Statistics and Its Application*, 10(1):1–23, 2023.

Udayan Khurana, Horst Samulowitz, and Deepak Turaga. Feature engineering for predictive modeling using reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Lou C Kohler-Voinov, Zahra N Sayyid, and Kathleen E Cullen. Multifactorial predictors of falls in older adults: a decade of data from the national health and aging trends study. *BMC geriatrics*, 25(1):950, 2025.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Ludmila I Kuncheva. A stability index for feature selection. In *Artificial intelligence and applications*, pages 421–427, 2007.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 95–104, 2018.

Susan S Levy, Katie J Thralls, and Shiloah A Kviatkovsky. Validity and reliability of a portable balance tracking system, btracks, in older adults. *Journal of geriatric physical therapy*, 41(2):102–107, 2018.

Fan Li, Wenhan Lu, Yuxuan Wang, Zehua Pan, Erich J Greene, Guanqun Meng, Can Meng, Ondrej Blaha, Yize Zhao, Peter Peduzzi, et al. A comparison of analytical strategies for cluster randomized trials with survival outcomes in the presence of competing risks. *Statistical Methods in Medical Research*, 31(7):1224–1241, 2022.

Susan Li and Kamalakar Surineni. Falls in hospitalized patients and preventive strategies: A narrative review. *The American Journal of Geriatric Psychiatry: Open Science, Education, and Practice*, 5:1–9, 2025.

Zhe Li, Shiyi Qi, Yiduo Li, and Zenglin Xu. Revisiting long-term time series forecasting: An investigation on linear mapping. *arXiv preprint arXiv:2305.10721*, 2023.

Zhe Khae Lim, Tee Connie, Michael Kah Ong Goh, and Nor 'Izzati Binti Saedon. Fall risk prediction using temporal gait features and machine learning approaches. *Frontiers in Artificial Intelligence*, 7:1425713, 2024.

Chang Liu, Ladda Thiamwong, Yanjie Fu, and Rui Xie. Diffusion policies with offline and inverse reinforcement learning for promoting physical activity in older adults using wearable sensors. In *2025 International Conference on Machine Learning and Applications (ICMLA)*, pages 275–282, 2025a. doi: 10.1109/ICMLA66185.2025.00043.

Minhao Liu, Ailing Zeng, Muxi Chen, Zhijian Xu, Qiuxia Lai, Lingna Ma, and Qiang Xu. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems*, 35:5816–5828, 2022.

Yi Liu, Chang Liu, Liqiang Ni, Wei Zhang, Chen Chen, Janet Lopez, Hao Zheng, Ladda Thiamwong, and Rui Xie. Effectiveness of peer intervention on older adults' physical activity time series using smoothing spline anova. *Mathematics*, 13(3):516, 2025b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

Yuankai Luo, Hongkang Li, Lei Shi, and Xiao-Ming Wu. Enhancing graph transformers with hierarchical distance structural encoding. *Advances in Neural Information Processing Systems*, 37:57150–57182, 2024.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

Agnieszka Maruszewska, Tadeusz Ambroży, and Łukasz Rydzik. Risk factors and socioeconomic determinants of falls among older adults. *Frontiers in Public Health*, 13:1571312, 2025.

Killian McManus, Barry R Greene, Lilian Genaro Motti Ader, and Brian Caulfield. Development of data-driven metrics for balance impairment and fall risk assessment in older adults. *IEEE Transactions on Biomedical Engineering*, 69(7):2324–2332, 2022.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. 2006.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

Sara Mortazavi, Ahmad Delbari, Mohsen Vahedi, Reza Fadayevatan, Mitra Moodi, Hossein Fakhrzadeh, Masoumeh Khorashadizadeh, Ameneh Sobhani, Moloud Payab, Mahbube Ebrahimpur, et al. Low physical activity and depression are the prominent predictive factors for falling in older adults: the birjand longitudinal aging study (blas). *BMC geriatrics*, 23(1):758, 2023.

Amitangshu Mukherjee, Isha Garg, and Kaushik Roy. Encoding hierarchical information in neural networks helps in subpopulation shift. *IEEE Transactions on Artificial Intelligence*, 5(2):827–838, 2023.

Tho Nguyen, Ladda Thiamwong, Qian Lou, and Rui Xie. Unveiling fall triggers in older adults: a machine learning graphical model analysis. *Mathematics*, 12(9):1271, 2024.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Yoshita Paliwal, Patricia W Slattum, and Scott M Ratliff. Chronic health conditions as a risk factor for falls among the community-dwelling us older adults: a zero-inflated regression modeling approach. *BioMed research international*, 2017(1):5146378, 2017.

Margaret Sullivan Pepe. *The statistical evaluation of medical tests for classification and prediction*. Oxford university press, 2003.

CA Pfortmueller, Gregor Lindner, and AK Exadaktylos. Reducing fall risk in the elderly: risk factors and fall prevention, a systematic review. *Minerva Med*, 105(4):275–81, 2014.

Pietro Picerno, Marco Iosa, Clive D'Souza, Maria Grazia Benedetti, Stefano Paolucci, and Giovanni Morone. Wearable inertial sensors for human movement analysis: A five-year update. *Expert review of medical devices*, 18(sup1):79–94, 2021.

Jennifer Pillay, Lindsay A Gaudet, Sabrina Saba, Ben Vandermeer, Ashiqur Rahman Ashiq, Aireen Wingert, and Lisa Hartling. Falls prevention interventions for community-dwelling older adults: Systematic review and meta-analysis of benefits, harms, and patient values and preferences. *Systematic Reviews*, 13(1):289, 2024.

Kaspar Riesen and Horst Bunke. Iam graph database repository for graph based pattern recognition and machine learning. In *Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR)*, pages 287–297. Springer, 2008.

Katherine Ritchey, Amanda Olney, Sunny Chen, and Elizabeth A Phelan. Steadi self-report measures independently predict fall risk. *Gerontology and geriatric medicine*, 8: 23337214221079222, 2022.

Anthony J Rosellini and Timothy A Brown. Developing and validating clinical questionnaires. *Annual review of clinical psychology*, 17(1):55–81, 2021.

Yuri Rykov, Thuan-Quoc Thach, Iva Bojic, George Christopoulos, and Josip Car. Digital biomarkers for depression screening with wearable devices: cross-sectional study with machine learning modeling. *JMIR mHealth and uHealth*, 9(10):e24872, 2021.

Joanne Salop and Steven Salop. Self-selection and turnover in the labor market. *The Quarterly Journal of Economics*, 90(4):619–627, 1976.

Stephanie Saunders, Cassandra D'Amore, Quikui Hao, Nabil Abd El-Moneim, Julie Richardson, Ayse Kuspinar, and Marla Beauchamp. Risk factors for falls in community-dwelling older adults: an umbrella review. *Journal of the American Medical Directors Association*, 26(9):105765, 2025.

Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl_1):D431–D433, 2004.

Hayley Smith, Michael Sweeting, Tim Morris, and Michael J Crowther. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagnostic and Prognostic Research*, 6(1):10, 2022.

Daina L Sturnieks, LY Lloyd, Maria Teresa Espinoza Cerda, Carmen Herrera Arbona, Beatriz Herrero Pinilla, Paula Santiago Martinez, Nigel Wei Seng, Natassia Smith, Jasmine C Menant, and R LORD Stephen. Cognitive functioning and falls in older people: A systematic review and meta-analysis. *Archives of Gerontology and Geriatrics*, 128:105638, 2025.

Krithika Suresh, Cameron Severn, and Debashis Ghosh. Survival prediction models: an introduction to discrete-time modeling. *BMC medical research methodology*, 22(1):207, 2022.

Nima Taheri, Luis Becker, Lena Fleig, Karolina Kolodziejczak, Lea Cordes, Bernhard U Hoehl, Ulrike Grittner, Lukas Mödl, Hendrik Schmidt, and Matthias Pumberger. Fear-avoidance beliefs are associated with changes of back shape and function. *Pain Reports*, 10(2):e1249, 2025.

Dipti Theng and Kishor K Bhoyar. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems*, 66 (3):1575–1637, 2024.

L. Thiamwong, R. Xie, J. H. Park, N. Lighthall, V. Loerzel, and J. Stout. Optimizing a technology-based body and mind intervention to prevent falls and reduce health disparities in low-income populations: Protocol for a clustered randomized controlled trial. *JMIR research protocols*, 12:e51899, 2023a.

Ladda Thiamwong, Mary Lou Sole, Boon Peng Ng, Gregory F Welch, Helen J Huang, and Jeffrey R Stout. Assessing fall risk appraisal through combined physiological and perceived fall risk measures using innovative technology. *Journal of gerontological nursing*, 46(4):41–47, 2020a.

Ladda Thiamwong, Jeffrey R Stout, Mary Lou Sole, Boon Peng Ng, Xin Yan, and Steven Talbert. Physio-feedback and exercise program (peer) improves balance, muscle strength, and fall risk in older adults. *Research in gerontological nursing*, 13(6):289–296, 2020b.

Ladda Thiamwong, Rui Xie, Norma E Conner, Justine M Renziehausen, Eunice Oladepe Ojo, and Jeffrey R Stout. Body composition, fear of falling and balance performance in community-dwelling older adults. *Translational medicine of aging*, 7:80–86, 2023b.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Prakasit Tonchoy, Katekaew Seangpraw, Parichat Ong-Artborirak, Supakan Kantow, Nisarat Auttama, Monchanok Choowanthanapakorn, and Sorawit Boonyathee. Mental health, fall prevention behaviors, and home environments related to fall experiences among older adults from ethnic groups in rural northern thailand. *Heliyon*, 10(17), 2024.

Md Palash Uddin, Md Al Mamun, and Md Ali Hossain. Pca-based feature reduction for hyperspectral remote sensing image classification. *IETE Technical Review*, 38(4):377–396, 2021.

Jane-Ling Wang and Qixian Zhong. Joint modeling of longitudinal and survival data. *Annual Review of Statistics and Its Application*, 12(1):449–476, 2025.

Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6202–6209, 2020.

Jixian Wang, Yongfang Li, Guo-Yuan Yang, and Kunlin Jin. Age-related dysfunction in balance: a comprehensive review of causes, consequences, and interventions. *Aging and disease*, 16(2):714, 2024.

Wei Wang, Dongqi Han, Xufang Luo, and Dongsheng Li. Addressing signal delay in deep reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Cecilia Xu, Peter R Ebeling, and David Scott. Body composition and falls risk in older adults. *Current Geriatrics Reports*, 8(3):210–222, 2019.

Tingyang Xu, Jiangwen Sun, and Jinbo Bi. Longitudinal lasso: Jointly learning features and temporal contingency for outcome prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1345–1354, 2015.

Lucy Yardley, Nina Beyer, Klaus Hauer, Gertrudis Kempen, Chantal Piot-Ziegler, and Chris Todd. Development and initial validation of the falls efficacy scale-international (fes-i). *Age and ageing*, 34(6):614–619, 2005.

AFOUDI Yassine, LAZAAR Mohamed, and Mohammed Al Achhab. Intelligent recommender system based on unsupervised machine learning and demographic attributes. *Simulation Modelling Practice and Theory*, 107:102198, 2021.

DongHyun Yi, SeungJun Jang, and JongEun Yim. Relationship between associated neuropsychological factors and fall risk factors in community-dwelling elderly. In *Healthcare*, volume 10, page 728. MDPI, 2022.

Wangyang Ying, Haoyue Bai, Kunpeng Liu, and Yanjie Fu. Topology-aware reinforcement feature space reconstruction for graph data. *ACM Transactions on Knowledge Discovery from Data*, 2024.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Wenxuan Zhong, Tingting Zhang, Yu Zhu, and Jun S Liu. Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(5):849–870, 2012.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Appendix A. Supplement Overview

This supplementary document presents additional information to support the main manuscript. It includes detailed descriptions of the datasets used, the full configuration of the PAFIR model, the training and evaluation procedures, and extended experimental results. These materials aim to enhance the reproducibility and transparency of our study and provide deeper insights into the implementation and performance of PAFIR across different settings. The code implementing the proposed PAFIR framework and all experimental pipelines is available at: https://github.com/changliu1993-cl/PAFIR

Appendix B. In-Field Evaluation: PEER Study on Fall Prevention in Older Adults

B.1. PEER Intervention and Control Across Visits

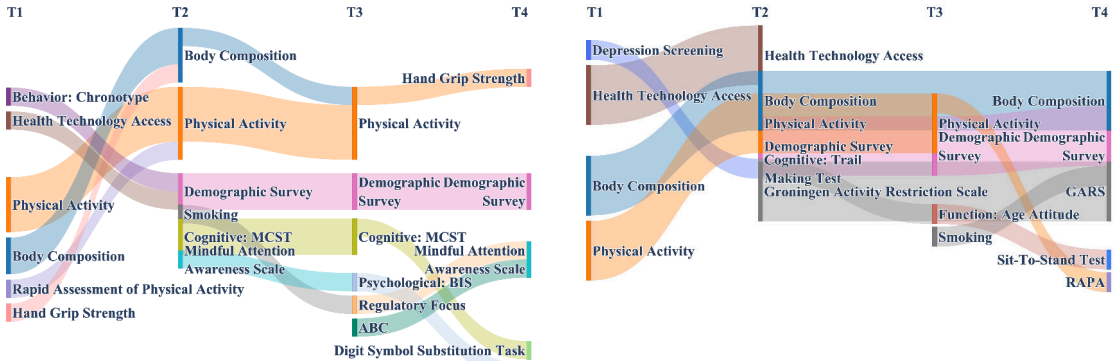
The top 10 most important fall risk factor groups and their evolving interconnections across four visits are visualized for both the PEER Intervention and Control in Fig. 6. In the PEER cluster (Fig. 6(a)), which received peer-led interventions, we observe greater temporal continuity in key domains such as cognitive function and physical activity. Fall risk factor groups like *MCST*, *Mindful Attention Awareness Scale*, and *balance confidence* appear repeatedly across multiple visits, suggesting a consistent focus on cognitive engagement and behavioral monitoring throughout the intervention period.

In contrast, the Control cluster shows more fragmented transitions, with higher variability in feature composition across visits in Fig. 6(b). Although core factors such as *physical activity* and *body composition* persist, other fall risk groups, particularly those related to cognition and behavior, tend to fluctuate more. This divergence highlights how the presence of a structured, peer-led program may help stabilize attention toward critical fall risk factors over time.

Despite these differences, both clusters consistently highlight domains like *physical activity* and *body composition*, reinforcing their foundational importance in fall risk factors identification. However, the PEER cluster captures a broader range of psychological and cognitive factors, which may support more holistic and individualized monitoring. These patterns underscore the role of community-based interventions in fostering consistent and interpretable health data trajectories across longitudinal assessments.

B.2. Initial Reference Sets for Evaluation

To evaluate feature- and group-level selection performance, we use visit-specific initial reference sets defined from the graph- and node-embedding results. Let $\mathcal{R}_\tau^{\text{feat}}$ and $\mathcal{R}_\tau^{\text{grp}}$ denote the feature- and group-level reference sets at visit τ, respectively. The reference sets are embedding-derived evaluation proxies constructed using the training partition only and independently of the evaluated methods. They provide a common reference for method comparison and should not be interpreted as exhaustive clinical ground truth.



(a) PEER Intervention (b) Control

Figure 6: Top 10 Fall Risk Factor Identification across 4 visits. (a) PEER Intervention, (b) Control.

Table 3: Initial feature- and group-level reference sets used for evaluation on the PEER study.

Reference Group	Reference Features	Visits
Demographic Survey	age, education, gender, health, living, number_of_falls, race, sleeping_pills, smoking	T1–T4
Fall Risk Screening: STEADI	ste8	T1–T4
Depression Screening: PHQ-9	phq	T1–T4
Anxiety Screening: GAI-SF	gai_sf_score	T1, T4
Fear of Falling: Short FES-I	fes	T1–T4
TUG, STS, and BTrackS	bbs, tug	T1–T4
Short Physical Performance Battery	balance, gait, speed_gait, sppb, sts_sppb_3	T1–T4
Hand Grip Strength	avg_hgs_kg_both_hands	T1–T3
Brief Aging Perceptions Questionnaire	b_apq_chronic, b_apq_conseqeunce_positive, b_apq_control_negative, b_apq_control_positive, b_apq_score	T2–T4
Behavioral Inhibition/Activation System	bas_drive, bas_fun, bas_reward, bis_score	T4
Regulatory Focus Questionnaire	rfq_prevention, rfq_promotion	T1, T4
FRAIL Questionnaire	fra	T4
Memory Impairment Screen	mis	T4
RUDAS	rudas_score	T4

B.3. Longitudinal Stability in the No-Fall Subgroup

To evaluate the longitudinal stability of PAFIR, we compare the selected feature groups across adjacent visits among participants with no recorded falls during follow-up. For each participant, Jaccard similarity (Hastie et al., 2009) is calculated as the number of feature groups selected at both visits divided by the total number of distinct feature groups selected across the two visits. Selection turnover (Salop and Salop, 1976; Barrick and Zimmerman, 2005) is defined as one minus the Jaccard similarity, with lower values indicating fewer changes between visits.

For each random seed, Jaccard similarity and turnover are calculated for all eligible participants with valid selections at both visits and then averaged within each adjacent visit pair. The overall result is obtained by first averaging across the available adjacent visit pairs for each participant and then averaging across eligible participants. Table 4 reports the mean and standard deviation of these results across 20 random seeds. Visit pairs with missing selections are excluded, and N denotes the number of eligible no-fall participants included in each comparison.

Table 4: Longitudinal stability of PAFIR group- and feature-level selections among participants with no recorded falls. Results are reported as mean ± standard deviation over 20 random seeds.

		Group-Level Selection		**Feature-Level Selection**	
Visit Pair	N	**Jaccard ↑**	**Turnover ↓**	**Jaccard ↑**	**Turnover ↓**
T1–T2	253	0.744 ± 0.007	0.256 ± 0.007	0.095 ± 0.007	0.905 ± 0.007
T2–T3	227	0.956 ± 0.014	0.044 ± 0.014	0.088 ± 0.006	0.912 ± 0.006
T3–T4	158	0.969 ± 0.015	0.031 ± 0.015	0.064 ± 0.007	0.936 ± 0.007
Overall	253	0.890 ± 0.009	0.110 ± 0.009	0.083 ± 0.004	0.917 ± 0.004

Table 5: Feature- and group-level fall-risk factor selection results on the PEER study. Precision, recall, and F1 score are computed relative to the predefined feature- and group-level reference sets.

Method	**Feature-Level Selection**			**Group-Level Selection**		
	Precision ↑	Recall ↑	F1 Score ↑	Precision ↑	Recall ↑	F1 Score ↑
PAFIR (Ours)	0.887 ± 0.008	**0.874 ± 0.007**	0.880 ± 0.007	**0.781 ± 0.011**	**0.793 ± 0.009**	**0.787 ± 0.010**
GraphGPS+HDSE (2024)	**0.891 ± 0.009**	**0.874 ± 0.008**	**0.882 ± 0.008**	0.736 ± 0.009	0.744 ± 0.010	0.740 ± 0.009
TAR (2024)	0.788 ± 0.010	0.795 ± 0.009	0.791 ± 0.010	0.767 ± 0.008	0.791 ± 0.008	0.779 ± 0.008
SAT (2022)	0.723 ± 0.009	0.731 ± 0.010	0.727 ± 0.009	0.677 ± 0.013	0.689 ± 0.011	0.683 ± 0.012
PCA (2021)	0.643 ± 0.013	0.624 ± 0.015	0.638 ± 0.014	0.573 ± 0.013	0.566 ± 0.012	0.569 ± 0.011
TTG (2018)	0.677 ± 0.008	0.664 ± 0.009	0.670 ± 0.008	0.643 ± 0.012	0.621 ± 0.012	0.611 ± 0.012

Among the 341 participants in the analyzed cohort, 48 had at least one recorded fall and 293 had no recorded falls during follow-up. Among the no-fall participants, 253, 227, and 158 had valid PAFIR feature-group selections at both visits for T1–T2, T2–T3, and T3–T4, respectively. Overall, 253 participants contributed to at least one adjacent-visit comparison. As shown in Table 4, PAFIR exhibits distinct stability patterns at the group and feature levels. The overall group-level Jaccard similarity is 0.890 ± 0.009, with a corresponding turnover of 0.110 ± 0.009, indicating that the broader selected risk domains remain highly consistent across adjacent visits. In contrast, the overall feature-level Jaccard similarity is 0.083 ± 0.004, corresponding to a turnover of 0.917 ± 0.004. These results indicate that PAFIR maintains stable group-level selections while adaptively updating the specific features selected within those groups across visits.

B.4. Feature- and Group-Level Benchmark and Ablation Results

Table 5 reports feature- and group-level selection performance on the PEER study. PAFIR achieves the highest group-level precision, recall, and F1 score, indicating that the hierarchical selection policy effectively identifies clinically relevant feature groups. At the individual-feature level, GraphGPS+HDSE obtains slightly higher precision and F1 score, while PAFIR achieves comparable recall. The ablation results show that removing hierarchical graph encoding, temporal encoding, or the three-agent comparison-driven selection mechanism reduces feature-level performance. Removing the three-agent mechanism also

Table 6: Computational cost of PAFIR on the PEER study.

Measurement	Result
GPU	NVIDIA Quadro RTX 6000
GPU memory capacity	24 GB
CUDA version	13.0
Runtime per random seed	∼101 s
Runtime for 20 seeds	∼33.7 min
Peak allocated GPU memory	∼240 MB

decreases group-level F1 from 0.787 to 0.716, supporting the contribution of the candidate, reference, and operation agents to group-level feature selection.

B.5. Computational Efficiency

Appendix C. Preprocessing

C.1. Hierarchical State Representation

To effectively model both the structural health assessment and temporal dynamics in longitudinal health data, we propose a state representation $\mathbf{s}_{i,t}$ that integrates graph-based and time-series modalities via a cross-attention mechanism. This architecture captures health status evolution across visits through three key components that jointly encode structurally informed features, such as fear of falling, at each visit, and temporally aligned physical activity patterns, such as step counts and vector magnitudes.

Hierarchical Structural Encoding First, structurally informed features are derived by applying the hierarchical distance structure encoding (HDSE) (Luo et al., 2024), which reveals and encodes latent multi-level feature relationships among features and their association with the primary outcome, *fall incidence*. At each visit τ, each node represents a feature (e.g., *BMI*, *weight*, and *height*), and each feature group (e.g., *body composition*) forms a fully connected graph $\mathcal{G}_m = (\mathcal{V}_m, \mathcal{E}_m)$, where $\mathcal{V}_m$ is the set of features belonging to group m. We construct a multi-level hierarchy for each feature group by iteratively coarsening the graph to generate $K+1$ levels. At each level k, we compute the graph hierarchy distance (GHD) (Luo et al., 2024) between nodes $v \in \mathcal{V}_m$ and $u \in \mathcal{V}_m$, denoted as,

$$\mathrm{GHD}^k(v,u) = \mathrm{SPD}(\psi_{k-1} \circ \cdots \circ \psi_0(v), \psi_{k-1} \circ \cdots \circ \psi_0(u)) \,,$$

where $\psi_{k-1} \circ \cdots \circ \psi_0(v)$ is the mapping of node v from level 0 to level k in the graph hierarchy, and $\mathrm{SPD}(\cdot,\cdot)$ denotes shortest path distance.

For each feature node v, we construct a structural distance tensor $\mathbf{D}_v \in \mathbb{R}^{(K+1)\times|\mathcal{V}_m|}$ by stacking its pairwise distances to all other nodes $u \in \mathcal{V}_m$ across all hierarchy levels. Each entry $\mathbf{D}_{v,u}$ captures the multi-level distances from node v to node u,

$$\mathbf{D}_{v,u} = \left[\mathrm{GHD}^0(v,u), \mathrm{GHD}^1(v,u), \cdots, \mathrm{GHD}^K(v,u)\right]. \tag{4}$$

This tensor captures the multi-level structural relationships of node v within group m. To obtain a compact representation, we apply positional encoding to $\mathbf{D}_v$, yielding a dense embedding that encodes the hierarchical position of the feature.

C.2. Temporal Evolving for Sequence and Time Series

To capture the physical activity dynamics observed after each clinical visit, we adopt a temporal state encoder based on the iTransformer architecture (Liu et al., 2023). For each feature, such as step counts and VMs, we extract a 7-day time series collected starting from the visit day, and treat it as an input token.

Let $\mathbf{X}_{i,:,b}^{(\tau)} \in \mathbb{R}^{L\times 1}$ denote the time series of feature b during the 7-day period following visit τ for participant i, where L is the sequence length (e.g., minute-level for 7 days). Each sequence is projected into a latent space via a shared embedding function,

$$\mathbf{z}_{i,t,b}^{(0)} = \mathrm{Embed}\Big(\mathbf{X}_{i,:,b}^{(\tau)}\Big) \in \mathbb{R}^{d}, \tag{5}$$

where d means the embedding dimension, which controls the capacity of the latent representation. We then apply a stack of M_{temp} Transformer encoder blocks to capture temporal dependencies,

$$\mathbf{z}_{i,t,b}^{(\ell+1)} = \mathrm{TrmBlock}^{(\ell)}\left(\mathbf{z}_{i,t,b}^{(\ell)}\right), \qquad \ell = 0,\ldots,M_{\text{temp}}-1. \tag{6}$$

The final encoder output $\mathbf{z}_{i,t,b}$ summarizes the temporal dynamics of feature b and serves as the query for the cross-attention mechanism.

Cross-attention Mechanism In the cross-attention mechanism, we regard the temporal embeddings of physical activity patterns as queries $\mathbf{Q}_{i,t} \in \mathbb{R}^{|\mathcal{B}|\times d'}$, the structural embeddings of feature groups as keys $\mathbf{K}_{i,t} \in \mathbb{R}^{M\times d'}$, and the corresponding within-group structural embeddings as values $\mathbf{V}_{i,t} \in \mathbb{R}^{M\times d'}$, where $\mathcal{B}$ is the set of temporal features, M is the number of feature groups, and d' is the hidden dimension per attention head. To incorporate structural priors, we introduce a bias matrix $\mathbf{B}_{i,t}^{\text{struct}} \in \mathbb{R}^{|\mathcal{B}|\times M}$ to encourage each temporal query to attend more strongly to structurally relevant feature groups, where each entry encodes the relative GHD-based proximity between temporal feature b and feature group $\mathcal{V}_m$. The attention is then computed as

$$\mathrm{CrossAttn}\left(\mathbf{Q}_{i,t},\mathbf{K}_{i,t},\mathbf{V}_{i,t}\right) = \mathrm{softmax}\left(\frac{\mathbf{Q}_{i,t}\mathbf{K}_{i,t}^{\top}}{\sqrt{d'}} + \mathbf{B}_{i,t}^{\text{struct}}\right)\mathbf{V}_{i,t}. \tag{7}$$

This fusion mechanism enables the model to align temporal physical activity patterns with the longitudinal structure of clinical survey data, allowing for more informed and interpretable inter-feature association.

C.3. Three-Agent RL Feature Selection Framework

We leverage a three-agent decision framework (Ying et al., 2024) that performs personalized and temporally adaptive feature selection under RL. At each clinical visit, the state representation $\mathbf{s}_{i,t}$ incorporates both temporal behavior dynamics from physical activity

and structural information from clinical feature topology, and is used to guide the decision-making in a three-agent RL structure composed of a candidate feature group agent, a reference feature group agent, and an operation agent. In the RL framework, time step t can correspond to a clinical visit τ or to the occurrence of a fall incident for participant i. Let $\mathcal{C} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_M\}$ denote the collection of clinically defined feature groups.

Group-Level Feature Selection At each iteration t, the framework operates

- *Candidate Feature Group Agent*: Selects a feature group $C_{i,t}^{\text{cand}} \in \mathcal{C}$ based on the current state $\mathbf{s}_{i,t}$.
- *Reference Feature Group Agent*: Selects a comparison feature group $C_{i,t}^{\text{ref}} \in \mathcal{C} \setminus \{C_{i,t}^{\text{cand}}\}$ from the remaining feature groups.
- *Operation Agent*: Selects an add, remove, or retain operation based on the relative incremental utility of $C_{i,t}^{\text{cand}}$ and $C_{i,t}^{\text{ref}}$ under the available outcome signal $y_{i,t}$.

The resulting action $a_{i,t} = (C_{i,t}^{\text{cand}}, C_{i,t}^{\text{ref}}, o_{i,t})$ governs the inclusion or exclusion of existing feature groups. The group with higher relevance to the outcome is retained in the selected feature set $F_{i,t}$, and the other is used as the reference in the next iteration.

Within-Group Feature Selection To further refine feature granularity, we extend the three-agent structure to operate within each selected group $\mathcal{V}_m \in F_{i,t}^{\text{group}}$. For each group, a second round of selection is performed at the individual feature level. A localized state representation $\mathbf{s}_{i,t}^{(\mathcal{V}_m)}$ encodes the temporal and structural context of individual features in $\mathcal{V}_m$.

Within-group agents operate similarly. The Candidate Feature Agent selects a feature $x_{i,t}^{\text{cand}} \in \mathcal{V}_m$ (e.g., *BMI*), while the Reference Feature Agent chooses a comparison feature $x_{i,t}^{\text{ref}} \in \mathcal{V}_m$ (e.g., *fat-free mass*) from the remaining features in the same group. Finally, the Operation Agent decides whether to retain $x_{i,t}^{\text{cand}}$. The temporal-difference (TD) learning and reward mechanisms follow the same procedure as the group level. The Q-networks are randomly initialized, while the HDSE node-level importance scores are used to initialize the feature-level rewards.

Appendix D. Benchmarking and Results

In this section, we evaluate our proposed method on ten public datasets and demonstrate state-of-the-art performance. We investigate the research problems: RQ1: Can our method effectively reconstruct high-quality feature spaces to improve performance on downstream tasks in both graph and time series domains? RQ2: Does our unified architecture outperform state-of-the-art baselines in terms of both structural and temporal evaluation metrics across diverse datasets?

D.1. Baselines

We compare our method against state-of-the-art baselines. For graph-based benchmarks, we include HDSE (Luo et al., 2024), TAR (Ying et al., 2024), SAT (Chen et al., 2022), PCA (Uddin et al., 2021), and TTG (Khurana et al., 2018). For time series forecasting, we consider iTransformer (Liu et al., 2023), RLinear (Li et al., 2023), PatchTST (Nie et al., 2022), TimesNet (Wu et al., 2022), and SCINet (Liu et al., 2022).

D.2. Benchmark Dataset Description

We evaluate our method on six public benchmark datasets across different domains. For graph-based comparisons, we use two bioinformatics datasets: ENZYMES (Schomburg et al., 2004) and PROTEINS (Dobson and Doig, 2003), and a small molecules dataset: AIDS (Riesen and Bunke, 2008). For time series evaluation, we use three datasets from the energy and transportation domains: Solar-Energy (Lai et al., 2018), Traffic (Wu et al., 2021), Weather (Wu et al., 2021), and ETT (Zhou et al., 2021) (details in Table 7.).

Table 7: Detailed Dataset Description. Dim denotes the variate number of each dataset. Dataset Size denotes the total number of time points in (Train, Validation, Test), respectively. Prediction Length denotes the future time points to be predicted in each dataset. Frequency denotes the sampling interval of timepoints. *Top Side*: Time Series Public Datasets. *Bottom Side*: Graph Public Datasets.

Time-Series Dataset	**Dim**	**Prediction Length**	**Dataset Size**	**Frequency**	**Information**
ETTh1, ETTh2	7	{96, 192, 336, 720}	(8545, 2881, 2881)	Hourly	Electricity
ETTm1, ETTm2	7	{96, 192, 336, 720}	(34465, 11521, 11521)	15min	Electricity
Traffic	862	{96, 192, 336, 720}	(12185, 1757, 3509)	Hourly	Transportation
Solar-Energy	137	{96, 192, 336, 720}	(36601, 5161, 10417)	Hourly	Energy
Weather	21	{96, 192, 336, 720}	(36792, 5271, 10540)	10min	Weather
Graph Dataset	**Graphs Counts**	**Nodes Counts**	**Graph Classes**	**Node Classes**	**Labels**
ENZYMES	600	19580	6	3	Yes
PROTEINS	1113	43471	2	3	Yes
AIDS	2000	31385	2	38	Yes

For the graph prediction, we follow the data processing and train-validate-test set split protocol used in TAR Ying et al. (2024). For the time-series prediction, we follow the same data processing and train-validate-test set split protocol used in iTransformer Liu et al. (2023), where the train, validation, and test datasets are strictly divided according to chronological order to make sure there are no data leakage issues.

D.3. Environment Setup and Metrics

For graph benchmarks, we follow the data-processing and split protocol of TAR (Ying et al., 2024). For time-series benchmarks, we use the predefined chronological train/validation/test splits adopted by iTransformer (Liu et al., 2023). To ensure robustness, each experiment is repeated 10 times with different random seeds, and the mean performance is reported. For graph-based tasks, we evaluate the quality of the transformed feature space using Precision, Recall, and F1 score. Higher values indicate better performance. For reinforcement feature space reconstruction, training was limited to 10 epochs, each with 10 exploration steps. All agents were implemented using a DQN with two ReLU-activated linear layers, optimized with Adam (learning rate 0.01), an experience replay memory size of 32, and a batch size of 8. For time series forecasting tasks, we report prediction accuracy using Mean Squared Error (MSE) and Mean Absolute Error (MAE). Lower values indicate better performance. For the forecasting settings, we fixed the lookback window length to 96 for the ETT, Solar-Energy, Traffic, and Weather datasets, while the prediction horizons were set to vary among {96, 192, 336, 720}.

Model Regularization and Generalization Given the high-dimensional feature space and sparse fall outcomes, we adopt multiple strategies to mitigate overfitting in the proposed PAFIR framework. First, strong regularization is applied throughout the model, including dropout (rate = 0.5) and residual connections to stabilize deep representations. Second, hierarchical and group-wise structural representations reduce the effective dimensionality by leveraging clinically meaningful feature groupings rather than treating all variables independently.

For temporal modeling, a shared encoder is used across longitudinal visits, preventing visit-specific overfitting and encouraging the learning of consistent temporal patterns. Finally, the adaptive policy is optimized using temporally aggregated reward signals rather than directly fitting individual fall events, reducing sensitivity to sparse and delayed outcome labels.

D.4. Benchmarks Results

To address RQ1, we evaluate the performance of our proposed method, PAFIR, in comparison with several state-of-the-art baselines across six benchmark datasets covering both graph classification and time series forecasting tasks. On the PROTEINS, ENZYMES, and AIDS datasets, as shown in Table 8, PAFIR achieves the highest scores across all evaluation metrics, including Precision, Recall, and F1 Score. These results confirm the effectiveness of PAFIR in capturing hierarchical structure patterns within graph data, consistently outperforming baselines such as GraphGPS+HDSE, TAR, SAT, PCA, and TTG.

Table 8: Overall Performance Comparison on Hierarchical Graph Datasets. PAFIR outperforms existing methods on both node and graph classification tasks across PROTEINS, ENZYMES, and AIDS datasets, following the TAR (Ying et al., 2024) setting.

Dataset	Method	Node Classification			Graph Classification		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
PROTEINS	PAFIR	**0.921 ± 0.007**	0.914 ± 0.007	**0.918 ± 0.009**	**0.843 ± 0.006**	**0.822 ± 0.007**	**0.832 ± 0.006**
	GraphGPS+HDSE (2024)	0.867 ± 0.055	0.785 ± 0.097	0.780 ± 0.180	0.830 ± 0.027	0.812 ± 0.055	0.821 ± 0.012
	TAR 2024	0.916 ± 0.017	**0.925 ± 0.017**	0.916 ± 0.017	0.767 ± 0.006	0.766 ± 0.007	0.765 ± 0.006
	SAT (2022)	0.779 ± 0.104	0.785 ± 0.097	0.780 ± 0.037	0.769 ± 0.037	0.643 ± 0.039	0.701 ± 0.104
	PCA 2021	0.781 ± 0.003	0.719 ± 0.003	0.738 ± 0.003	0.643 ± 0.001	0.648 ± 0.002	0.644 ± 0.001
	TTG (2018)	0.847 ± 0.005	0.856 ± 0.005	0.847 ± 0.005	0.752 ± 0.004	0.751 ± 0.004	0.751 ± 0.004
ENZYMES	PAFIR	**0.937 ± 0.004**	**0.940 ± 0.007**	**0.941 ± 0.006**	**0.866 ± 0.006**	**0.848 ± 0.007**	**0.857 ± 0.006**
	GraphGPS+HDSE (2024)	0.816 ± 0.006	0.862 ± 0.277	0.885 ± 0.121	0.842 ± 0.003	0.819 ± 0.004	0.831 ± 0.106
	TAR 2024	0.936 ± 0.006	0.934 ± 0.005	0.937 ± 0.006	0.324 ± 0.053	0.358 ± 0.029	0.325 ± 0.029
	SAT (2022)	0.744 ± 0.007	0.782 ± 0.006	0.753 ± 0.004	0.694 ± 0.008	0.833 ± 0.037	0.757 ± 0.104
	PCA (2021)	0.753 ± 0.000	0.768 ± 0.000	0.756 ± 0.000	0.239 ± 0.000	0.292 ± 0.000	0.260 ± 0.000
	TTG (2018)	0.919 ± 0.003	0.928 ± 0.002	0.920 ± 0.003	0.257 ± 0.042	0.308 ± 0.015	0.261 ± 0.026
AIDS	PAFIR	0.987 ± 0.003	**0.991 ± 0.007**	**0.989 ± 0.002**	**0.986 ± 0.001**	**0.985 ± 0.006**	**0.986 ± 0.001**
	GraphGPS+HDSE (2024)	0.982 ± 0.002	0.989 ± 0.097	0.985 ± 0.021	0.984 ± 0.006	**0.985 ± 0.004**	0.984 ± 0.096
	TAR 2024	**0.988 ± 0.002**	**0.991 ± 0.001**	0.988 ± 0.002	0.984 ± 0.002	0.984 ± 0.002	0.984 ± 0.002
	SAT (2022)	0.944 ± 0.007	0.982 ± 0.006	0.953 ± 0.004	0.946 ± 0.005	0.933 ± 0.017	0.957 ± 0.094
	PCA (2021)	0.381 ± 0.000	0.615 ± 0.000	0.471 ± 0.000	0.899 ± 0.000	0.899 ± 0.000	0.893 ± 0.000
	TTG (2018)	0.925 ± 0.023	0.947 ± 0.016	0.933 ± 0.020	0.896 ± 0.000	0.899 ± 0.000	0.893 ± 0.000

For time series forecasting, we conduct experiments on Solar-Energy, Traffic, ETT (including ETTh1, ETTh2, ETTm1, and ETTm2), and Weather datasets with a fixed input

sequence length of 96 and prediction lengths varying in {96, 192, 336, 720}. As summarized in Table 9, PAFIR achieves the lowest MSE and MAE in most cases across all datasets, outperforming recent baselines including iTransformer, RLinear, PatchTST, TimesNet, and SCINet. These results demonstrate the model's strong ability to align temporal resolution with evolving feature dynamics. Overall, the consistent superiority of PAFIR in both graph and time series domains highlights its robustness and adaptability for complex learning tasks involving heterogeneous and temporal data.

Table 9: Overall Performance for Time Series Forecasting. We compare extensive competitive methods following the setting of iTransformer (Liu et al., 2023). The input sequence length is set to 96 for all baselines.

Dataset	Seq.	PAFIR (Ours)		iTransformer(2023)		RLinear(2023)		PatchTST(2022)		TimesNet(2022)		SCINet(2022)	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Solar Energy	96	0.213	**0.236**	**0.203**	0.237	0.322	0.339	0.271	0.307	0.250	0.292	0.237	0.344
	192	**0.226**	**0.244**	0.233	0.261	0.359	0.356	0.267	0.310	0.296	0.416	0.281	0.383
	336	**0.243**	**0.258**	0.248	0.273	0.397	0.369	0.290	0.315	0.319	0.432	0.302	0.400
	720	**0.246**	**0.268**	0.250	0.276	0.397	0.356	0.289	0.317	0.338	0.425	0.310	0.400
Traffic	96	**0.412**	**0.272**	0.417	0.276	0.649	0.398	0.462	0.304	0.593	0.321	0.804	0.509
	192	**0.424**	**0.269**	0.428	0.282	0.601	0.366	0.466	0.296	0.617	0.336	0.789	0.505
	336	**0.428**	**0.277**	0.433	0.283	0.609	0.369	0.482	0.304	0.629	0.335	0.800	0.508
	720	**0.451**	**0.296**	0.467	0.301	0.647	0.387	0.514	0.322	0.645	0.351	0.841	0.523
ETTh1	96	**0.381**	**0.398**	0.383	0.405	0.386	0.400	0.414	0.419	0.384	0.402	0.654	0.599
	192	**0.434**	**0.422**	0.441	0.436	0.437	0.424	0.460	0.445	0.436	0.429	0.719	0.631
	336	**0.433**	0.450	0.487	0.458	0.479	**0.446**	0.501	0.466	0.491	0.469	0.778	0.659
	720	0.498	0.487	0.503	0.491	**0.481**	**0.470**	0.500	0.488	0.521	0.500	0.836	0.697
ETTh2	96	**0.288**	0.340	0.297	0.349	**0.288**	**0.338**	0.302	0.348	0.340	0.374	0.707	0.621
	192	0.377	0.398	0.380	0.400	**0.374**	**0.390**	0.388	0.400	0.402	0.414	0.860	0.689
	336	0.422	0.429	0.428	0.432	**0.415**	**0.426**	0.426	0.433	0.452	0.452	1.000	0.744
	720	**0.419**	**0.438**	0.427	0.445	0.420	0.440	0.431	0.446	0.462	0.468	1.249	0.838
ETTm1	96	0.330	**0.366**	0.334	0.368	0.355	0.376	**0.329**	0.367	0.338	0.375	0.418	0.438
	192	0.370	**0.385**	0.377	0.391	0.391	0.392	**0.367**	**0.385**	0.374	0.387	0.439	0.450
	336	0.408	0.413	0.426	0.420	0.424	0.415	**0.399**	**0.410**	0.410	0.411	0.490	0.485
	720	0.484	0.448	0.491	0.459	0.487	0.450	**0.454**	**0.439**	0.478	0.450	0.595	0.550
ETTm2	96	**0.173**	**0.259**	0.180	0.264	0.182	0.265	0.175	**0.259**	0.187	0.267	0.286	0.377
	192	**0.241**	0.305	0.250	0.309	0.246	0.304	**0.241**	**0.302**	0.249	0.309	0.399	0.445
	336	**0.304**	0.343	0.311	0.348	0.307	**0.342**	0.305	0.343	0.321	0.351	0.369	**0.342**
	720	0.407	0.403	0.412	0.407	0.407	**0.398**	**0.402**	0.400	0.408	0.403	0.960	0.735
Weather	96	**0.170**	**0.211**	0.174	0.214	0.192	0.232	0.177	0.218	0.172	0.220	0.221	0.306
	192	**0.206**	**0.244**	0.221	0.254	0.240	0.271	0.225	0.259	0.219	0.261	0.261	0.340
	336	**0.277**	**0.295**	0.278	0.296	0.292	0.307	0.278	0.297	0.280	0.306	0.309	0.378
	720	**0.351**	**0.340**	0.358	0.349	0.364	0.353	0.354	0.348	0.365	0.359	0.377	0.427

Appendix E. Ablation Study

To evaluate the contribution of each key component in our unified architecture and address RQ2, we perform a comprehensive ablation study across both graph-based and time series forecasting tasks. Specifically, we examine the individual impact of (1) the hierarchical graph encoder, (2) the time series encoder, and (3) the multi-agent policy mechanism. Each component is systematically removed or replaced, and the resulting performance is compared against the full model. Experiments are conducted on representative benchmarks,

including node classification, graph classification, and multivariate time series forecasting, to assess the effectiveness and generalizability of each architectural element.

E.1. Impact of Hierarchical Graph Encoding

To evaluate the effectiveness of the hierarchical graph encoder in capturing structural dependencies, we perform an ablation by removing this component from the PAFIR architecture. As shown in Table 10, the performance drops significantly across both node and graph classification tasks on the PROTEINS, ENZYMES, and AIDS datasets.

Table 10: Ablation Experiment for Hierarchical Graph Encoding.

Dataset	Model Variant	Node Classification			Graph Classification		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
PROTEINS	PAFIR (Full)	**0.921 ± 0.007**	**0.914 ± 0.007**	**0.918 ± 0.009**	**0.843 ± 0.006**	**0.822 ± 0.007**	**0.832 ± 0.006**
	w/o Hierarchical Graph Encoding	0.678 ± 0.005	0.673 ± 0.013	0.675 ± 0.020	0.765 ± 0.033	0.757 ± 0.005	0.766 ± 0.003
ENZYMES	PAFIR (Full)	**0.937 ± 0.004**	**0.940 ± 0.007**	**0.941 ± 0.006**	**0.866 ± 0.006**	**0.848 ± 0.007**	**0.857 ± 0.006**
	w/o Hierarchical Graph Encoding	0.926 ± 0.006	0.923 ± 0.005	0.927 ± 0.006	0.340 ± 0.009	0.344 ± 0.007	0.348 ± 0.001
AIDS	PAFIR (Full)	**0.987 ± 0.003**	**0.991 ± 0.007**	**0.989 ± 0.002**	**0.986 ± 0.001**	**0.985 ± 0.006**	**0.986 ± 0.001**
	w/o Hierarchical Graph Encoding	0.896 ± 0.004	0.863 ± 0.006	0.884 ± 0.001	0.808 ± 0.006	0.820 ± 0.003	0.814 ± 0.009

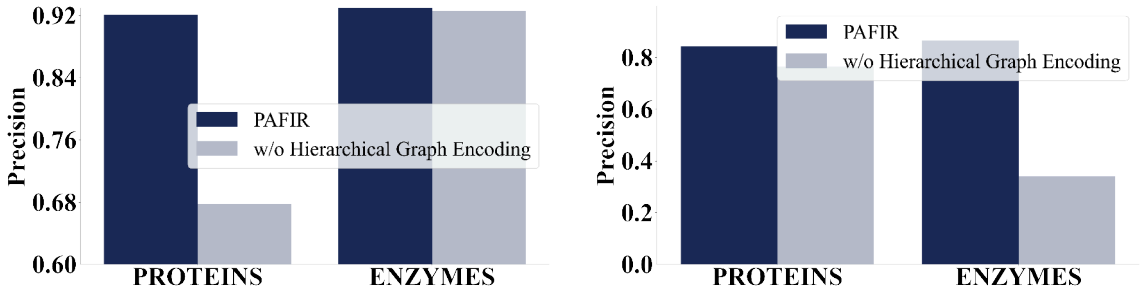
On PROTEINS, the F1 score for node classification drops from 0.918 to 0.675, while graph classification performance also declines from 0.832 to 0.766. The impact is even more pronounced on the ENZYMES dataset, where graph classification F1 decreases drastically from 0.857 to 0.348. On the AIDS dataset, although the overall performance remains high, the consistent drop in F1 scores for both node and graph classification after removing hierarchical encoding further confirms its role in stabilizing and enhancing structural representation learning. This suggests that hierarchical structural information plays a crucial role in enabling PAFIR to identify task-relevant topological and semantic patterns.

The results confirm that incorporating multi-level structural representations allows the model to generalize better across diverse graph distributions and enhances its capacity to model complex relationships between features and labels.

E.2. Impact of Temporal Encoding

To assess the contribution of the time series encoding module in capturing temporal dynamics, we remove it from the PAFIR architecture and compare performance across four forecasting benchmarks: Solar-Energy, Traffic, ETT, and Weather. As shown in Table 11, removing the temporal encoder results in substantial performance degradation across all datasets.

For instance, on the Solar-Energy dataset with a sequence length of 96, removing time series encoding causes the MSE to increase from 0.213 to 0.716 and the MAE from 0.236 to 0.754. Consistent and substantial performance degradations are also observed across Traffic, ETTh1, ETTh2, ETTm1, ETTm2, and Weather datasets, where both MSE and MAE increase markedly, in several cases by nearly two to three times. These results demonstrate



(a) Node Classification (b) Graph Classification

Figure 7: Ablation experiment on the graph domain. We report Precision under two tasks: node classification and graph classification.

Table 11: Ablation Experiment for Time Series Encoding. The input sequence length is fixed to 96 for all methods.

Model Variant	Solar-Energy		Traffic		ETTh1		ETTh2		ETTm1		ETTm2		Weather	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PAFIR (Full)	**0.213**	**0.236**	**0.412**	**0.272**	**0.381**	**0.398**	**0.288**	**0.340**	**0.330**	**0.366**	**0.173**	**0.259**	**0.170**	**0.211**
w/o Time Series Encoding	0.716	0.754	0.788	0.761	0.725	0.739	0.733	0.748	0.656	0.671	0.459	0.470	0.433	0.451

that the temporal encoding module plays a crucial role in capturing sequential dependencies and aligning feature evolution with predictive accuracy across diverse time-series forecasting tasks.

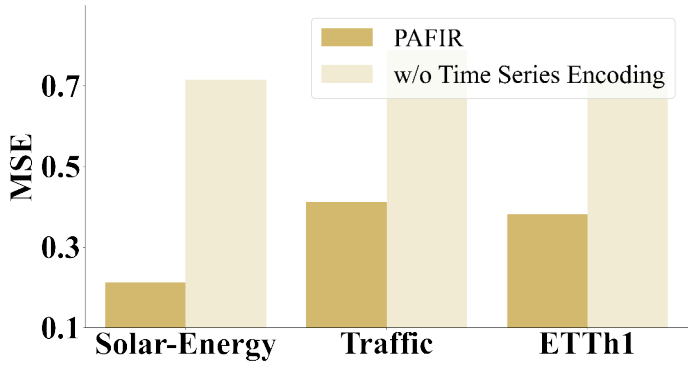


Figure 8: Ablation Experiment on Time Series Domain. We report MSE under Three Datasets.

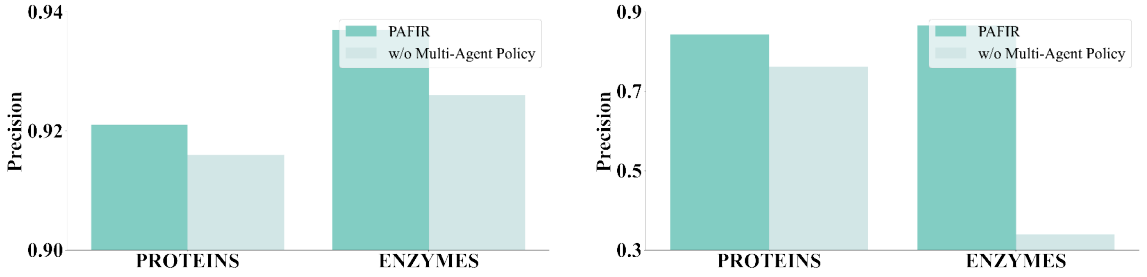
E.3. Impact of Multi-Agent Policy

To investigate the role of the multi-agent policy in PAFIR, we conduct an ablation by removing this component and evaluating performance on both node and graph classification tasks. As presented in Table 12, removing the multi-agent policy leads to a notable drop

in graph classification performance, while the impact on node classification is relatively modest.

Table 12: Ablation Experiment for Multi-Agent Module in RL Framework.

Dataset	Model Variant	Node Classification			Graph Classification		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
PROTEINS	PAFIR (Full)	**0.921 ± 0.007**	**0.914 ± 0.007**	**0.918 ± 0.009**	**0.843 ± 0.006**	**0.822 ± 0.007**	**0.832 ± 0.006**
	w/o Multi-Agent Policy	0.916 ± 0.003	0.913 ± 0.002	0.917 ± 0.002	0.761 ± 0.010	0.760 ± 0.007	0.759 ± 0.007
ENZYMES	PAFIR (Full)	**0.937 ± 0.004**	**0.940 ± 0.007**	**0.941 ± 0.006**	**0.866 ± 0.006**	**0.848 ± 0.007**	**0.857 ± 0.006**
	w/o Multi-Agent Policy	0.926 ± 0.003	0.932 ± 0.002	0.927 ± 0.003	0.339 ± 0.040	0.348 ± 0.002	0.306 ± 0.004
AIDS	PAFIR (Full)	**0.987 ± 0.003**	**0.991 ± 0.007**	**0.989 ± 0.002**	**0.986 ± 0.001**	**0.985 ± 0.006**	**0.986 ± 0.001**
	w/o Multi-Agent Policy	0.934 ± 0.006	0.948 ± 0.002	0.940 ± 0.001	0.453 ± 0.011	0.488 ± 0.002	0.474 ± 0.005



(a) Node Classification (b) Graph Classification

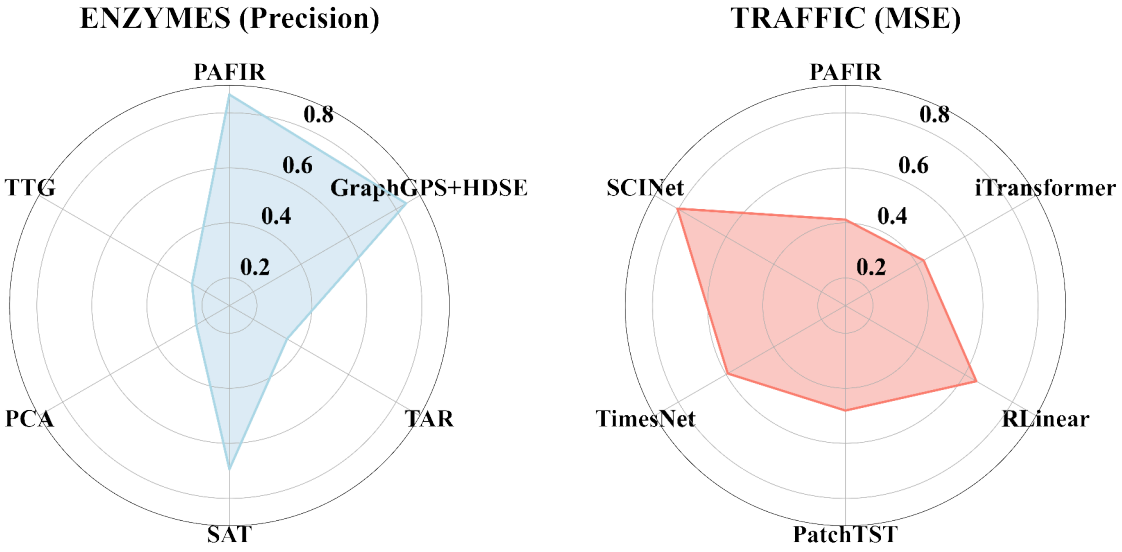
Figure 9: Ablation experiment on the graph domain. We report Precision under two tasks: node classification and graph classification.

For example, on the PROTEINS dataset, the F1 score for graph classification drops from 0.832 to 0.759 when the multi-agent policy is excluded. A similar but more severe degradation is observed on the ENZYMES dataset, where the F1 score decreases from 0.857 to 0.306. Similarly, on the AIDS dataset, removing the multi-agent policy leads to a substantial decline in graph classification performance from 0.986 to 0.474. This contrast indicates that the multi-agent strategy plays a critical role in guiding feature transformation and aggregation at the graph level, where complex interactions between feature subsets are more pronounced.

These results confirm that the cooperative mechanism among agents enables PAFIR to better explore and select informative features, ultimately enhancing its capacity for structured representation learning.

Summary of Ablation Findings Across all ablation studies, we observe that each core component of PAFIR contributes uniquely and substantially to the model's overall performance. The hierarchical graph encoder is essential for capturing multi-level structural patterns, with its removal leading to significant performance degradation in both node and graph classification. The time series encoding module plays a key role in modeling temporal patterns; excluding it results in dramatically increased forecasting errors across all benchmarks. Finally, the multi-agent policy proves particularly effective in guiding feature

selection at the graph level, as evidenced by the sharp drop in graph classification performance when this component is removed. Collectively, these results demonstrate that the integrated design of structural encoding, temporal modeling, and the multi-agent module in the reinforcement learning framework is key to PAFIR's potential in learning expressive and generalizable representations across diverse data modalities.



(a) ENZYMES: Precision

(b) TRAFFIC: MSE

Figure 10: Experiment Result of Different Methods on Two Datasets.

Appendix F. Cross-Domain Generality Results

We apply the proposed PAFIR framework to two representative tasks, graph-based classification and multivariate time series forecasting, to evaluate its generality across heterogeneous data modalities. Specifically, we test PAFIR on the ENZYMES dataset, which involves structural learning over biological graph data, and the TRAFFIC dataset, which captures temporal dynamics in real-world sensor measurements (Fig. 10). These benchmarks represent two fundamentally different problem settings and data characteristics, allowing us to assess the robustness and flexibility of our unified framework. These two tasks not only differ in terms of data modality, structured graphs versus sequential time series, but also in the nature of their feature dependencies and temporal or topological patterns. On ENZYMES, PAFIR is compared with state-of-the-art graph learning models, including GraphGPS+HDSE, TAR, SAT, PCA, and TTG. On TRAFFIC, we benchmark PAFIR against competitive forecasting models such as iTransformer, RLinear, PatchTST, TimesNet, and SCINet.

On the ENZYMES dataset (Fig. 10(a)), PAFIR achieves the highest precision among all competing graph-based models. This result highlights PAFIR's ability to capture discriminative structural features and maintain high predictive accuracy in complex biological graph classification tasks. On the TRAFFIC dataset (Fig. 10(b)), where the evaluation

metric is MSE, PAFIR again outperforms temporal forecasting baselines. Its consistently lower MSE reflects the model's robustness in capturing temporal dependencies and adapting to traffic dynamics.

The consistent gains across both structural and temporal settings highlight the effectiveness of PAFIR's unified design in handling heterogeneous data. This cross-domain generality underscores its practical value in real-world scenarios where multimodal health, behavioral, or sensor data are commonly encountered.