JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Panneerselvam, A., Yaakob, R., Perumal, T., & Marlisah, E. (2026). A cross-dataset adaptive hybrid ensemble framework for robust inertial fall detection. *Journal of Information and Communication Technology*, *25*(4), 92-119. https://doi.org/10.32890/jict2026.25.4.5

A Cross-Dataset Adaptive Hybrid Ensemble Framework for Robust Inertial Fall Detection

[1]Athimahendran Panneerselvam, [2]Razali Yaakob,
[3]Thinagaran Perumal & [4]Erzam Marlisah
[1,2,3&4]Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, Malaysia

[*1]gs49868@upm.edu.my
[2]razaliy@upm.edu.my
[3]thinagaran@upm.edu.my
[4]erzam@upm.edu.my
*Corresponding author

Received: 15/4/2026 Revised: 24/9/2026 Accepted: 12/10/2026 Published: 31/10/2026

ABSTRACT

Falls remain a major healthcare concern, particularly among older adults, since they frequently result in serious injuries, prolonged rehabilitation, and reduced quality of life. Although numerous inertial sensor-based fall detection methods have been reported, many are validated on a single dataset, limiting their ability to generalise across diverse sensing environments in real-world applications. This study proposes a cross-dataset Adaptive Hybrid Ensemble Framework (AHEF) for robust wearable fall detection. The framework combines heterogeneous Machine Learning (ML) classifiers Random Forest (RF), XGBoost (XGB), Extra Trees (ET), and Elastic Net Logistic Regression (ENet-LR) within a probability-calibrated stacking architecture, followed by adaptive decision refinement using activity-aware thresholds and temporal confirmation. Performance was evaluated using three publicly available benchmark datasets (SisFall, MobiFall & KFall) under both subject-independent cross-validation and Leave-One-Dataset-Out (LODO) protocols. Experimental results indicate that the proposed framework achieved 98.7% accuracy, 99.1% sensitivity, and 98.3% specificity during within-dataset evaluation while maintaining strong cross-dataset performance, with average sensitivity exceeding 98% across unseen datasets. The framework also reduced the False Positive Rate (FPR) to 1.7%, compared with 4.0–6.4% obtained by the individual base classifiers, without compromising detection sensitivity.

Overall, these findings demonstrate that combining cross-dataset evaluation with adaptive ensemble learning improves both prediction reliability and generalisation, supporting deployment of wearable fall detection systems in heterogeneous real-world environments.

Keywords: Ensemble learning, fall detection, feature engineering, hybrid model, stacking classifier.

INTRODUCTION

Falls are a leading cause of injury among older adults and remain a significant public health concern, resulting in substantial numbers of emergency department visits, hospitalisations, and fall-related deaths each year (Centers for Disease Control and Prevention, CDC, 2026a, 2026b). Beyond their immediate physical consequences, falls often lead to long-term psychological and behavioural effects, including fear of falling, reduced physical activity, and loss of independence, which further increase the risk of subsequent falls (CDC, 2026b; Gießler et al., 2024). As wearable technologies and smart healthcare systems become increasingly integrated into daily life, demand is growing for fall detection systems that provide accurate, continuous, and reliable monitoring under real-world conditions while balancing detection sensitivity with false alarm rates (Subramaniam et al., 2022).

Although wearable fall detection has advanced considerably, the evaluation strategies adopted in many studies remain limited. Most models are trained and validated on a single dataset, which may allow classifiers to learn dataset-specific characteristics and produce optimistic performance estimates while limiting their ability to generalise to unseen environments (Hong et al., 2024; Qin et al., 2023). Recent machine learning research has raised similar concerns regarding limited and imbalanced datasets, where restricted training distributions can contribute to overfitting and reduced model generalisation (Yusof & Fajila, 2026). In practice, variations in participant characteristics, sensor placement, acquisition devices, and activity protocols introduce substantial dataset shift, often degrading performance when models are deployed beyond the data on which they were trained (Hong et al., 2024; Kasim & Sheppard, 2023).

Confidence calibration presents another challenge for safety-critical fall detection systems. Many machine learning classifiers produce probability estimates that do not accurately reflect the likelihood that their predictions are correct (Silva Filho et al., 2023; scikit-learn developers, 2026). In healthcare applications, poorly calibrated confidence estimates may lead to excessive false alarms that contribute to alarm fatigue or insufficient confidence in genuine fall events, delaying timely intervention (Silva Filho et al., 2023; Pavlovic, 2025).

To address these challenges, this study proposes a cross-dataset AHEF that improves both prediction robustness and decision reliability. The proposed architecture combines heterogeneous ensemble learning with probability calibration to exploit complementary classifier characteristics while producing more reliable confidence estimates for safety-critical decision-making (Li et al., 2025; Priya & Amshakala, 2025; Silva Filho et al., 2023; scikit-learn developers, 2026). Unlike conventional approaches that rely primarily on within-dataset evaluation, AHEF is assessed using subject-independent cross-validation and cross-dataset validation. The cross-dataset evaluation is conducted across three publicly available inertial datasets (SisFall, MobiFall, & KFall) to provide a more realistic assessment of robustness across heterogeneous sensing environments (BIng2325, 2023; BMI-HMU, 2026; Hong et al., 2024; Qin et al., 2023).

This study's primary contribution is a hierarchical adaptive learning architecture that integrates heterogeneous multi-model base learning, probability calibration, Adaptive Learning Module I (Artificial Intelligence (AI) Base Refinement), stacking meta-learning, and Adaptive Learning Module II (AI Meta Refinement) within a unified prediction framework. It also incorporates uncertainty-aware meta-features and physics-informed motion descriptors, including the Fall Morphology Index (FMI) and Spectral Burstiness (SB), to strengthen ensemble fusion while reducing false-positive detections. Finally, the proposed framework applies activity-aware decision refinement with temporal confirmation. It is comprehensively evaluated using a Leave-One-Dataset-Out (LODO) protocol across the SisFall, MobiFall, and KFall datasets, providing a rigorous assessment of generalisation under dataset shift (Hong et al., 2024; Qin et al., 2023).

RELATED WORKS

Fall detection research has evolved from conventional threshold-based techniques to machine learning and Deep Learning (DL) approaches using wearable inertial sensor data. Although threshold-based methods are computationally efficient and well suited for resource-constrained devices, their dependence on predefined decision rules limits adaptability to diverse movement patterns. Furthermore, this often increases susceptibility to false alarms during real-world deployment (Subramaniam et al., 2022; Li et al., 2025).

Traditional machine learning methods, particularly Support Vector Machines (SVMs) and Random Forest (RF), improve discrimination between falls and Activities of Daily Living (ADLs). However, models trained on a single dataset often generalise poorly to unseen sensing environments (Subramaniam et al., 2022; Hong et al., 2024). Recent deep learning models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM), learn discriminative motion representations directly from raw sensor signals, though they require larger labelled datasets and greater computational resources. This, in turn, limits their suitability for wearable and edge devices (Su et al., 2024; Li et al., 2025).

Improving model generalisation across heterogeneous datasets has therefore become a key research focus. Variations in sensor placement, acquisition devices, participant populations, and data collection protocols can introduce domain shifts that degrade predictive performance when models are evaluated on unseen datasets that differ from the training data (Hong et al., 2024; Kasim & Sheppard, 2023; Palli et al., 2024). To address this challenge, recent studies in human activity recognition have investigated domain generalisation, transferable feature representations, and representation learning techniques. These aim to preserve discriminative motion characteristics across diverse sensing environments (Qin et al., 2023; Hong et al., 2024).

Beyond generalisation challenges, effective fall detection also depends on capturing the biomechanical characteristics that distinguish genuine falls from high-intensity ADLs. Falls commonly involve abrupt motion transitions, elevated impact intensity, and periods of post-impact inactivity. In contrast, many ADLs exhibit only partial components of this sequence. Although many fall detection approaches use statistical and spectral signal descriptors, these fall-specific motion characteristics have received comparatively less attention in adaptive machine learning frameworks. Thus, integrating motion morphology information may provide additional discriminative power to reduce false alarms while preserving sensitivity to genuine fall events.

Ensemble learning improves predictive robustness by integrating the complementary strengths of multiple classifiers (Sainin et al., 2021). Previous studies have reported that ensemble-based methods can reduce false-positive detections and improve classification performance relative to individual models, particularly in complex real-world environments (Priya & Amshakala, 2025; Li et al., 2025). Nevertheless, many existing ensemble frameworks primarily emphasise classification accuracy while giving limited attention to probability calibration and the reliability of prediction confidence estimates. Consequently, uncalibrated prediction probabilities may remain overconfident or underconfident despite high classification performance, reducing their suitability for safety-critical healthcare applications (Silva Filho et al., 2023; scikit-learn developers, 2026).

To provide a structured overview of recent developments in cross-dataset generalisation, adaptive learning, and ensemble-based fall detection, Table 1 summarises representative studies, their methodological focus, key findings, and remaining research gaps.

Table 1

Comparison of Cross-Dataset Generalisation, Adaptive Learning, and Ensemble-Based Fall Detection Studies

Study	Category	Method/ Approach	Key findings	Identified research gap
Qin et al. (2023)	Cross-Dataset Generalisation	Adaptive Feature Fusion (AFFAR)	Improved transferability across heterogeneous HAR datasets	Not fall-specific; no calibrated ensemble learning
Kasim & Sheppard (2023)	Cross-Domain Adaptation	BLSTM + similarity-based domain adaptation	Improved cross-domain HAR performance	Requires target-domain information; limited fall-detection focus
Hong et al. (2024)	Cross-Dataset Generalisation	CrossHAR hierarchical self-supervised learning	Improved robustness under dataset shift	Focused on HAR rather than fall detection
Silva et al. (2024)	Cross-Dataset Fall Detection Evaluation	Evaluation of 13 AI-based wearable fall-detection models across datasets	Significant performance degradation under cross-dataset and realistic conditions	Identifies the generalisation problem but does not provide an adaptive ensemble solution
Sainin et al. (2021)	Ensemble Learning	Meta-classifier with sampling and feature selection	Improved classification under imbalance	Does not address fall detection or dataset shift
Palli et al. (2024)	Adaptive Learning	Review of concept drift and class-imbalance adaptation	Highlights challenges of adaptation in evolving data streams	No fall-specific adaptive ensemble framework
Gangadhar et al. (2025)	Ensemble Fall Detection	RF–SVM ensemble evaluated on KFall	Demonstrated benefits of classifier diversity	Limited to single-dataset evaluation; no calibration or LODO validation
Priya & Amshakala (2025)	Adaptive Ensemble Fall Detection	YOLOv8–RetinaNet–DETR ensemble	Improved robustness through detector fusion	Vision-based; no inertial cross-dataset evaluation or probabilistic calibration

The studies summarised in Table 1 demonstrate substantial progress in cross-dataset learning, adaptive learning, and ensemble-based fall detection. Cross-dataset research has shown that domain shifts caused by differences in sensing devices, participant populations, and acquisition protocols can significantly degrade model performance when models are evaluated on unseen datasets. In particular, Silva et al. (2024) reported substantial performance deterioration when wearable fall-detection algorithms were evaluated under cross-dataset and realistic deployment conditions, highlighting the limitations of relying solely on within-dataset validation.

Meanwhile, ensemble learning and adaptive learning studies have demonstrated that combining complementary learners and adaptive mechanisms can improve predictive robustness. However, limited research has integrated cross-dataset evaluation, heterogeneous ensemble learning, probability calibration, uncertainty-aware refinement, and fall-specific motion morphology within a unified wearable inertial fall-detection framework. Collectively, these limitations motivate the development of the proposed AHEF.

Despite these advances, three important research gaps remain. First, although cross-dataset learning and domain generalisation have received increasing attention, many wearable fall-detection systems continue to rely on within-dataset validation, providing limited evidence of robustness when deployed across heterogeneous sensing environments. Consequently, the ability of many models to maintain performance under unseen sensing conditions remains insufficiently investigated. Second, ensemble and adaptive-learning approaches have shown considerable promise, yet limited work has combined heterogeneous ensemble learning with probability calibration and uncertainty-aware adaptive refinement. Third, existing inertial fall-detection systems primarily employ conventional statistical and spectral descriptors, while fall-specific motion morphology information remains underutilised in adaptive ensemble decision-making. These gaps motivate the development of a unified framework capable of integrating cross-dataset robustness, heterogeneous ensemble learning, probabilistic reliability, adaptive refinement, and fall-specific motion modelling within a single wearable inertial fall-detection architecture.

To address these limitations, this study proposes a cross-dataset Adaptive Hybrid Ensemble Framework (AHEF) that combines heterogeneous ensemble learning, probability calibration, uncertainty-aware adaptive refinement, and physics-informed motion descriptors. The framework is further evaluated using subject-independent validation and Leave-One-Dataset-Out (LODO) protocols across the SisFall, MobiFall, and KFall benchmark datasets to assess predictive robustness and decision reliability under heterogeneous sensing conditions.

THE PROPOSED METHOD

Adaptive Hybrid Ensemble Framework

This study proposes an AHEF for inertial fall detection using wearable sensor data. The framework follows a hierarchical learning structure that transforms raw inertial measurements into fall predictions through sequential signal processing, feature representation, probabilistic learning, and decision refinement stages. Instead of relying on a single classifier or direct ensemble aggregation, AHEF combines multi-domain feature engineering, heterogeneous base learning, probability calibration, stacking-based fusion, and adaptive refinement. These aim to improve robustness and reduce false alarms under heterogeneous sensing conditions.

Figure 1

High-Level Flow of the Proposed AHEF

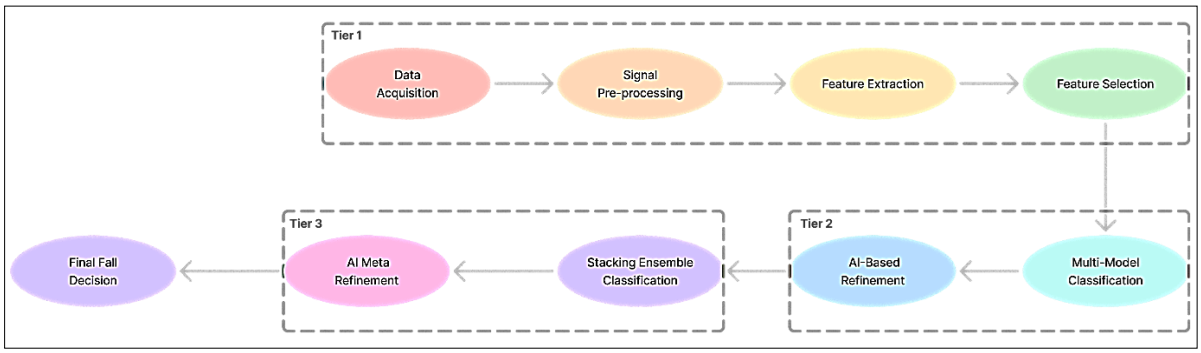


Figure 1 presents a high-level flowchart of the proposed framework and illustrates AHEF's three-tier structure. Tier 1 performs signal pre-processing and feature engineering, while Tier 2 conducts multi-model learning and AI-based refinement. In contrast, Tier 3 performs stacking-based fusion and adaptive decision refinement to generate the final fall/non-fall prediction.

Data Acquisition and Pre-processing

The proposed framework utilises tri-axial accelerometer data collected using publicly available wearable and smartphone-based inertial sensors. These measurements provide information on the dynamic behaviour of human movement relevant to fall detection. However, variations in sensing hardware, device orientation, sampling conditions, and sensor placement introduce substantial heterogeneity, which may compromise feature consistency and classification performance. Hence, to minimise these sources of variability, a pre-processing stage establishes a standardised signal representation before feature engineering.

The pre-processing pipeline first normalises the raw accelerometer measurements to reduce scale variations while preserving the motion patterns associated with fall events. The pipeline then segments the continuous time-series data into fixed-length, overlapping windows, generating motion segments suitable for supervised learning. This windowing strategy isolates short-duration fall events while retaining the temporal information required for subsequent analysis. The continuous inertial signal is represented as Equation 1.

$$X = x_1, x_2, \dots, x_N, \tag{1}$$

where X = continuous inertial signal, x_i = inertial measurement at the ith sampling instant, and N = total number of sampled observations. Each segmented motion window is defined as Equation 2.

$$W_i = x_i, x_{(i+1)}, \dots, x_{(i+L-1)}, \tag{2}$$

where $W_i = i^{th}$ segmented motion window, x_i = starting sample of the window, L = window length, and i = window index.

To reduce the influence of sensor orientation and device placement, the resultant acceleration magnitude is computed from the three orthogonal accelerometer axes, presented as Equation 3.

$$M(t) = \sqrt{\left(a_x(t)^2 + a_y(t)^2 + a_z(t)^2\right)}, \quad (3)$$

where $a_x(t), a_y(t), and\ a_z(t)$ denote the acceleration measurements along the x-, y-, and z-axes, respectively. The magnitude representation provides an orientation-independent measure of motion intensity, allowing the subsequent feature engineering stage to focus on intrinsic movement characteristics rather than sensor orientation. The resulting standardised motion windows are subsequently used as input for the multi-domain feature engineering process.

Multi-Domain Feature Engineering

Following signal pre-processing, each motion segment is converted into a numerical feature vector using a multi-domain feature engineering strategy. Rather than relying on a single descriptor category, the framework combines time-domain, frequency-domain, and motion-dynamic features to characterise movement patterns associated with falls and ADLs. Integrating these complementary feature domains provides a more informative representation by jointly capturing statistical properties, spectral information, and biomechanical motion characteristics. The feature vector represents each segmented motion window as depicted in Equation 4.

$$x_i = [f_1, f_2, \dots, f_n]^T, d = 29, \quad (4)$$

where x_i = feature vector extracted from the i^{th} motion window, $f_j = j^{th}$ engineered feature, $d =$ total number of engineered features $(d = 29)$, and $T =$ transpose operator. The resulting feature vector offers a compact description of the extracted motion characteristics. It is subsequently supplied to the AHEF for model learning.

Table 2 summarises the complete feature set extracted from each segmented inertial window.

Table 2

Extracted Feature Set Used for Model Training

Category	Features	Count
Time-domain	Mean, variance, standard deviation, skewness, kurtosis, minimum, maximum, range, Root Mean Square (RMS), zero-crossing rate, Signal Magnitude Area (SMA)	11
Frequency-domain	Spectral energy, spectral entropy, dominant frequency, spectral centroid, first 10 Fast Fourier Transform (FFT) coefficients	14
Motion-dynamic	Impact intensity (peak acceleration), post-impact inactivity duration, FMI, SB	4
Total		29

The extracted features capture complementary statistical, spectral, and biomechanical characteristics of human motion. Together, these feature domains provide a comprehensive representation of both fall and non-fall activities, including high-intensity ADLs, by capturing statistical signal properties and

higher-level biomechanical behaviour. This integrated feature representation improves discrimination between fall events and non-fall activities and provides robust, generalisable inputs for the AHEF across heterogeneous sensing environments.

Feature Selection

After feature extraction, a feature selection stage eliminates redundant and weakly informative attributes while preserving the most discriminative motion descriptors for classification. The selection process combines statistical filtering with model-based selection to improve feature compactness without sacrificing predictive performance. This approach aligns with recent Journal of Information and Communication Technology (JICT) research demonstrating that appropriate feature selection can improve the effectiveness and efficiency of multiple machine learning classifiers, including RF, XGBoost (XGB), and Logistic Regression (Rafrastara et al., 2025).

Initially, low-variance features and highly correlated feature pairs are removed to reduce redundancy and minimise multicollinearity. Next, Logistic Regression with L1 regularisation (LASSO) promotes sparse feature selection by shrinking insignificant feature coefficients toward zero. Feature importance scores generated by RF and XGB are then used to validate the contribution of the retained descriptors. The optimisation objective is defined in Equation 5.

$$\frac{\min}{w} L(w) + \lambda \|w\|_1, \tag{5}$$

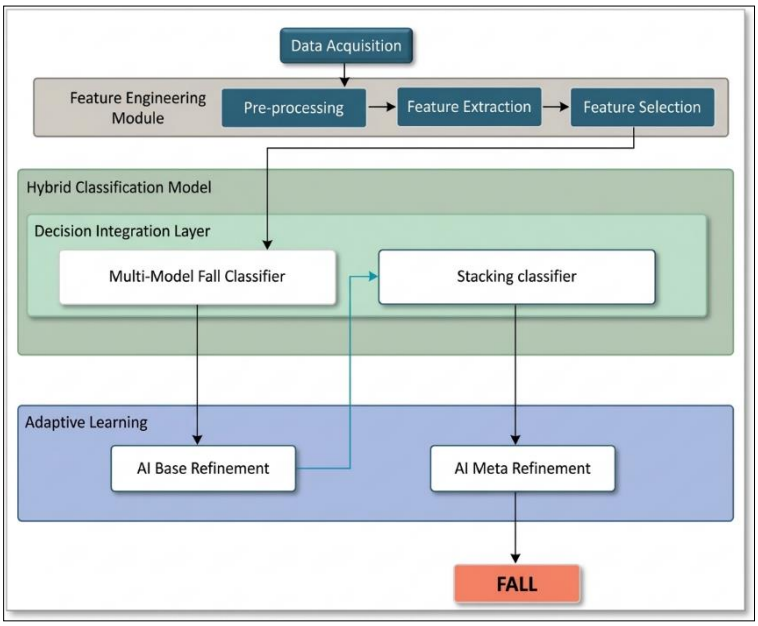
where L(w) = logistic loss function, w = model coefficient vector, λ = regularisation parameter controlling the sparsity, and $\| w \|_1$ = L1 norm of the coefficient vector that encourages sparse feature selection.

Finally, the selected feature subset is evaluated using the heterogeneous base learners to verify that the retained descriptors preserve both discriminative capability and probability calibration. The resulting feature vector provides a compact and robust representation of human motion for the proposed AHEF.

The AHEF serves as the core prediction component of the fall detection framework. Using the selected motion features, the model applies a hierarchical learning strategy to generate fall predictions through a sequence of learning and refinement stages. Unlike conventional ensemble methods that directly aggregate classifier outputs, AHEF performs probabilistic refinement before and after ensemble fusion through two adaptive learning modules. This design enhances prediction robustness, probability calibration, and cross-dataset generalisation. Figure 2 illustrates the architectural organisation of AHEF, which comprises three main modules: Feature Engineering, Hybrid Classification, and Adaptive Learning. Together, these modules progressively transform inertial sensor measurements into probabilistic fall predictions through feature extraction, ensemble learning, and adaptive decision refinement.

Figure 2

Architectural Overview of the Proposed AHEF



The AHEF is organised into three functional architectural layers. The Feature Engineering Module transforms raw inertial sensor measurements into structured feature representations through data acquisition, signal pre-processing, feature extraction, and feature selection. The Hybrid Classification Module performs multi-model base learning and ensemble integration to generate reliable probabilistic predictions from heterogeneous machine learning classifiers. The Adaptive Learning Module incorporates AI Base Refinement and AI Meta Refinement, which progressively enhance prediction reliability through uncertainty-aware refinement, contextual reasoning, temporal verification, and physics-informed decision support.

Table 3 summarises each architectural component's role, including its inputs, internal processing mechanisms, outputs, and methodological contributions. Together, these components progressively refine information from engineered motion features to the final fall-detection decision.

Table 3

Functional Architecture of the Proposed AHEF

Computational layer	Input	Internal processing	Output	Methodological contribution
Multi-Model Base Learning	Selected feature vector	Heterogeneous machine learning classifiers	Calibrated probability vector	Exploits complementary classifier diversity for robust probability estimation

(continued)

Computational layer	Input	Internal processing	Output	Methodological contribution
Adaptive Learning Module I (AI Base Refinement)	Calibrated probability vector	Consensus analysis, uncertainty quantification, entropy estimation, contextual, and morphology-aware feature enrichment	Enriched meta-feature vector	Performs uncertainty-aware probabilistic refinement before ensemble fusion
Stacking Meta-Learning	Calibrated probabilities and enriched meta-features	Regularised Logistic Regression	Hybrid probability estimate	Learns adaptive weighting of complementary classifier behaviours
Adaptive Learning Module II (AI Meta Refinement)	Hybrid probability estimate	Activity-aware threshold adaptation, temporal consistency verification, and FMI-based physics-informed validation	Final fall prediction	Performs adaptive decision reasoning for robust and reliable fall detection

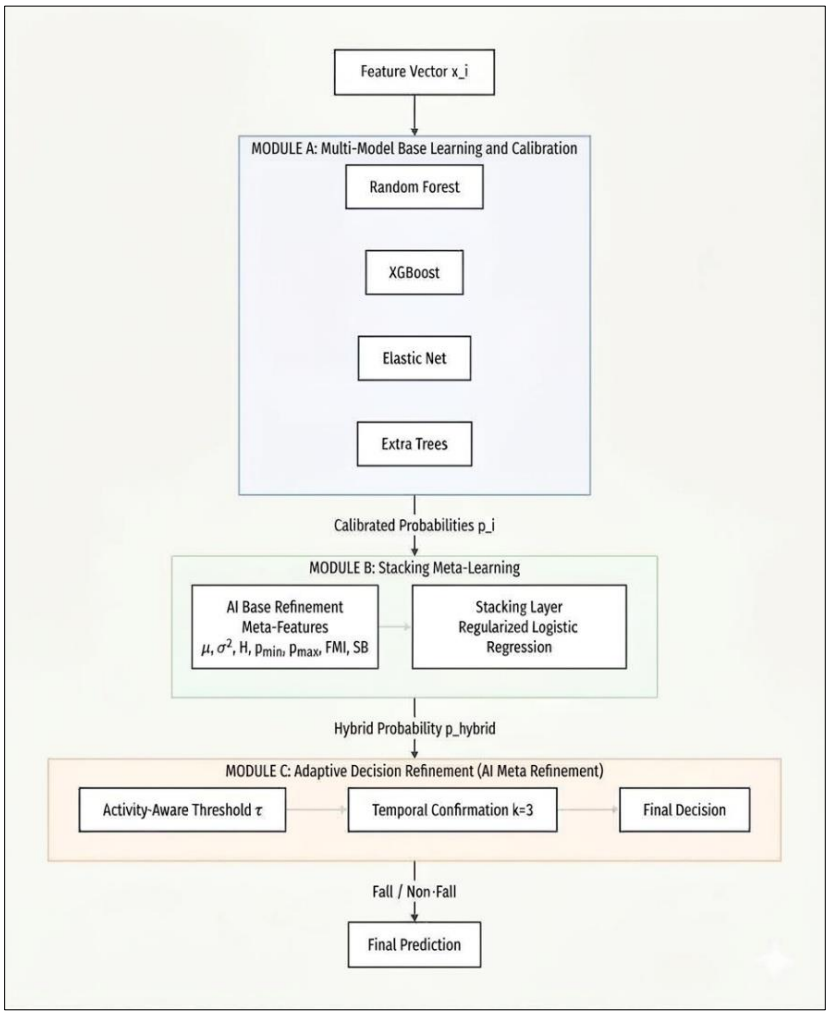
Framework Implementation

The AHEF follows the hierarchical architecture illustrated in Figure 2, with each module performing a specific function in the prediction pipeline. Rather than directly aggregating classifier outputs, the framework applies successive adaptive learning stages to improve probability reliability and decision confidence before generating the final fall prediction.

Figure 3 presents the detailed implementation architecture of AHEF. Module A performs heterogeneous base learning and probability calibration, while Module B performs uncertainty-aware feature enrichment and stacking meta-learning. Module C then applies adaptive decision refinement to generate the final fall prediction. Figures 2 and 3 present the proposed AHEF at two complementary levels of abstraction. Figure 2 presents the framework's conceptual architecture, illustrating the major computational layers and their functional roles, while Figure 3 details the technical implementation through three operational modules. Specifically, Module A implements the Multi-Model Base Learning layer shown in Figure 2 and handles heterogeneous classifier learning and probability calibration. Module B integrates Adaptive Learning Module I (AI Base Refinement) with the Stacking Meta-Learning layer, where calibrated classifier probabilities are transformed into uncertainty-aware and context-informed meta-features before being integrated through the stacking classifier. Module C implements Adaptive Learning Module II (AI Meta Refinement), which performs activity-aware threshold adaptation, temporal consistency verification, and final fall-decision generation. Therefore, Figure 3 represents the implementation-level expansion of the conceptual architecture presented in Figure 2, with the three modules collectively implementing the computational layers of the proposed AHEF.

Figure 3

Detailed Technical Architecture of the Proposed AHEF



The technical implementation of AHEF is organised into three functional modules. Module A performs multi-model base learning and probability calibration using heterogeneous machine-learning classifiers. Module B implements Adaptive Learning Module I (AI Base Refinement) and stacking meta-learning, where calibrated classifier probabilities are transformed into uncertainty-aware meta-features and subsequently fused through a stacking classifier. Module C implements Adaptive Learning Module II (AI Meta Refinement), which refines the hybrid prediction using activity-aware threshold adaptation, temporal consistency verification, and physics-informed decision reasoning to generate the final fall prediction.

Module A: Multi-Model Base Learning and Calibration

Module A generates calibrated probability estimates using four heterogeneous classifiers: RF, XGB, Extra Trees (ET), and Elastic Net Logistic Regression (ENet-LR). The resulting probability outputs provide complementary representations of fall likelihood for subsequent ensemble fusion.

Rather than relying on a single classifier, the proposed framework leverages the diversity of multiple learning paradigms to generate complementary decision boundaries and probability-estimation behaviours. The outputs from the heterogeneous base learners are then calibrated to obtain reliable confidence estimates that more accurately reflect the likelihood of fall events. These calibrated probabilities are then provided as inputs to the subsequent adaptive learning and ensemble fusion processes implemented within the framework.

Module B: Stacking Meta-Learning and Adaptive Learning Module I

Module B performs adaptive probabilistic refinement and ensemble fusion. Adaptive Learning Module I (AI Base Refinement) transform calibrated classifier outputs into uncertainty-aware and morphology-informed meta-features, including the ensemble mean (μ), variance (σ^2), entropy (H), minimum, and maximum probabilities ($p_{\min}, p_{\max}$), FMI, and SB. These enriched meta-features are subsequently integrated using a Regularized Logistic Regression stacking classifier to generate a fused probability estimate for the subsequent adaptive decision refinement stage.

Module C: Adaptive Decision Refinement and Adaptive Learning Module II

Module C represents the final decision-making stage of the AHEF. The stacking classifier's hybrid probability is further refined through Adaptive Learning Module II (AI Meta Refinement), which incorporates activity-aware threshold adaptation and temporal consistency verification to improve decision reliability. The refined prediction is subsequently used to generate the final fall/non-fall prediction.

Table 4 summarises each classifier's learning paradigm, primary strengths, and methodological justification. All base learners are independently trained using the subject-independent cross-validation protocol described in the experimental setup.

Table 4

Heterogeneous Base Learners Used in AHEF

Classifier	Paradigm	Justification
Random Forest (RF)	Bagging Ensemble	Robust to noise and overfitting through bootstrap aggregation; well-suited to high-dimensional handcrafted inertial features with correlated subspaces.
XGBoost	Gradient Boosting	Sequential residual correction with regularisation; consistently robust performance on structured tabular data and widely adopted in recent fall detection studies (Subramaniam et al., 2022; Li et al., 2025).
Elastic Net LR (ENet-LR)	Regularised Linear	Combined L1 and L2 regularisation enforces sparsity and implicit feature selection; it provides an interpretable, computationally lightweight linear baseline.
Extra Trees (ET)	Randomised Trees	Extreme randomisation of both feature selection and split thresholds reduces variance beyond standard RF; it is efficient and stable under class imbalance.

As presented in Table 4, the selected classifiers provide complementary predictive capabilities that collectively improve ensemble diversity and generalisation performance. The heterogeneous learning strategy enables the proposed framework to exploit multiple decision perspectives before adaptive refinement and ensemble fusion. This, in turn, reduces dependence on the limitations of any individual classifier (Wan Azman et al., 2024).

Probability Calibration

The heterogeneous learning strategy incorporates RF, XGB, ET, and ENet-LR. These classifiers represent complementary learning paradigms, including bagging, boosting, randomised ensemble learning, and regularised linear modelling (Bhanusree et al., 2023).

Among the available calibration techniques, Platt Scaling is adopted for its computational efficiency and effectiveness with heterogeneous machine learning models. The method transforms raw classifier scores into calibrated posterior probabilities that more accurately reflect the likelihood of fall events, providing consistent probabilistic inputs for the subsequent adaptive learning stages.

Let $h_m(\cdot)$ denote the m^{th} base classifier and $s_m(x_i)$ represent the corresponding raw prediction score for the input feature vector x_i. The calibrated posterior probability is computed using Platt Scaling, presented as Equation 6.

$$p_m(x_i) = \frac{1}{1 + \exp(A_m s_m(x_i) + B_m)}, \tag{6}$$

where (A_m) and (B_m) are calibration parameters estimated from the validation data. If the sigmoid assumption is violated, isotonic regression may be used instead during model optimisation.

The calibrated probabilities generated by all heterogeneous base classifiers are subsequently organised into a unified probability vector, given as Equation 7.

$$p_i = [p_1(x_i), p_2(x_i), \dots, p_M(x_i)]^\top, \tag{7}$$

where M denotes the total number of base classifiers. This calibrated probability vector serves as input to Adaptive Learning Module I (AI Base Refinement), enabling the subsequent refinement stage to analyse classifier agreement, confidence, and uncertainty using a common probabilistic representation.

Adaptive Learning Module I (AI Base Refinement)

It represents the first adaptive reasoning stage within the AHEF. Rather than forwarding calibrated classifier probabilities directly to the stacking meta-learner, this module performs uncertainty-aware probabilistic feature enrichment by analysing the collective behaviour of heterogeneous base learners before ensemble fusion. This process captures complementary information, including prediction reliability, classifier consensus, and uncertainty, which individual probability estimates do not explicitly represent.

The module begins by evaluating the agreement among the heterogeneous base learners through the ensemble mean and confidence dispersion of the calibrated probability vector (refer to Equation 8). These statistical descriptors quantify both overall consensus and variability in the predicted probabilities, providing informative measures of prediction confidence and disagreement.

$$\mu_i = \frac{1}{M}\sum_{m=1}^{M} p_m(x_i), \quad \sigma_i^2 = \frac{1}{M}\sum_{m=1}^{M}(p_m(x_i) - \mu_i)^2, \tag{8}$$

where μ_i = ensemble mean probability, σ_i^2 = ensemble probability variance, $p_m(x_i)$ = calibrated probability from the m^{th}classifier, and M = total number of base classifiers.

Prediction uncertainty is further quantified using Shannon entropy, presented as Equation 9.

$$H_i = -\frac{1}{M}\sum_{m=1}^{M}\left[p_m(x_i)\log p_m(x_i) + \left(1 - p_m(x_i)\right)\log\left(1 - p_m(x_i)\right)\right], \tag{9}$$

where higher entropy values indicate greater disagreement among base classifiers, while lower entropy reflects stronger prediction consensus.

To further improve discriminative capability, the proposed framework augments these uncertainty descriptors with contextual and morphology-based information. Specifically, the calibrated probability vector is enriched using ensemble statistics, minimum and maximum probability values, together with physics-informed descriptors, namely the FMI and SB, to construct the final meta-feature representation.

$$z_i = \left[p_i, \mu_i, \sigma_i^2, H_i, p_i^{min}, p_i^{max}, FMI_i, SB_i\right], \tag{10}$$

where p_i= calibrated probability vector, μ_i = ensemble mean probability, σ_i^2 = ensemble variance, H_i = Shannon entropy, p_i^{min} = minimum classifier probability, and p_i^{max} = maximum classifier probability.

Unlike conventional ensemble methods that rely solely on calibrated classifier probabilities, Adaptive Learning Module I enriches the probabilistic representation by incorporating classifier consensus, uncertainty, contextual information, and fall morphology before ensemble fusion. This enriched representation offers the stacking classifier more informative decision features, improving adaptive fusion and enhancing generalisation across heterogeneous inertial sensing datasets.

Stacking Meta-Learning

The enriched meta-feature vector feeds into the stacking meta-learning stage, where the heterogeneous base-learner outputs are fused into a unified probability estimate. Rather than applying fixed voting or averaging strategies, the stacking classifier learns the relationship between calibrated probabilities and uncertainty-aware meta-features through supervised learning, enabling adaptive ensemble fusion.

To prevent information leakage, the stacking classifier is trained using Out-of-Fold (OOF) predictions generated during subject-independent cross-validation. For each iteration, the base learners are fitted using $K - 1$ folds and validated on the remaining fold, providing unbiased probability estimates for

training the meta-learner while preserving subject independence. The resulting OOF predictions are subsequently combined with the enriched meta-features to construct the input feature vector for the stacking classifier.

A Regularised Logistic Regression model is adopted as the stacking meta-learner due to its robustness, computational efficiency, and well-calibrated probabilistic outputs. Given the enriched meta-feature vector z_i, the hybrid probability is estimated as Equation 11.

$$p_{\text{hybrid},i} = \sigma(w^{\top} z_i + b), \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, w represents the learned coefficient vector, z_i = enriched meta-feature vector, and b is the bias parameter. The resulting hybrid probability integrates complementary information from heterogeneous classifiers together with uncertainty-aware descriptors generated by Adaptive Learning Module I, providing a more reliable estimate than any individual base learner.

Adaptive Learning Module II (AI Meta Refinement)

Although the stacking classifier produces a well-calibrated hybrid probability, isolated false positives may still occur during high-intensity ADLs. To improve prediction reliability, Adaptive Learning Module II (AI Meta Refinement) performs the final decision verification by integrating the hybrid probability with contextual, temporal, and biomechanical information.

Unlike Adaptive Learning Module I, which enriches probabilistic representations before ensemble fusion, this module operates after stacking meta-learning to refine the resulting ensemble output. The hybrid probability is integrated with activity-aware contextual information and morphology-based descriptors through the adaptive refinement function, presented as Equation 12.

$$\mathrm{r_i} = \psi\left(\mathrm{p_{hybrid}}, \mathrm{i}, \mathrm{c_i}, \mathrm{m_i}\right), \tag{12}$$

where $p_{\text{hybrid}}, i,$ denotes the hybrid probability generated by the stacking meta-learner, c_i represents context-aware information derived from activity-normalised motion intensity, m_i denotes morphology-based descriptors including the FMI and SB, and $\psi(\cdot)$ represents the adaptive refinement function.

The activity-aware threshold is estimated during model validation from the probability distributions of fall and non-fall activities. This allows the framework to adapt the confidence threshold according to the activity context. The refined decision representation r_i is then verified through activity-aware thresholding and temporal confirmation.

To reduce isolated false alarms, the framework requires consistent predictions across consecutive motion windows before confirming a fall event, rather than treating a single high-confidence prediction as sufficient evidence of a fall. The temporal confirmation process is defined as Equation 13.

$$\sum_{j=i-k+1}^{i} \mathbb{1}\left(p_{hybrid,j} \geq \tau_{a_i}\right) = k, \tag{13}$$

where I(·) is the indicator function, $p_{hybrid,j}$ = hybrid probability at window j, τ_{a_i} = activity-aware decision threshold, k = minimum number of consecutive windows required to confirm a fall, i = current motion window index, and j = verification window index.

Finally, the refined prediction is verified using the FMI to assess whether the detected event exhibits motion characteristics consistent with a genuine fall. A fall is confirmed only when both the adaptive confidence criteria and temporal-morphological verification are satisfied, thereby reducing false-positive detections arising from vigorous ADLs.

EVALUATION AND RESULTS

Experimental Setup

This section describes the experimental evaluation conducted to assess the proposed AHEF for inertial fall detection. The evaluation uses three publicly available benchmark datasets, namely SisFall, KFall, and MobiFall (Sucerquia et al., 2017; Yu et al., 2021; Vavoulas et al., 2013), under subject-independent cross-validation and cross-dataset validation protocols to evaluate classification accuracy, robustness, and generalisation across heterogeneous sensing environments. Performance is measured using accuracy, sensitivity (recall), specificity, precision, F1-score, and False Positive Rate (FPR) to provide a comprehensive assessment of fall-detection capability and prediction reliability for practical fall-detection applications.

As summarised in Table 5, the SisFall, MobiFall, and KFall datasets encompass heterogeneous sensing platforms, participant populations, and activity protocols, enabling rigorous assessment of both classification performance and cross-dataset generalisation. Within-dataset evaluation measures classification performance under dataset-specific conditions. In contrast, cross-dataset evaluation assesses the framework's robustness by evaluating its performance on data originating from unseen sensing environments.

Table 5

Summary of the Characteristics and Experimental Roles of the Benchmark Datasets

Dataset	Sensor platform	Subjects	Fall windows	Non-Fall/ADL windows	Total	Sampling rate
SisFall	Wearable IMU	38 (23 young adults + 15 older adults)	1,799	2,706	4,505	200 Hz
MobiFall	Smartphone IMU	11 volunteers	476	1,137	1,613	87 Hz
KFall	Wearable IMU	32	2,346	2,728	5,074	100 Hz

Although the benchmark datasets exhibited varying class distributions, particularly MobiFall, which contains 476 Fall windows and 1,137 Non-Fall/ADL windows, the original class distributions were retained throughout the experiments to preserve the natural characteristics of the benchmark datasets and to provide a realistic evaluation scenario. To minimise potential bias arising from unequal class distributions, model performance was not assessed using Accuracy alone. Instead, Sensitivity, Specificity, Precision, F1-score, and False Positive Rate (FPR) were reported in addition to Accuracy,

providing complementary assessments of the model's ability to recognise both Fall and Non-Fall/ADL classes. Furthermore, subject-independent cross-validation and Leave-One-Dataset-Out (LODO) validation were employed to evaluate the robustness and generalisation capability of the proposed framework across independent subjects and heterogeneous datasets. The LODO evaluation further examines whether the learned decision patterns remain effective when the framework is evaluated on a completely unseen dataset, thereby reducing the likelihood that performance is driven by dataset-specific characteristics, including class distribution effects. Despite the moderate imbalance observed in MobiFall, the framework achieved consistently high Sensitivity, Specificity, and low FPR across all evaluation settings, suggesting balanced discrimination between Fall and Non-Fall/ADL classes.

Two evaluation protocols were adopted. Within-dataset evaluation was performed independently on each dataset using subject-independent 10-fold cross-validation. To assess generalisation, a LODO protocol was employed, in which the model was trained on two datasets and tested on the remaining unseen dataset. This provides an experimental approximation of deployment across different sensing environments and participant populations.

For all experiments, inertial signals were segmented into fixed-length windows with 50% overlap before undergoing the signal pre-processing and feature engineering procedures described in Section 3. The AHEF processed the resulting feature vectors, comprising multi-model base learning, probability calibration, Adaptive Learning Module I (AI Base Refinement), stacking meta-learning, and Adaptive Learning Module II (AI Meta Refinement). Platt Scaling was applied before ensemble fusion to calibrate the probability outputs of the base learners, while OOF predictions were used during stacking to prevent information leakage and provide unbiased inputs for the meta-learner.

Performance was evaluated using six metrics: accuracy, sensitivity (recall), specificity, precision, F1-score, and FPR. Together, these metrics assess overall classification performance, fall-detection capability, and false-alarm control for wearable fall-detection applications.

Within-Dataset Performance

Within-dataset evaluation assesses AHEF performance using subject-independent cross-validation on the SisFall, MobiFall, and KFall datasets. These results establish the baseline before evaluating cross-dataset generalisation. Table 6 reports the quantitative results, while Figure 4 provides a comparative visualisation across the three benchmark datasets.

Table 6

Within-Dataset Evaluation Results of the AHEF

Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	FPR (%)
SisFall	98.5	99.0	98.0	98.8	2.0
KFall	97.8	98.2	97.4	98.0	2.6
MobiFall	98.9	99.3	98.5	99.1	1.5

Figure 4

Comparison of Within-Dataset Performance across the SisFall, MobiFall, and KFall Benchmark Datasets

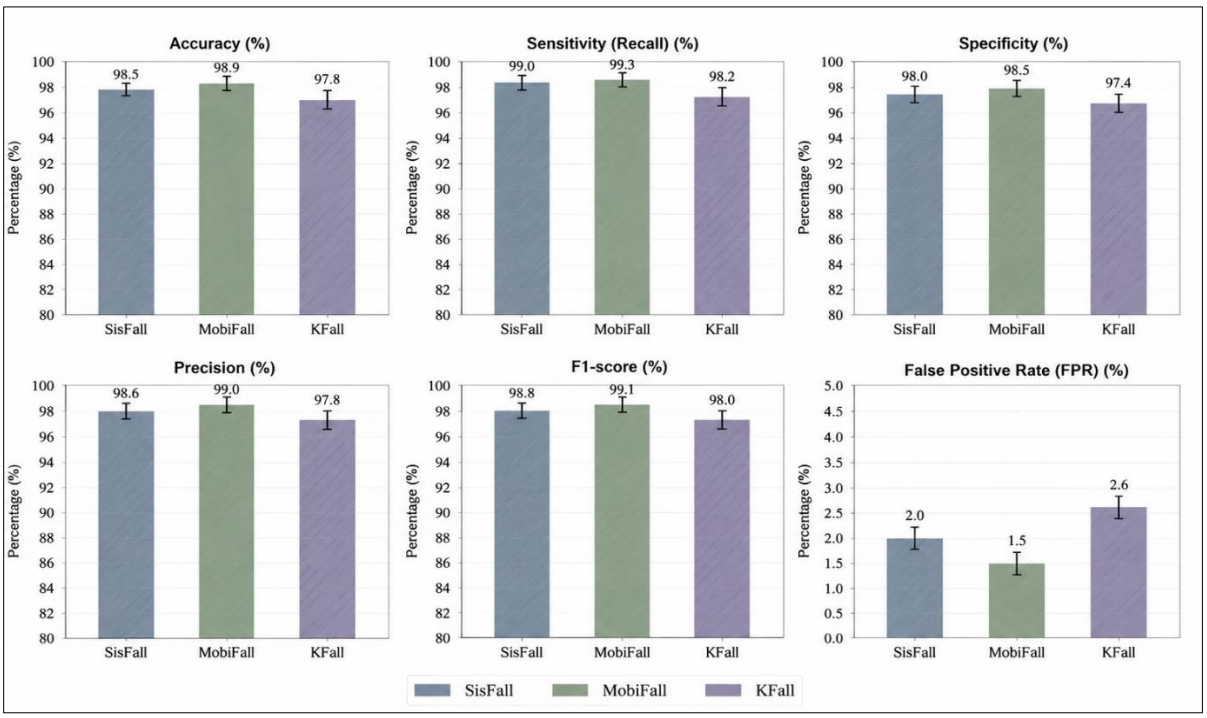


Table 6 and Figure 4 illustrate that AHEF consistently achieves high classification performance across the SisFall, MobiFall, and KFall datasets despite differences in sensing platforms, participant demographics, and activity protocols. The consistently high accuracy, sensitivity, specificity, and F1-score indicate that the hierarchical learning architecture effectively distinguishes fall events from ADLs under diverse sensing conditions. These findings establish a strong baseline for the subsequent cross-dataset generalisation study.

Cross-Dataset Generalisation

Cross-dataset evaluation assesses how well the AHEF generalises to unseen inertial sensing environments. Unlike within-dataset evaluation, where training and testing data come from the same dataset, this protocol evaluates the model under distributional shifts caused by differences in sensing devices, participant populations, activity protocols, and data acquisition conditions. As a result, it provides a more practical measure of deployment robustness when the model encounters previously unseen motion patterns.

Generalisation was evaluated using a LODO protocol. For each experiment, two benchmark datasets were used for model training, while the remaining dataset served as an independent test set. The procedure was repeated until every dataset had been used once for testing, enabling evaluation under heterogeneous sensing environments while avoiding information leakage between the training and test data.

Table 7

Cross-Dataset Evaluation of the AHEF Using LODO Validation

Training	Validate	Accuracy (%)	Sensitivity (%)	FPR (%)
SisFall + MobiFall	KFall	97.5	98.1	2.7
SisFall + KFall	MobiFall	98.3	99.0	1.8
MobiFall + KFall	SisFall	98.1	98.7	2.1

Table 7 indicates that AHEF maintains consistently high performance across all LODO scenarios, indicating robust generalisation despite variations in sensing devices, participant populations, and activity protocols. To further investigate the proposed framework's classification behaviour, Figures 5-7 present the confusion matrices for each LODO evaluation scenario.

Figure 5

Confusion Matrix for LODO: Train (SisFall + MobiFall), Validate (KFall)

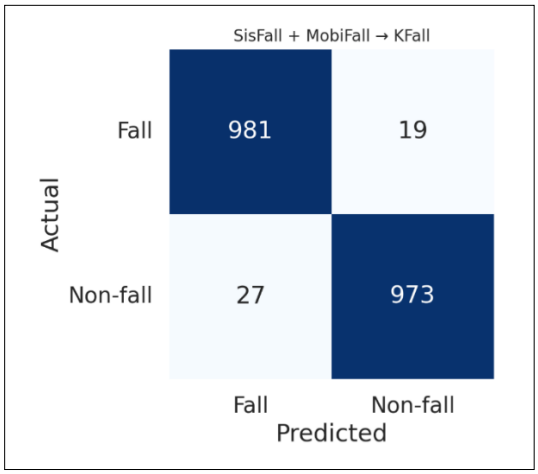


Figure 6

Confusion Matrix for LODO: Train (SisFall + KFall), Validate (MobiFall)

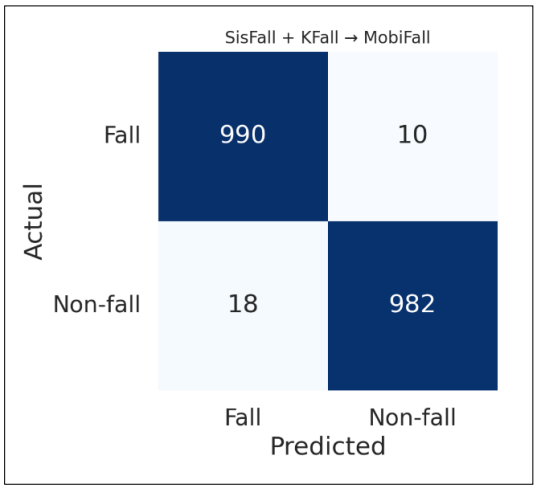
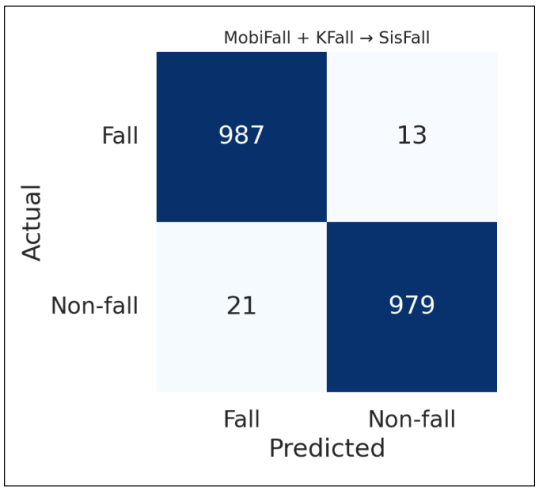


Figure 7

Confusion Matrix for LODO: Train (MobiFall + KFall), Validate (SisFall)



The confusion matrices presented in Figures 5-7 provide a detailed assessment of classification performance across the evaluated sensing environments. Across all evaluation scenarios, the AHEF achieves a high proportion of correctly classified fall and non-fall events, with relatively few false-positive and false-negative predictions. The consistent distribution of true positives and true negatives indicates stable decision boundaries despite differences in sensing hardware, participant characteristics, and activity protocols.

Only modest performance variation is observed across the three evaluation scenarios, suggesting that the framework maintains classification performance under dataset shift. The low number of misclassifications indicates the proposed architecture's ability to preserve both detection sensitivity and specificity across heterogeneous inertial sensing environments.

Comparison with Base Learners

To further assess the performance of the AHEF, its performance was compared with the four heterogeneous base learners incorporated in the framework: RF, XGB, ET, and ENet-LR. This comparison quantifies the performance gains achieved through probability calibration, Adaptive Learning Module I (AI Base Refinement), stacking meta-learning, and Adaptive Learning Module II (AI Meta Refinement).

Table 8

Performance Comparison of Base Classifiers and the Adaptive Hybrid Model

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	FPR (%)
Random Forest (RF)	96.1	96.8	95.2	96.3	4.8
XGBoost	96.8	97.5	96.0	97.0	4.0
ENet-LR	94.5	95.2	93.6	94.8	6.4
Extra Trees (ET)	95.9	96.4	95.1	96.1	4.9
Adaptive Hybrid Model	98.7	99.1	98.3	98.9	1.7

Table 8 demonstrates that the AHEF consistently outperforms all heterogeneous base learners across every evaluation metric. These results indicate that integrating complementary learning paradigms within a hierarchical ensemble architecture systematically improves performance over individual classifiers.

Among the baseline classifiers, XGB achieved the best overall performance, with 96.8% accuracy, 97.5% sensitivity, and a 97.0% F1-score, making it the strongest standalone model. RF and ET achieved similar results, suggesting that tree-based ensemble methods can capture complex nonlinear motion patterns. In comparison, ENet-LR recorded the lowest performance, with 94.5% accuracy, a 94.8% F1-score, and the highest FPR of 6.4%, implying that a regularised linear classifier is less effective in modelling the complex characteristics of inertial fall data.

AHEF delivered the best overall results, achieving 98.7% accuracy, 99.1% sensitivity, 98.3% specificity, a 98.9% F1-score, and the lowest FPR of 1.7%. Compared with XGB, the strongest individual classifier, the proposed framework improved accuracy by 1.9 percentage points while reducing the FPR by 2.3 percentage points. The reduction in false alarms, together with the high sensitivity achieved by AHEF, indicates its potential for practical fall detection applications.

The observed performance gains demonstrate that progressively refining classifier outputs through probability calibration, adaptive ensemble fusion, and decision refinement produces more reliable probability estimates than standalone classifiers. Collectively, these results demonstrate the effectiveness of the hierarchical ensemble architecture in improving predictive accuracy and decision reliability for inertial fall detection.

Figure 8

Comparison of Classification Accuracy and FPR for AHEF and the Individual Base Learners

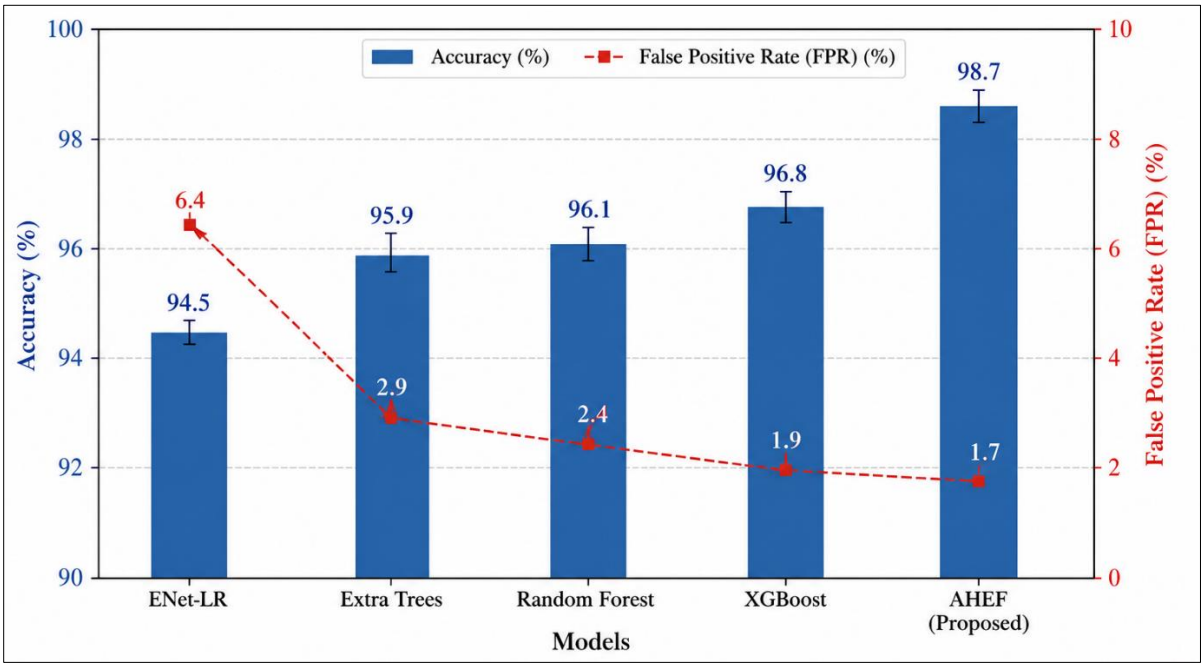


Figure 8 compares AHEF performance with that of the individual heterogeneous base learners in terms of accuracy and FPR. AHEF achieves the highest classification accuracy while maintaining the lowest FPR, suggesting that improved predictive performance comes with a substantial reduction in false alarms.

Compared with the individual base learners, the framework provides a better balance between fall-detection sensitivity and false-alarm suppression, producing more reliable predictions under diverse activity conditions. This balance is particularly crucial for practical fall detection systems, where excessive false alarms may reduce user confidence and long-term system acceptance.

Performance Stability

Performance stability was assessed using the mean and standard deviation from subject-independent 10-fold cross-validation. Table 9 summarises the average performance and corresponding standard deviations of AHEF and the individual base learners across all evaluation metrics.

Table 9

Mean Performance and Standard Deviation across Subject-Independent Cross-Validation Folds

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)	FPR (%)
Random Forest	96.1 ± 0.8	96.8 ± 0.7	95.2 ± 0.8	96.3 ± 0.7	4.8 ± 0.6
XGBoost	96.8 ± 0.6	97.5 ± 0.5	96.0 ± 0.6	97.0 ± 0.5	4.0 ± 0.5
ENet-LR	94.5 ± 1.0	95.2 ± 0.9	93.6 ± 1.0	94.8 ± 0.9	6.4 ± 0.8
Extra Trees	95.9 ± 0.7	96.4 ± 0.8	95.1 ± 0.7	96.1 ± 0.7	4.9 ± 0.6
Adaptive Hybrid	98.7 ± 0.4	99.1 ± 0.3	98.3 ± 0.4	98.9 ± 0.3	1.7 ± 0.3

Table 9 indicates that the AHEF consistently recorded the highest average performance while exhibiting the lowest standard deviations across all evaluation metrics. The limited variability indicates that the model is robust to changes in training data partitions and produces consistent predictions across different validation folds. Compared with the individual base learners, AHEF indicates greater stability in accuracy, sensitivity, specificity, F1-score, and FPR, demonstrating improved prediction consistency under subject-independent cross-validation.

The consistently low FPR variation (1.7 ± 0.3%) indicates that the framework reliably suppresses false alarms without compromising fall detection sensitivity. These findings demonstrate that the proposed framework delivers strong predictive performance and consistent behaviour across validation folds, supporting its applicability to practical wearable fall detection systems.

Ablation Study

An ablation analysis was conducted to quantify the contribution of the key architectural components of the AHEF. Three configurations were assessed under identical subject-independent 10-fold cross-validation using the SisFall, MobiFall, and KFall datasets: (i) Base Ensemble, (ii) Stacking Only, and (iii) the complete AHEF.

The Base Ensemble applies simple probability averaging; Stacking only introduces stacking meta-learning with uncertainty-aware meta-features; and the complete AHEF additionally incorporates probability calibration, Adaptive Learning Module I (AI Base Refinement), and Adaptive Learning Module II (AI Meta Refinement), with activity-aware thresholding and temporal confirmation (k = 3). Table 10 presents the corresponding results.

Table 10

Summary of the Performance of the Three Configurations

Dataset	Configuration	Acc (%)	Sens (%)	Spec (%)	F1 (%)	FPR (%)
SisFall	Base Ensemble	96.2	96.8	95.6	96.2	4.4
	Stacking Only	97.5	97.9	97.1	97.5	2.9
	Adaptive Hybrid Model	98.5	99.0	98.0	98.8	2.0
MobiFall	Base Ensemble	96.8	97.2	96.4	96.8	3.6
	Stacking Only	98.0	98.5	97.5	98.0	2.5
	Adaptive Hybrid Model	98.9	99.3	98.5	99.1	1.5
KFall	Base Ensemble	94.2	94.8	93.6	94.2	6.4
	Stacking Only	96.0	96.5	95.5	96.0	4.5
	Adaptive Hybrid Model	97.8	98.2	97.4	98.0	2.6

As Table 10 indicates, performance improves consistently as additional components are incorporated into the framework. Although the Base Ensemble provides competitive classification results, it produces relatively higher FPRs, particularly on the KFall dataset, where complex transition activities and pre-impact movements increase classification difficulty. Introducing stacking meta-learning consistently improves accuracy and F1-score while reducing FPR, indicating that adaptive fusion of heterogeneous classifier outputs is more effective than simple probability averaging.

Across all three datasets, the AHEF consistently delivers the highest performance, with the greatest improvement observed in FPR reduction. Compared with the Stacking Only configuration, incorporating Adaptive Learning Module I (AI Base Refinement) and Adaptive Learning Module II (AI Meta Refinement) reduces FPR by 0.9–1.9 percentage points, with the largest reduction on KFall. Relative to the Base Ensemble, the FPR decreases by approximately 55–64%, from 4.4% to 2.0% on SisFall, 3.6% to 1.5% on MobiFall, and 6.4% to 2.6% on KFall. These results suggest that adaptive probabilistic refinement, together with decision verification, effectively reduces false alarms while maintaining high fall-detection sensitivity.

Figure 9 visually compares the three configurations across the evaluation metrics. The progressive improvement from the Base Ensemble to Stacking Only and finally AHEF confirms that each architectural component contributes positively to the framework's overall performance.

Figure 9

Performance Progression across Ablation Configurations for the AHEF

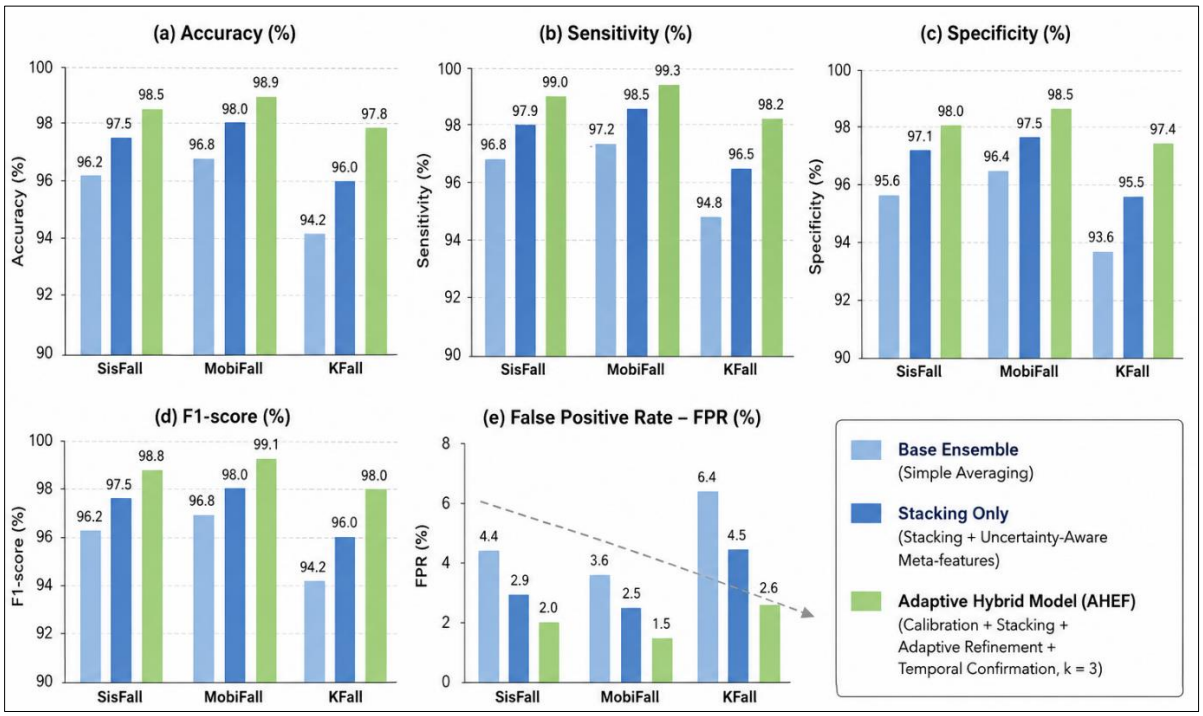


Figure 9 illustrates the incremental contribution of each architectural component. Incorporating stacking meta-learning consistently improves accuracy, sensitivity, specificity, and F1-score while reducing the FPR across all three datasets. The complete AHEF achieves the best overall performance, with the largest FPR reduction on the KFall dataset. This consistent progression confirms the effectiveness of the hierarchical adaptive learning strategy.

These improvements are maintained despite differences in sensing devices, sensor placement, sampling frequencies, participant characteristics, and fall protocols. Although the KFall dataset yields slightly lower performance than SisFall and MobiFall, this is expected as it contains more complex pre-impact transitions and motion patterns that closely resemble ADLs. Consequently, this makes fall discrimination more challenging.

Statistical analysis further supports the observed improvements. Paired *t*-tests across the subject-independent cross-validation folds indicate that AHEF significantly outperforms both the Base Ensemble ($p < 0.01$) and the Stacking Only configuration ($p < 0.05$). This confirms that the performance gains are unlikely to be attributable to random variation.

Overall, the ablation analysis indicates that AHEF's superior performance stems from the complementary interaction of its architectural components. Stacking meta-learning enhances the adaptive fusion of heterogeneous classifier outputs. In contrast, the two adaptive refinement modules further refine the resulting predictions and suppress false alarms. Collectively, these findings provide strong empirical support for the proposed hierarchical adaptive learning architecture for robust cross-dataset inertial fall detection.

DISCUSSION

The evaluation results indicate that AHEF consistently improves fall detection performance in both within-dataset and cross-dataset evaluations. Compared with the individual base learners, AHEF achieves higher classification accuracy while substantially reducing FPRs, demonstrating the effectiveness of the proposed hierarchical adaptive learning architecture.

The cross-dataset experiments further demonstrate strong generalisation despite variations in sensing devices, participant populations, sensor placement, and activity protocols. The limited performance degradation under dataset shift indicates that the learned feature representations remain robust across heterogeneous sensing environments. This, in turn, supports the framework's suitability for practical wearable fall detection.

The effect of class distribution was also considered when interpreting the classification results. Although MobiFall contains fewer Fall windows (476) than Non-Fall/ADL windows (1,137), the reported results do not indicate a substantial bias toward the majority class. In the within-dataset evaluation, the proposed AHEF achieved 99.1% Sensitivity and 98.3% Specificity, demonstrating strong recognition of both Fall and Non-Fall/ADL instances rather than favouring the majority class. Furthermore, the Leave-One-Dataset-Out (LODO) evaluation-maintained Sensitivity above 98% across all cross-dataset evaluation scenarios, indicating that the framework consistently detected Fall events even when evaluated on unseen datasets. These findings suggest that the proposed framework learned generalisable fall-discrimination patterns rather than dataset-specific class characteristics. Nevertheless, the present study retained the original class distributions to preserve the characteristics of the benchmark datasets and did not explicitly evaluate class-rebalancing techniques such as cost-sensitive learning or synthetic oversampling. Consequently, while the observed results suggest that class imbalance did not materially affect the framework's predictive performance, its potential influence remains an important consideration and should be further investigated in future studies.

Although recent deep learning models have reported competitive performance by automatically learning representations from inertial signals, they generally require larger labelled datasets and greater computational resources (Su et al., 2024; Li et al., 2025). Moreover, many studies have primarily evaluated their models on a single dataset, providing limited evidence of robustness under heterogeneous sensing conditions. In contrast, AHEF emphasises cross-dataset generalisation, probability calibration, and false-alarm reduction while maintaining computational efficiency for wearable and edge-based deployment.

The ablation analysis further indicates that each architectural component contributes to AHEF's overall effectiveness. Stacking meta-learning enhances predictive performance through adaptive fusion, while the two adaptive refinement modules further reduce false alarms without compromising detection sensitivity. Together, these results provide empirical support for the proposed hierarchical architecture's contribution to robust cross-dataset inertial fall detection.

CONCLUSION

This study proposed a cross-dataset AHEF for wearable inertial fall detection. By integrating heterogeneous ensemble learning with hierarchical adaptive refinement, AHEF improves prediction reliability while maintaining strong generalisation across heterogeneous sensing environments.

Experimental evaluation using the SisFall, MobiFall, and KFall benchmark datasets indicated that AHEF generally outperformed the individual base learners and conventional ensemble configurations. Under within-dataset evaluation, AHEF achieved 98.7% accuracy, 99.1% sensitivity, 98.3% specificity, and a 98.9% F1-score, while maintaining the lowest FPR among all evaluated models. Cross-dataset evaluation further confirmed its robustness under dataset shift, and the ablation analysis demonstrated that each architectural component contributed to the overall performance improvement.

A limitation of the present study is the presence of moderate class imbalance within some of the benchmark datasets, particularly MobiFall, which exhibits the largest difference between Fall and Non-Fall/ADL windows. The experiments retained the original benchmark distributions and did not apply explicit class-rebalancing techniques such as class weighting, oversampling, or undersampling. Although the use of class-sensitive evaluation metrics and cross-dataset validation provides evidence of robust Fall and Non-Fall/ADL recognition, the independent effect of class-rebalancing strategies was not evaluated in the present study. Future work should therefore investigate cost-sensitive learning, class-weighted optimisation, and appropriate minority-class oversampling techniques, together with comparative experiments, to quantify their influence on cross-dataset fall detection performance.

Overall, the results indicate that AHEF provides a reliable, generalisable solution for wearable inertial fall detection. By combining high detection sensitivity with a low FPR, the proposed approach indicates strong potential for practical healthcare monitoring applications.

Future work may focus on validating the proposed approach using continuous real-world monitoring data, assessing its deployment on resource-constrained wearable and edge-computing platforms, and extending the adaptive learning mechanism through online and continual learning to support long-term personalisation and adaptation to changes in user behaviour over time.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors thank the editor and the anonymous referees for their valuable and constructive comments, which greatly improved the paper.

REFERENCES

Bhanusree, Y., Kumar, S. S., & Rao, A. K. (2023). Time-distributed attention-layered convolution neural network with ensemble learning using random forest classifier for speech emotion recognition. *Journal of Information and Communication Technology, 22*(1), 49–76. https://doi.org/10.32890/jict2023.22.1.3

BIng2325. (2023, July 12). *SisFall dataset* [Data set]. GitHub. https://github.com/BIng2325/SisFall/releases

BMI-HMU. (2026). *The MobiFall and MobiAct datasets*. Biomedical Informatics & eHealth Laboratory, Hellenic Mediterranean University. https://bmi.hmu.gr/the-mobifall-and-mobiact-datasets-2/

Centers for Disease Control and Prevention. (2026a, February 26). *Older adult falls data*. https://www.cdc.gov/falls/data-research/index.html

Centers for Disease Control and Prevention. (2026b, January 27). *Facts about falls*. https://www.cdc.gov/falls/data-research/facts-stats/index.html

Gießler, M., Werth, J., Waltersberger, B., & Karamanidis, K. (2024). A framework to automatically detect near-falls using a wearable inertial measurement cluster. *Communications Engineering, 3*, Article 181. https://doi.org/10.1038/s44172-024-00325-x

Hong, Z., Li, Z., Zhong, S., Lyu, W., Wang, H., Ding, Y., He, T., & Zhang, D. (2024). CrossHAR: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8*(2), Article 64. https://doi.org/10.1145/3659597

Kasim, S., & Sheppard, J. W. (2023, June). Cross-domain similarity in domain adaptation for human activity recognition. In *2023 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN54540.2023.10191305

Li, Y., Liu, P., Fang, Y., Wu, X., Xie, Y., Xu, Z., Ren, H., & Jing, F. (2025). A decade of progress in wearable sensors for fall detection (2015–2024): A network-based visualization review. *Sensors, 25*(7), 2205. https://doi.org/10.3390/s25072205

Palli, A. S., Jaafar, J., Gilal, A. R., Alsughayyir, A., Gomes, H. M., Alshanqiti, A., & Omar, M. (2024). Online machine learning from non-stationary data streams in the presence of concept drift and class imbalance: A systematic review. *Journal of Information and Communication Technology, 23*(1), 105–139. https://doi.org/10.32890/jict2024.23.1.5

Pavlovic, M. (2025). *Understanding model calibration: A gentle introduction and visual exploration of the calibration and the expected calibration error (ECE)* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.19047

Priya, S., & Amshakala, K. (2025). An adaptive fall detection system based on ensemble learning using variants of YOLOv8, RetinaNet and DETR. *Scientific Reports, 15*, Article 33161. https://doi.org/10.1038/s41598-025-97634-8

Qin, X., Wang, J., Chen, Y., Lu, W., & Jiang, X. (2023). Domain generalization for activity recognition via adaptive feature fusion. *ACM Transactions on Intelligent Systems and Technology, 14*(1), 9:1–9:21. https://doi.org/10.1145/3552434

Rafrastara, F. A., Ghozi, W., Sani, R. R., Handoko, L. B., Abdussalam, A., Pramudya, E. R., & Abdollah, F. M. (2025). Integrating information gain and chi-square for enhanced malware detection performance. *Journal of Information and Communication Technology, 24*(1), 79–101. https://doi.org/10.32890/jict2025.24.1.4

Sainin, M. S., Alfred, R., & Ahmad, F. (2021). Ensemble meta classifier with sampling and feature selection for data with imbalance multiclass problem. *Journal of Information and Communication Technology, 20*(2), 103–133. https://doi.org/10.32890/jict2021.20.2.1

Scikit-learn developers. (2026). *Probability calibration* [User guide]. https://scikit-learn.org/stable/modules/calibration.html

Silva Filho, T. M., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning, 112*, 3211–3260. https://doi.org/10.1007/s10994-023-06336-7

Silva, C. A., Casilari, E., & García-Bermúdez, R. (2024). Cross-dataset evaluation of wearable fall detection systems using data from real falls and long-term monitoring of daily life. *Measurement, 235*, 114992. https://doi.org/10.1016/j.measurement.2024.114992

Su, C., Wei, J., Lin, D., Kong, L., & Guan, Y. (2024). A novel model for fall detection and action recognition combined lightweight 3D-CNN and convolutional LSTM networks. *Pattern Analysis and Applications, 27*(1), Article 3. https://doi.org/10.1007/s10044-024-01224-9

Subramaniam, S. S., Faisal, A. A. I., & Deen, M. J. (2022). Wearable sensor systems for fall risk assessment: A review. *Frontiers in Digital Health, 4*, 921506. https://doi.org/10.3389/fdgth.2022.921506

Sucerquia, A., López, J. D., & Vargas-Bonilla, J. F. (2017). SisFall: A fall and movement dataset. *Sensors, 17*(1), 198. https://doi.org/10.3390/s17010198

Vavoulas, G., Pediaditis, M., Spanakis, E. G., & Tsiknakis, M. (2013). The MobiFall dataset: An initial evaluation of fall detection algorithms using smartphones. In *2013 13th IEEE International Conference on BioInformatics and BioEngineering (BIBE)* (pp. 1–4). IEEE. https://doi.org/10.1109/BIBE.2013.6701629

Wan Azman, W. N. F. S., Ku Azir, K. N. F., & Mohd Khairuddin, A. (2024). An embedded machine learning-based spoiled leftover food detection device for multiclass classification. *Journal of Information and Communication Technology, 23*(2), 253–292. https://doi.org/10.32890/jict2024.23.2.4

Yu, X., Jang, J., & Xiong, S. (2021). A large-scale open motion dataset (KFall) and benchmark algorithms for detecting pre-impact fall of the elderly using wearable inertial sensors. *Frontiers in Aging Neuroscience, 13*, Article 692865. https://doi.org/10.3389/fnagi.2021.692865

Yusof, Y., & Fajila, F. (2026). Conditional tabular generative adversarial network-based synthetic data generation for model generalisation improvement. *Journal of Information and Communication Technology, 25*(1), 1–16. https://doi.org/10.32890/jict2026.25.1.1