electronics

MDPI

Article

Non-Invasive Showering Estimation Utilizing Household-Adaptive Models and Washing Time Data

Takuya Nakata [1,*], Jiro Hashizume [2], Akihiro Yanada [2] and Masahide Nakamura [1,3]

[1] Center for Mathematical and Data Sciences, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan; masa-n@cmds.kobe-u.ac.jp
[2] NORITZ Corporation, 93 Edomachi, Chuo-ku, Kobe 650-0033, Japan
[3] RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
* Correspondence: tnakata@bear.kobe-u.ac.jp; Tel.: +81-78-803-6295

Abstract

This study introduces a **dual-proxy framework** for household-adaptive, non-invasive shower detection using standard water-heater logs. The framework leverages proxy at two complementary levels: a *feature-level proxy* (`washing_seconds`) that captures washing duration, and a *scheme-level proxy* (proxy-driven training) that enables learning in periods without direct shower labels. The proxy feature (`washing_seconds`) serves as an indirect descriptor of washing behavior, enabling effective inference even under label scarcity. We investigated three research questions: (RQ1) the effectiveness of proxy features in improving shower detection, (RQ2) how proxy-driven evaluation identifies compact yet reliable feature subsets, and (RQ3) the robustness of these subsets in long-term, real-world scenarios. Experiments on two households showed that `washing_seconds` consistently improved discrimination (raising summer PR-AUC, lowering non-summer false alarms), and that compact subsets of only two or three features, anchored by the proxy feature, achieved stable performance across households. The evaluation represents an illustrative example based on two cooperating households, providing practical evidence of the framework's real-world applicability. Evaluation in real-world conditions confirmed robustness: representative subsets maintained **micro PR-AUC 0.724–0.728**, **micro F1 0.66–0.69 (macro F1 0.55–0.58)**, and summer PR-AUC near 0.87, with generalization gaps within ± 0.01 for discrimination and small positive shifts for F1 ($+0.02$–$+0.05$). These results demonstrate that proxy can function both as a feature and as a methodological principle, and that the proposed framework is model-agnostic and transferable to other learning architectures. It provides a foundation for adaptive, privacy-preserving smart home applications that can scale to broader household and healthcare contexts.

Keywords: dual-proxy framework; shower detection; non-invasive sensing; household-adaptive modeling; proxy feature; proxy-driven scheme; feature selection; Pareto analysis; calibration; smart home

Academic Editor: Simeone Marino

Received: 17 October 2025
Revised: 28 October 2025
Accepted: 4 November 2025
Published: 5 November 2025

Citation: Nakata, T.; Hashizume, J.; Yanada, A.; Nakamura, M. Non-Invasive Showering Estimation Utilizing Household-Adaptive Models and Washing Time Data. *Electronics* **2025**, *14*, 4336. https://doi.org/10.3390/electronics14214336

1. Introduction

Human-centered smart living environments increasingly rely on non-invasive sensing to recognize daily activities [1–4]. Among these activities, shower detection represents a critical but challenging case. It is relevant to hygiene monitoring, energy management, and elderly care [5–7]. Accurate detection can inform personalized health support, optimize hot-water energy usage, and prevent hygiene-related risks in vulnerable populations.

However, existing methods often rely on multiple dedicated sensors, which increase installation cost, hinder scalability, and raise privacy concerns [8,9]. Moreover, models trained under controlled conditions may fail to generalize across households and seasons, leading to unstable performance and high false alarm rates [10–12]. In this study, we focus on typical domestic bathrooms equipped with both a bathtub and a hand shower—a configuration common in Japan and also found elsewhere. Within this context, shower-only use (washing without soaking) is distinguished from bathtub soaking and other non-bathing stays.

In such daily bathing routines, we define the notion of washing time (W_i), representing the estimated duration of washing activities such as body or hair washing. This quantity serves as a key proxy descriptor, indirectly capturing shower-related behaviors from standard water-heater logs without additional sensors. It provides a practical basis for scalable, privacy-preserving behavioral recognition.

A central difficulty is the scarcity of reliable shower labels. While bathtub usage or non-bathing stays can be inferred more directly from temperature and flow signals, showering events are difficult to annotate consistently, especially in long-term, in-the-wild deployments. This motivates the use of **proxy** information: features or schemes that indirectly capture the target behavior. We argue that proxy can function at two complementary levels. At the *feature level*, we define `washing_seconds`, which approximates washing duration by subtracting tub immersion time from the total stay. At the *scheme level*, we design a proxy-driven training strategy, in which models are trained in periods lacking direct shower labels (e.g., winter) and rely on proxy-derived signals to support learning under unlabeled conditions rather than human supervision.

We propose a dual-proxy framework: proxy features guide the model, while a proxy-driven scheme enables learning without direct shower labels. This dual perspective extends the role of proxy beyond variable design, positioning it as a methodological principle for robust behavioral sensing under label scarcity.

Building on this framework, we investigate three research questions:

RQ1. How effective is the proxy feature `washing_seconds` in improving shower detection?

RQ2. How can a proxy-driven scheme, without direct shower labels, identify compact and reliable feature subsets?

RQ3. How robustly do proxy features and proxy-driven schemes generalize across households and seasons in shower detection?

By addressing these questions, this study contributes both methodological and practical insights. Methodologically, it demonstrates that proxy can function at both the feature and scheme levels, forming a dual-proxy framework for activity recognition under label-scarce or unlabeled conditions. Practically, it shows that household-adaptive models can maintain reliable performance with only two to three features, supporting scalable, cost-efficient, and privacy-preserving smart home deployment.

2. Related Work

Research on shower detection spans diverse sensing modalities. Dedicated devices (e.g., Aguardio) rely on vibration, acoustic, temperature, and humidity sensors [13], while smart metering [14], acoustic classifiers [15], wearables [16], and environmental signals such as indoor CO_2 [17] have also been explored. Additional approaches include motion and water-level sensors for tub monitoring [18,19]. These studies demonstrate feasibility but raise concerns of cost, scalability, and privacy due to the need for new or wearable sensors.

A second line of work exploits existing infrastructure. HydroSense showed that single-point pressure sensing can disaggregate household water events [20], and plumbing

vibrations have also been leveraged [21]. More recently, acoustic [15], vision [17], Wi-Fi CSI [22], and smart water metering [23] have been applied for unobtrusive monitoring. These studies highlight the promise of infrastructure-mediated sensing, but distinguishing showers from baths remains difficult.

Beyond individual appliances, prior work in household behavior recognition and human activity recognition (HAR) has emphasized the long-term reliability of multi-sensor and context-aware systems. For instance, Hiremath and Plötz highlighted that the lifespan of smart-home HAR models often degrades over time due to user and environmental variability [24]. Such systems frequently rely on sensor fusion or transfer learning to restore performance, but these solutions require additional instrumentation and retraining, limiting scalability and privacy in ordinary households. Our study instead focuses on infrastructure-native modeling—recognizing bathing behaviors directly from existing heater logs—bridging infrastructure-mediated sensing with general activity-recognition approaches.

Related perspectives arise from non-intrusive load monitoring (NILM), where energy disaggregation is used to infer activities. Reviews discuss the challenges of trustworthy NILM pipelines [25,26] and applications in energy accountability and sustainability [27]. Such work underscores both the potential and limitations of infrastructure-based recognition.

Proxy or surrogate features provide a complementary strategy when direct labels are scarce. Athey et al. [28] and Tripuraneni et al. [29] formalized surrogate metrics, while Pinnow et al. [30] reviewed validation across domains. Beyond conventional machine learning, proxy- or surrogate-based optimization has also been explored in privacy-sensitive and label-scarce domains such as healthcare, where blockchain-enabled frameworks and the Mother Optimization Algorithm (MOA) demonstrate that surrogate indicators can enhance decision reliability under limited data [31–33].

Multi-objective feature selection has been widely studied to balance accuracy, stability, and interpretability [34,35]. Representative approaches include interactive evolutionary methods [36], Pareto-based algorithms [37], and joint optimization [38]. Other works explored sparsity constraints [39], metaheuristics [40], and systematic reviews [41–43]. These methods provide the methodological basis for our multi-criteria evaluation of proxy-driven subsets.

Finally, robustness and calibration are critical for deployment. Dawadi et al. [44] and Hiremath et al. [24] emphasized longitudinal validation, while surveys have documented calibration metrics and strategies [45,46]. Guo et al. [47] further showed that modern networks are often miscalibrated. These findings motivate our focus on calibration robustness when evaluating dual-proxy subsets.

Collectively, prior research highlights that proxy and surrogate mechanisms support learning under data and privacy constraints, motivating the present dual-proxy framework that unifies feature-level and scheme-level proxies for robust behavioral sensing.

In summary, prior research has demonstrated the promise of sensor-based approaches, infrastructure-mediated sensing, NILM disaggregation, proxy features, multi-objective feature selection, and long-term validation. Yet few studies have integrated proxy features and proxy-driven schemes into a unified framework while explicitly validating robustness under real-world conditions. Our work addresses this gap by evaluating a dual-proxy framework that combines feature-level proxies, scheme-level subset selection, and long-term robustness in household deployments.

3. Problem Definition and Challenges

3.1. Problem Formulation

We consider household bathrooms equipped with both a bathtub and a hand shower, a configuration common in Japan and also found elsewhere. Within this context, residents alternate between bathtub soaking and shower-only washing, and we aim to classify each bathroom stay according to these behaviors. This problem setting represents a typical but constrained domestic configuration (two cooperating households in our dataset), providing a basis for examining generalizable modeling strategies rather than population-level statistics.

We formulate the task of showering estimation as a supervised classification problem based on standard water-heater logs. Each bathroom stay i is categorized into one of three mutually exclusive classes: tub bathing (C_1), showering without tub immersion (C_2), and non-bathing stays such as cleaning or laundry (C_3). While tub bathing (C_1) can be directly detected by water-level sensors, the central difficulty lies in distinguishing showering (C_2) from non-bathing (C_3), since both involve bathroom occupancy without tub immersion. The relationship among these classes is illustrated in Figure 1.

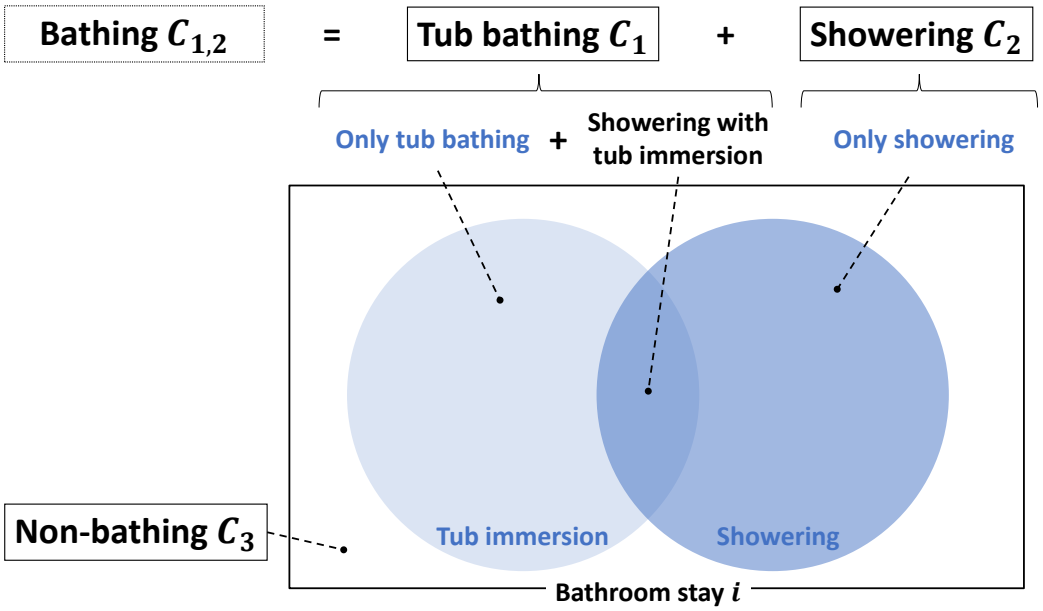


Figure 1. Class definition in the showering estimation problem.

To capture the behavioral distinction, we introduce the notion of *washing time*, denoted as W_i. Intuitively, this quantity represents the duration of washing activities such as body or hair washing, which are typical in bathing behaviors (C_1, C_2) but largely absent in non-bathing stays (C_3). If a bathroom stay included both showering and tub immersion, it was categorized as C_1 (tub bathing), because soaking dominates in energy and water use. Thus, "shower-only" strictly refers to washing without tub immersion. In practice, W_i is not directly measured but rather approximated from available water-heater logs (implemented as the feature `washing_seconds`), serving as a proxy descriptor of washing duration. While other aggregate features such as `water_hour` or `gas_hour` are computed at the household level and may include concurrent non-bathroom usage (e.g., kitchen or laundry), the proxy W_i is grounded in bathroom stay segmentation and tub immersion information, making it less sensitive to such confounding activities. This proxy-based definition is essential because direct annotation of showering events is rarely scalable; it allows the model to leverage measurable quantities derived from standard heater logs

without relying on manual labeling. The detailed formulation and general properties of W_i are described later in Section 4.1.

3.2. Challenges and Research Questions

Despite the apparent simplicity of the problem formulation, several challenges hinder reliable shower detection:

- **Shower vs. non-bathing ambiguity.** Both showering (C_2) and non-bathing stays (C_3) involve bathroom occupancy without tub immersion. Their water-heater usage patterns can overlap, making them difficult to separate using generic presence or flow signals.
- **Household dependency.** Bathing frequency, duration, and preferred times of day vary widely between households. Models trained globally may fail to generalize, highlighting the need for household-adaptive approaches.
- **Sensor and privacy constraints.** Many prior works rely on additional sensors (e.g., acoustic, motion, environmental). Such solutions increase cost, hinder scalability, and raise privacy concerns. A practical method must exploit signals already available in standard infrastructure.
- **Seasonal variation and label scarcity.** Showering prevalence is highest in summer and lowest in winter, while tub bathing shows the opposite trend. Moreover, direct shower labels are often scarce or unavailable, making it difficult to train models in some periods. Robust methods must therefore balance detection quality during high-prevalence seasons, suppress nuisance alarms in low-prevalence seasons, and remain effective even when direct labels are missing.

These challenges are further compounded by well-documented issues in model evaluation under class imbalance, particularly when the imbalance varies across time [48]. Recent work has also highlighted the pitfalls of large-scale imbalanced scenarios, where traditional accuracy metrics become misleading [49]. Alternative evaluation tools such as the MCC-F1 curve have been proposed to better represent classifier performance under skewed distributions [50].

To address these challenges, we propose a **dual-proxy framework**, leveraging proxy features and a proxy-driven scheme. This design targets scalable and privacy-preserving household adaptation without relying on manual labeling. The study is structured around the following research questions:

RQ1. How effective is the proxy feature `washing_seconds` in improving shower detection?

RQ2. How can a proxy-driven scheme, without direct shower labels, identify compact and reliable feature subsets?

RQ3. How robustly do proxy features and proxy-driven schemes generalize across households and seasons in shower detection?

4. Proposed Method

This section presents a dual-proxy framework for shower detection that builds on standard water-heater logs without requiring additional sensors. The framework consists of two complementary components: (A1) a proxy feature that captures washing-related activities, and (A2) a proxy-driven scheme that enables learning despite label scarcity and household-specific variation. These components are formulated in a way that is broadly applicable to residential monitoring contexts.

4.1. (A1) Proxy Feature: Washing Time

A bathroom stay can be represented by two basic quantities: the total stay duration A_i and the cumulative tub immersion time $B_{i,j}$. We define the *washing time* feature W_i as

$$W_i = A_i - \sum_j B_{i,j}.$$

This quantity approximates the duration of washing activities such as body washing and hair washing. It is expected to be large for both tub bathing and showering, while remaining small for non-bathing stays.

Washing time is attractive as a proxy feature for two reasons. First, its distribution is similar between tub bathing and showering, making it effective for separating bathing-related from non-bathing behaviors. Second, it exhibits relatively small seasonal variation, which supports its robustness for long-term monitoring. Importantly, W_i can be derived directly from standard water-heater logs without additional sensors, making it cost-effective and privacy-preserving.

4.2. (A2) Proxy-Driven Scheme: Household-Adaptive Classification

While W_i serves as a proxy feature, household-specific variation and the scarcity of direct shower labels pose additional challenges. To address this, we adopt a *proxy-driven scheme* that combines household-level adaptation with learning from proxy signals without relying on manual C_2 labels. This design enables training even when explicit shower labels are unavailable, supporting scalable deployment across large numbers of households where manual annotation would be impractical.

Each household is modeled independently to avoid bias that may arise from pooling heterogeneous users. Formally, given a set of features $\mathbf{x}_i$ for stay i and the corresponding class label $y_i \in \{C_1, C_2, C_3\}$, the goal is to learn a classifier $f_h : \mathbf{x}_i \mapsto y_i$ for each household h.

The labeling procedure is as follows:

- **Tub bathing (C_1)** and **non-bathing (C_3)** can be determined directly from sensor data (tub usage vs. no usage).
- **Showering (C_2)** cannot be inferred from sensors and was manually annotated by participants. However, C_2 labels were not used for training. Models were trained only on C_1 and C_3, while proxy features such as `washing_seconds` indirectly separate C_2 from C_3 during evaluation.

Thus, only C_1 and C_3 supervise training, while C_2 is inferred indirectly via proxy-driven learning. By relying on proxy-derived cues rather than manual C_2 labels, the framework enables label-free training that remains feasible for long-term, real-world sensing. Because the framework relies only on generic heater-log variables such as temperature, flow, and stay duration, it is compatible with diverse household water-heater models and plumbing architectures without requiring device-specific adjustments. This design enables models to be trained even in periods with sparse or absent shower labels (e.g., winter) and applied robustly across seasons.

The framework is model-agnostic: what matters is the household-level adaptation and the reliance on proxy-driven principles to capture washing activities without invasive sensing. Although LightGBM was used in this study for its interpretability, the proposed proxy-driven design is conceptually transferable to other classifiers such as SVMs or shallow neural networks, underscoring its model-agnostic nature. As illustrated in Figure 2, the proxy-driven design leverages tub vs. non-tub distinctions and proxy features to resolve the ambiguity of shower detection.

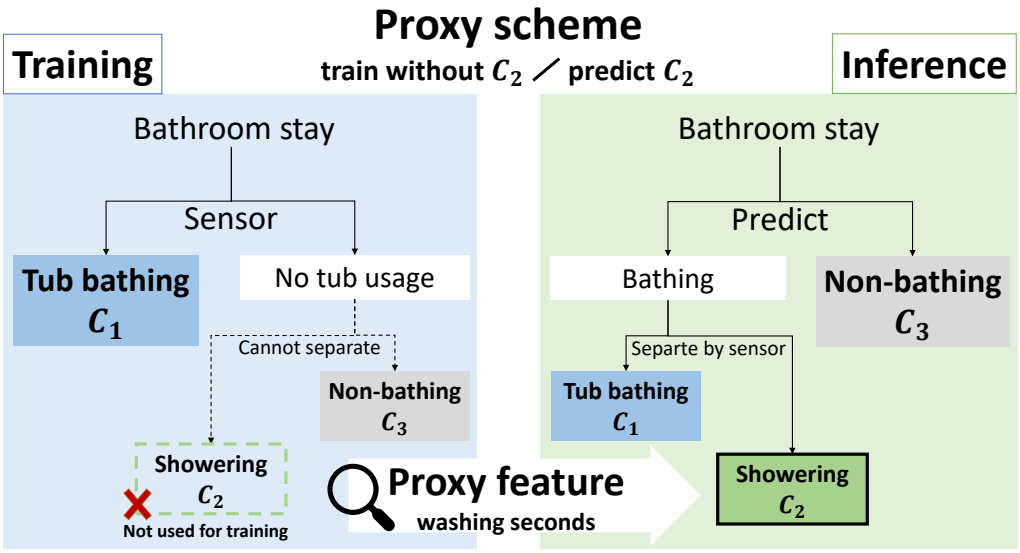


Figure 2. Proxy-driven scheme. During training (**left**), only C_1 (tub bathing) and C_3 (non-bathing) are used, while C_2 (shower) labels are excluded. During inference (**right**), proxy features such as `washing_seconds` separate C_2 from C_3.

5. Experimental Setup

This section details the dataset, data partitioning, implementation, and evaluation protocol used to assess the proposed method. All design choices were made to ensure strict reproducibility and to prevent information leakage.

5.1. Dataset

We conducted our study using year-long water-heater logs collected from two cooperating households, denoted Household 1 and Household 2. Each household contributed twelve consecutive months of data, spanning from May to the following April (Household 1: May 2023–April 2024, Household 2: May 2024–April 2025), providing a one-year basis for long-term evaluation of seasonal and household variability. The data were obtained from standard residential water heaters equipped with built-in temperature and flow sensors. Each household produced several hundred bathroom stays per year, covering diverse routines and seasonal bathing behaviors. Research participation was voluntary, and both households provided explicit consent for the use of their logs in this study. The data were collected under a corporate–academic research agreement; therefore, the raw logs are not publicly distributable, although all feature definitions and analysis procedures are fully disclosed to ensure reproducibility. As the data were collected from consenting collaborators rather than the general public, no additional ethical review was required.

Each bathroom stay was identified and labeled through a semi-automatic procedure. Entry and exit times were detected using a human-presence sensor embedded in the wall-mounted water-heater controller inside the bathroom, a configuration common in Japan but not universal elsewhere. This controller records motion-based occupancy signals together with temperature and flow data, enabling precise stay segmentation. Tub bathing (C1) was automatically detected from water-level readings, while showering without tub immersion (C2) was manually annotated through participant review to establish ground-truth labels for evaluation, and all remaining non-bathing stays (C3) were determined by exclusion.

Table 1 summarizes the features extracted from the raw logs. All features were derived directly from existing heater sensors without additional instrumentation.

Table 1. Extracted features from water-heater logs.

Feature	Abbr.	Definition
`washing_seconds`	WSec	Approximate duration of washing activities, computed as stay duration minus tub immersion time.
`water_hour`	WH	Quantity of hot-water usage during the stay, aggregated at 1-h granularity.
`gas_hour`	GH	Quantity of gas consumption during the stay, aggregated at 1-h granularity.
`max_temp_diff`	TmaxDiff	Maximum instantaneous rise in bathroom room temperature observed during the stay (1-h granularity).
`in_out_temp_diff`	Tin–Tout	Change in bathroom room temperature between the beginning and end of the stay (exit minus entry).
`time_max_min_temp_diff`	Tmax–Tmin	Forward-direction room-temperature rise during the stay, computed as $\max_{t_2>t_1}\{T(t_2) - T(t_1)\}$.
`since`	–	Numeric encoding of the stay start time (converted from wall-clock timestamp).

Note that `water_hour` and `gas_hour` are computed at the household level and may include concurrent non-bathroom activities such as kitchen or laundry usage during the stay window. This overlap reflects realistic household operation and increases the difficulty of distinguishing showering events from other hot-water use. The bathroom-specific proxy feature `washing_seconds`, derived from stay segmentation and tub immersion information, is therefore less affected by such confounds and serves as a more reliable indicator of washing behavior. These characteristics motivate the use of proxy-anchored, household-adaptive feature selection to achieve stable detection despite overlapping signals. Across the two households, the dataset serves as a representative yet privacy-conscious testbed for household-adaptive modeling, providing sufficient variability to examine seasonal and behavioral differences while respecting the confidentiality of industrial data. As this dataset involves only two cooperating households, the findings should be regarded as illustrative rather than statistically generalizable.

5.2. Data Partitioning

For each household, the dataset was divided into three disjoint subsets:

- **Winter (training).** Three consecutive winter months (November–January) were reserved exclusively for model training and hyperparameter selection. As illustrated in Figure 2, winter stays contain tub bathing (C_1) and non-bathing (C_3) events that can be directly distinguished by the sensor. Showering (C_2), which is hard to separate in other seasons, does not occur in winter and therefore does not contaminate the training set. This allowed us to train a proxy-driven model without relying on C_2 labels. For each household, this training subset covers about three months of data (approximately two to three hundred stays), yielding six months in total across both households.
- **Development (Dev).** The remaining spring–autumn data were split month-wise into two halves, with even-indexed samples assigned to Dev. This set was used to analyze feature utility (RQ1) and feature-set preference (RQ2). It spans roughly eight to nine months per household, ensuring that seasonal variations and routine differences were represented in feature evaluation.
- **Evaluation (Eval).** Odd-indexed samples from the same spring–autumn months were assigned to Eval, which was strictly held out until final assessment (RQ3). This held-out set provides an independent basis for cross-season and cross-household validation, confirming the robustness of proxy-driven learning.

All partitions were created independently for each household to prevent data leakage and to reflect the privacy and contractual constraints described in Section 5.1. This protocol

ensured that no information from Dev or Eval leaked into training, while exploiting the unique label availability in winter to realize the proxy scheme.

5.3. Implementation Details

We employed LightGBM as the classifier. All preprocessing steps (imputation, scaling, encoding) were omitted except for converting timestamps into numeric values. Hyperparameter selection was conducted exclusively on winter data. This choice reflects the proxy scheme (Figure 2): winter stays contain only tub bathing (C_1) and non-bathing (C_3), so models can be trained without ambiguity from showering (C_2). Freezing the resulting models and applying them directly to Dev and Eval sets ensured that evaluation reflects true generalization to seasons where C_2 occurs. No probability calibration was performed, as our primary aim was to test whether raw LightGBM scores—without post hoc tuning—can separate showering from non-bathing. A fixed threshold of 0.5 was applied for generating binary predictions when required (e.g., F1, false-positive counts). Each household was modeled independently rather than pooling across households, to account for household-specific usage patterns and to test adaptability at the household level.

5.4. Evaluation Metrics

Performance was assessed using both threshold-free and threshold-dependent metrics:

- **Primary metrics.** Summer PR-AUC quantified shower detection quality during peak prevalence. Non-summer false-positive rate (FPR) captured nuisance levels when showering was rare.
- **Auxiliary metrics.** Daily false alarms, macro-averaged PR-AUC and F1, worst-month F1, and calibration measures (Brier score, expected calibration error, ECE).

Metrics were aggregated using both micro- and macro-averaging. To summarize results across households, we report both household-specific values and pooled distributions. Effect sizes were defined as paired differences between models with and without the proxy feature. Uncertainty was represented by empirical distributions (median and interquartile range), rather than formal bootstrap resampling or meta-analysis.

6. Results for RQ1: Proxy Feature Utility

6.1. Setup for RQ1

To isolate the contribution of feature-level proxies, we evaluated `washing_seconds`, a duration feature derived from water-heater logs. Two conditions were compared:

- **Base:** Any subset of six candidate features excluding `washing_seconds`.
- **+Proxy:** The same subset with `washing_seconds` added.

Evaluation was conducted under a seasonal split, with July–September considered as summer and the remaining months as non-summer. The primary metrics were

- Summer: ΔPR-AUC = PR-AUC(+Proxy) $-$ PR-AUC(Base), capturing improvements in detection quality when shower prevalence is high.
- Non-summer: ΔFPR = FPR(+Proxy) $-$ FPR(Base), capturing changes in false alarms when shower prevalence is low.

Supplementary metrics included the minimum monthly F1, the standard deviation of monthly F1, and false alarms per day. Two households were analyzed individually and in a pooled (POOL) setting. For each metric, 95% bootstrap confidence intervals were computed across monthly samples to quantify the stability of observed differences.

6.2. Analysis for RQ1

Figure 3 shows the joint distribution of ΔPR-AUC (summer) and ΔFPR (non-summer) for all feature sets. Most points lie in the favorable upper-left quadrant (ΔPR-AUC > 0, ΔFPR < 0), indicating a consistent benefit of adding the proxy feature.

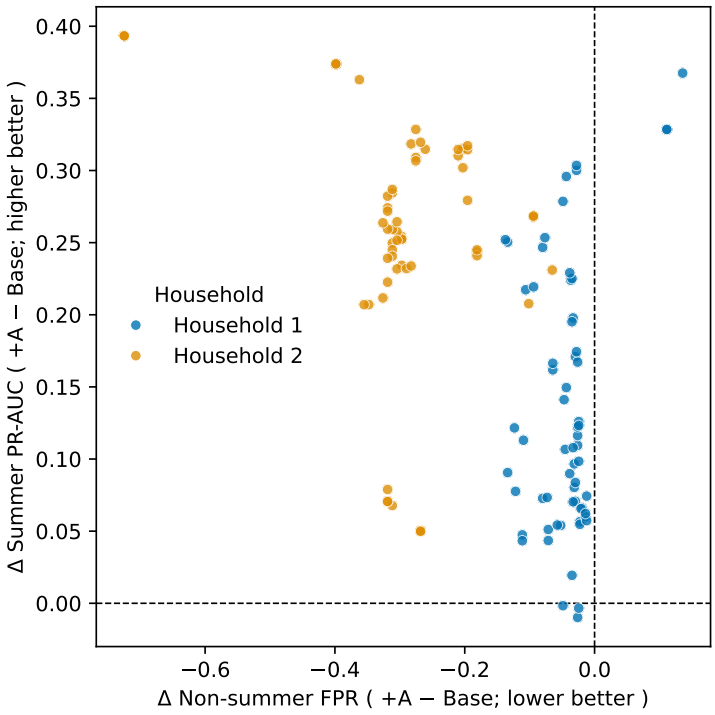


Figure 3. Scatter plot of ΔPR-AUC (summer) vs. ΔFPR (non-summer) for each household.

Household 1 exhibited moderate gains (median ΔPR-AUC $= +0.14$ [95% CI: $+0.11$, $+0.16$], ΔFPR $= -0.04$ [95% CI: -0.06, -0.03]), whereas Household 2 showed larger but more variable improvements (median ΔPR-AUC $= +0.25$ [95% CI: $+0.23$, $+0.28$], ΔFPR $= -0.29$ [95% CI: -0.32, -0.26]). The pooled analysis confirmed robust gains (median ΔPR-AUC $= +0.20$ [95% CI: $+0.18$, $+0.21$], ΔFPR $= -0.17$ [95% CI: -0.20, -0.14]). All confidence intervals excluded zero, confirming that the proxy-induced improvements were statistically significant and consistent across seasons and households.

Violin plots further illustrate these tendencies:

- **Summer:** ΔPR-AUC distributions were shifted upward (Figure 4a).
- **Non-summer:** ΔFPR distributions were shifted below zero (Figure 4b). Daily false alarms (FA/day) also tended to decrease, although a few feature sets showed slight increases (Figure 4c).

Overall, the scatter highlights that proxy gains were consistent across households, while the pooled analysis revealed stronger stability than household-specific results.

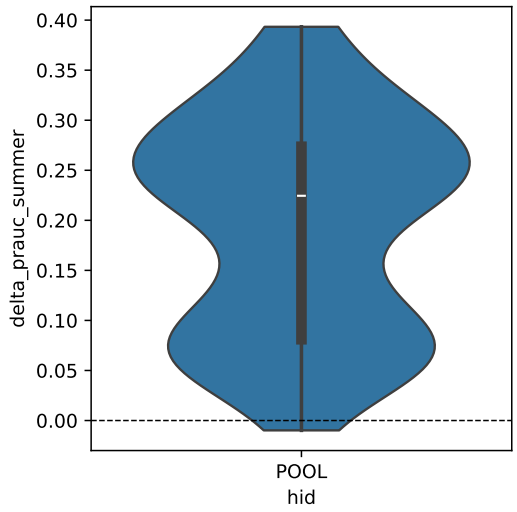
6.3. Discussion for RQ1

The proxy feature `washing_seconds` enhanced model performance in two complementary ways: it improved discrimination during high-prevalence summer periods and suppressed false alarms in non-summer months. These results demonstrate that a simple proxy feature, derived without new sensors, can resolve the ambiguity between showering and non-bathing stays.

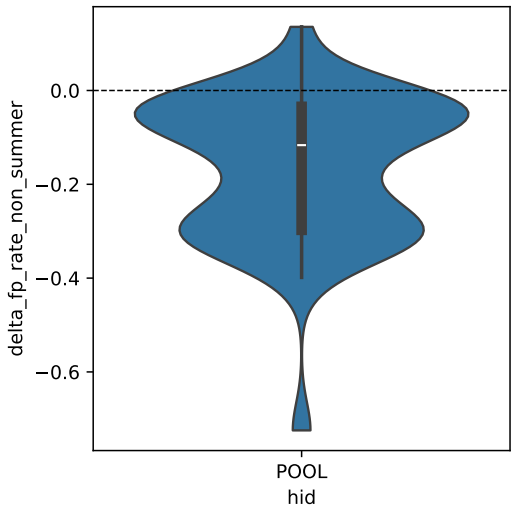
Nevertheless, residual variation across households and feature sets indicates that proxy features alone are insufficient. This motivates the design of a proxy-driven scheme

that adapts to household-specific conditions and leverages compact feature subsets (RQ2). In other words, RQ1 validates the utility of proxy features at the *feature level*, providing the foundation for scheme-level proxy design addressed in the next section.

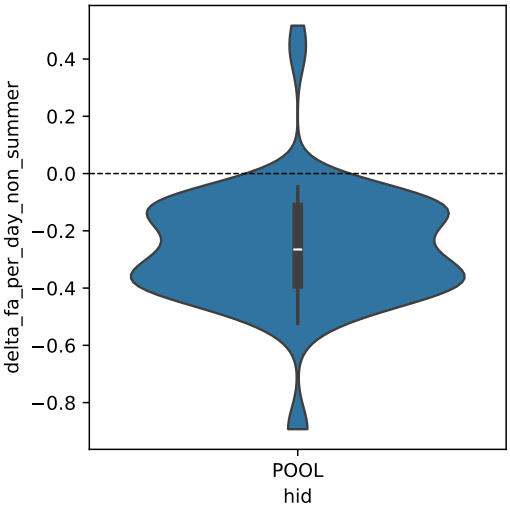
Implication for RQ1: Feature-level proxy design, exemplified by `washing_seconds`, provides a simple yet powerful way to enhance discrimination without requiring new sensors. The quantitative analysis with 95% confidence intervals confirms that the proxy effect is statistically reliable across seasons and households, supporting its use as a dependable foundation for household-adaptive proxy schemes (RQ2).



(**a**) ΔPR-AUC (summer, pooled)



(**b**) ΔFPR (non-summer, pooled)



(**c**) ΔFA/day (non-summer, pooled)

Figure 4. Violin plots of performance changes by adding Proxy. (**a**) Summer ΔPR-AUC for each household and pooled data. (**b**) Non-summer ΔFPR for each household and pooled data. (**c**) Supplementary non-summer ΔFA/day for pooled data.

7. Results for RQ2: Proxy-Driven Feature Set Selection

7.1. Setup for RQ2

RQ2 examines how a proxy-driven scheme can guide the selection of compact yet effective feature sets under multiple criteria. We enumerated all $2^7 - 1 = 127$ non-empty subsets from seven candidate features, one of which was `washing_seconds`, and evaluated them on the Dev portion of each month (spring–autumn) using winter-trained models. All

subsets were exhaustively evaluated without applying heuristic selection methods such as mutual information, recursive feature elimination, or LASSO regression. Compactness was defined strictly by the number of features used as a tie-breaker in the lexicographic ranking.

The ranking used a prioritized scheme (lexicographic key) defined as

1. **Summer PR-AUC** (higher is better),
2. **Non-summer FPR** (lower is better),
3. **Number of features** (fewer is better; tie-breaker).

Minimum monthly F1 and the standard deviation of monthly F1 were reported as auxiliary stability descriptors, but were *not* used as ranking keys.

In parallel, Pareto-front analysis was applied to highlight trade-offs between non-summer FPR (x-axis, lower is better) and summer PR-AUC (y-axis, higher is better).

7.2. Analysis for RQ2

The rank-stability analysis summarized in Table 2 focuses on the top-3 feature sets for each household and the pooled data under the prioritized ranking scheme.

Household 1. The best set, `washing_seconds+water_hour+in_out_temp_diff`, achieved PR-AUC 0.819 with FPR 0.061. Comparable 3–4 feature sets showed similar trade-offs, while a low-FPR option (`since+water_hour+time_max_min_temp_diff`) reduced FPR to 0.037 at the cost of PR-AUC (0.782).

Household 2. The top 2-feature set, `washing_seconds+max_temp_diff`, reached PR-AUC 0.857 with FPR 0.116. Gas-related sets offered a different trade-off, lowering FPR to about 0.094 while reducing PR-AUC to around 0.844.

Pooled (POOL). The best 3-feature set, `washing_seconds+water_hour+in_out_temp_diff`, achieved PR-AUC 0.802 and FPR 0.065. A gas-heavy set shifted the balance to PR-AUC 0.755 and FPR 0.052, indicating that proxy-driven subsets maintained stable trade-offs between precision and false alarms across households.

Table 2. Top-3 ranked feature sets per evaluation target (abbreviations: see Table 1).

Evaluation Target	Rank	Feature Set (Abbr.)	Summer PR-AUC	Non-Summer FPR
Household 1	1	WSec + WH + Tin–Tout	0.819	0.061
	2	WSec + WH + Tin–Tout + Tmax–Tmin	0.811	0.061
	3	WSec + WH	0.811	0.061
Household 2	1	WSec + TmaxDiff	0.857	0.116
	2	WSec + TmaxDiff + Tin–Tout	0.850	0.116
	3	WSec + TmaxDiff + Tmax–Tmin	0.850	0.116
POOL (Eval)	1	WSec + WH + Tin–Tout	0.802	0.065
	2	WSec + WH + Tin–Tout + Tmax–Tmin	0.795	0.065
	3	WSec + WH	0.794	0.070

Pareto fronts (Figure 5a–c) make these trade-offs explicit: compact 2–3 feature sets cluster in the favorable high-PR-AUC region, while gas-related sets shift toward lower FPR but with a measurable loss in PR-AUC.

7.3. Discussion for RQ2

Several principles emerge from these results:

- **Proxy as anchor:** Feature sets containing `washing_seconds` consistently ranked highest, underscoring the role of proxy features in guiding selection.
- **Compact yet effective:** Two to three feature sets achieved strong performance, suggesting a practical balance between accuracy and simplicity.

- **Consistency through proxy-driven learning:** While household-specific rankings varied, pooled results emphasized stable subsets, showing that proxy-driven evaluation can reveal generalizable configurations across households.
- **Lexico vs. Pareto:** Lexicographic ranking emphasized stability, whereas Pareto fronts highlighted explicit trade-offs.

These findings align with recent studies framing feature selection as a multi-objective optimization task [51–53]. The proxy-driven scheme demonstrates that compact, proxy-anchored subsets can be systematically identified even under label scarcity. Overall, proxy-guided evaluation maintained stable trade-offs across households and criteria, demonstrating reliable subset identification without heuristic bias. This sets the stage for RQ3, which evaluates their robustness in independent households and seasons.

Implication for RQ2: Proxy-driven selection highlights that compact, proxy-anchored subsets (2–3 features) can achieve both accuracy and stability. This systematic identification of generalizable sets paves the way for evaluating their robustness under real-world conditions (RQ3).

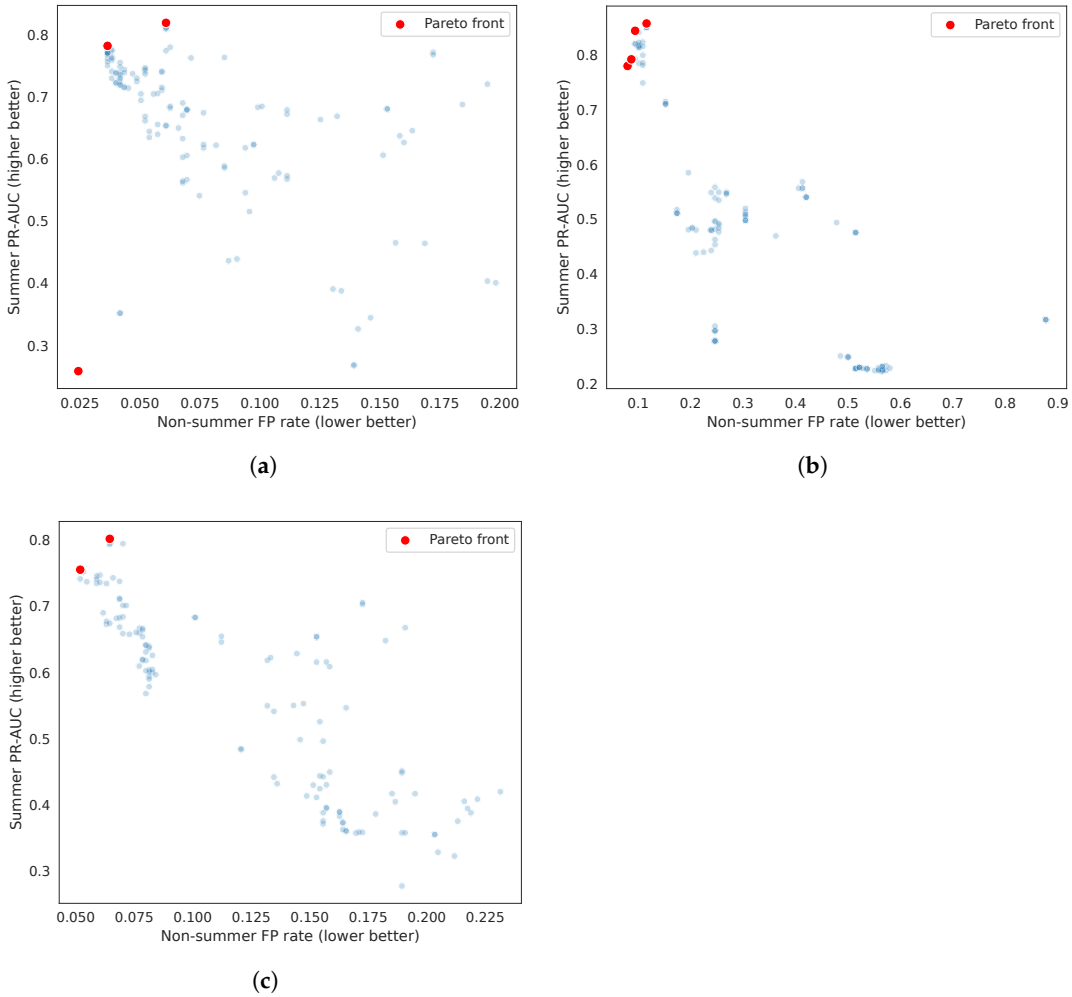


Figure 5. Pareto fronts of feature subset selection. Each point represents a candidate subset, with the axes denoting summer PR-AUC and non-summer FPR. Red points indicate front subsets, while blue points denote the others. (**a**) Household 1. (**b**) Household 2. (**c**) Pooled data (POOL).

8. Results for RQ3: Robustness of Proxy-Driven Subsets

8.1. Setup for RQ3

RQ3 evaluates the robustness of feature sets selected under proxy-driven conditions (RQ2) when deployed in real-world, long-term evaluation scenarios. This section provides an illustrative evaluation based on two cooperating households, serving as an example of real-world deployment rather than a population-level analysis. Representative subsets identified by lexicographic ranking and Pareto fronts were trained on winter data and applied to Eval portions (spring–autumn, within-month even–odd split). The Eval period covered nearly one year of seasonal data, enabling assessment of long-term robustness under natural environmental and behavioral variation. Metrics included

- **Overall performance:** micro/macro F1, micro PR-AUC,
- **Stability:** minimum monthly F1, monthly F1 standard deviation,
- **Calibration:** Brier score, Expected Calibration Error (ECE),
- **KPI:** summer PR-AUC, non-summer FPR, Recall@Prec ≥ 0.8, Prec@Rec ≥ 0.8, FA/day,
- **Generalization gap:** Dev$\rightarrow$Eval differences for the above.

8.2. Analysis for RQ3

Compact proxy-anchored subsets maintained strong performance on Eval. As shown in Figure 6, micro PR-AUC reached 0.724–0.728 (95% CI covering 0.60–0.88) for the compact sets, while the gas-including set lagged at 0.668 (CI 0.48–0.84). Macro F1 was around 0.55–0.57 across subsets (Figure 7).

Overall, compact subsets maintained stable performance across seasonal and household variations. Generalization gaps were small: micro PR-AUC differences were within ± 0.01, and macro F1 increased slightly ($+0.02$–$+0.03$). Calibration metrics (Brier, ECE) were preserved or improved (Figure 8). These consistent trends indicate that proxy-driven subsets were resilient to minor sensor noise and household-specific differences, sustaining comparable accuracy across diverse operating conditions.

Summer PR curves confirmed that compact subsets sustained PR-AUC around 0.87, while the gas-including set fell to 0.81 (Figure 9). KPI trade-offs (Figure 10) further showed that compact subsets achieved higher summer PR-AUC (0.865–0.870) with modest non-summer FPR (0.082–0.086), suppressing daily false alarms to 0.23. The gas-including set minimized FPR (0.054) and FA/day (0.15), but at the cost of lower summer PR-AUC (0.806).

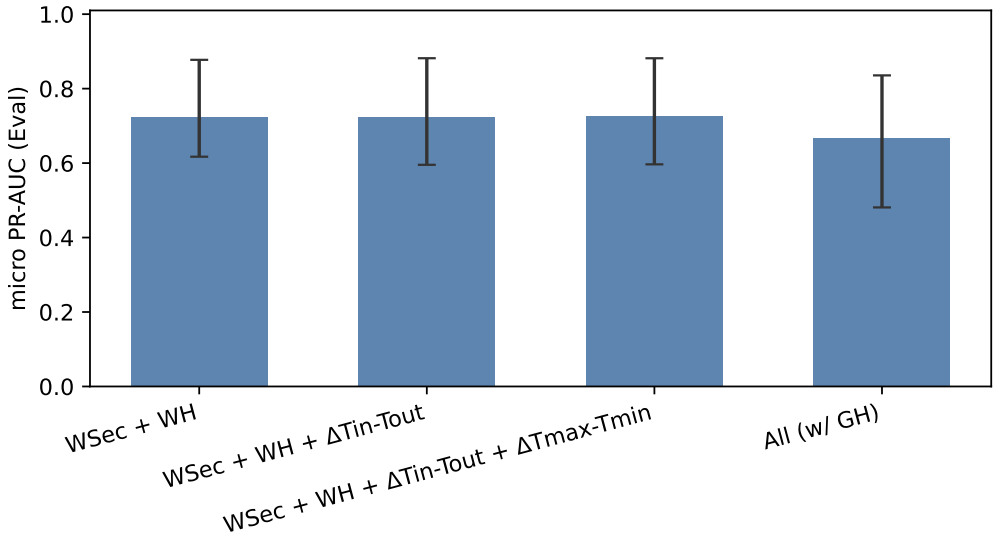


Figure 6. Micro PR-AUC on Eval data for representative feature sets (abbreviations: see Table 1). Error bars indicate 95% confidence intervals aggregated by monthly bootstrap.

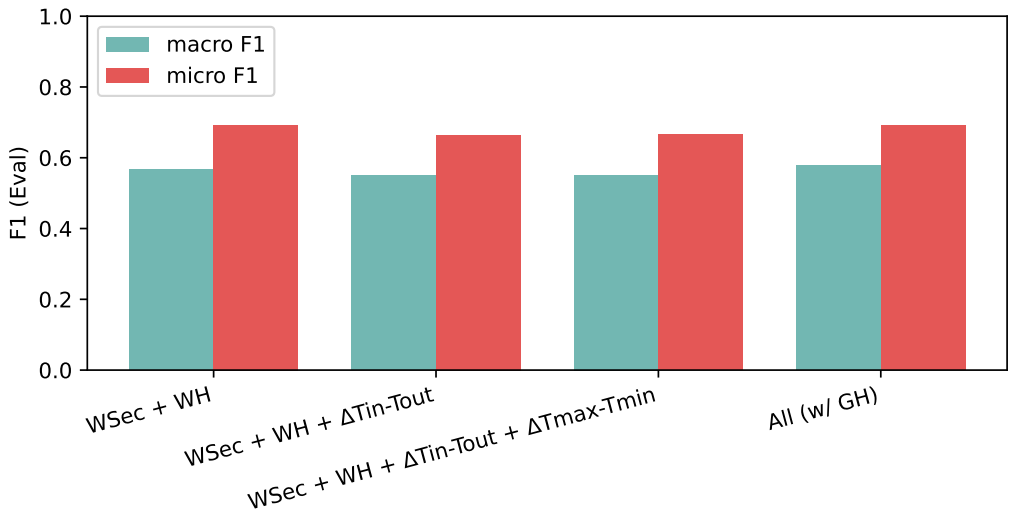


Figure 7. Macro and micro F1 on Eval data for representative feature sets (abbreviations: see Table 1).

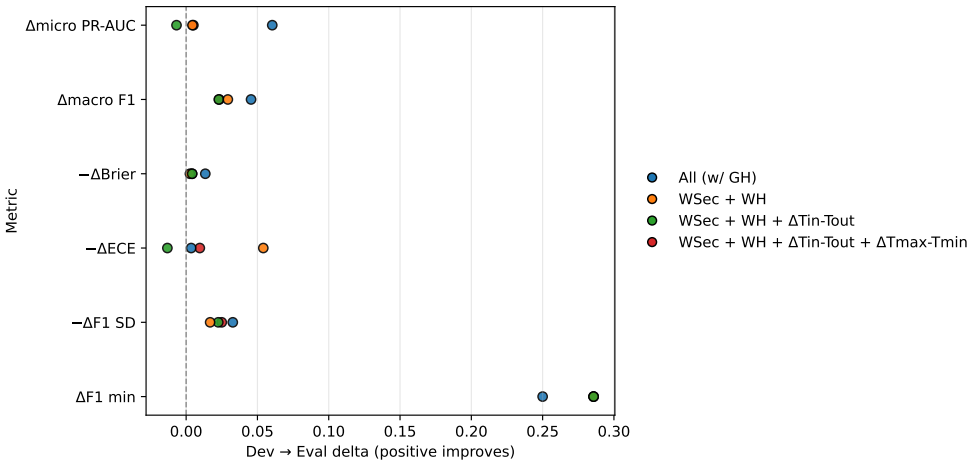


Figure 8. Generalization gap (Dev → Eval) across multiple metrics (abbreviations: see Table 1).

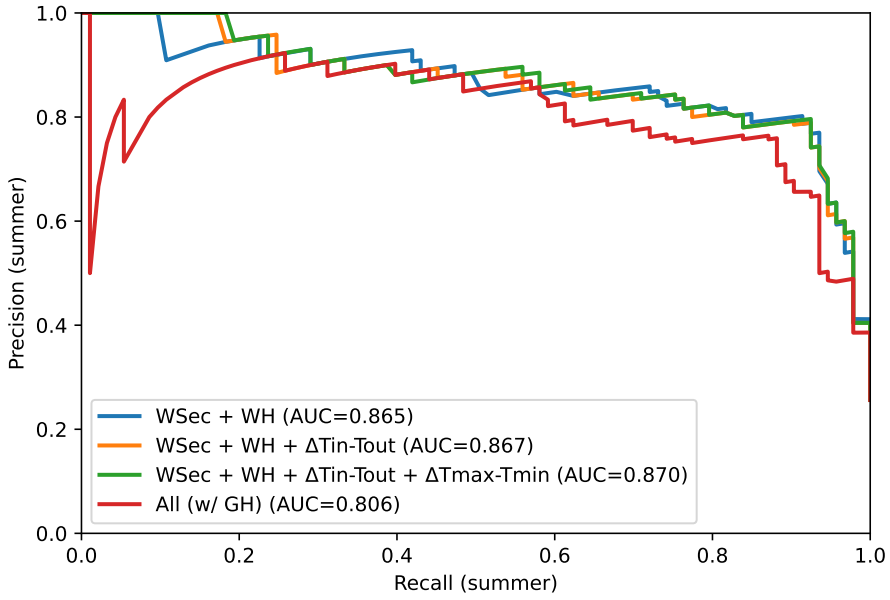


Figure 9. Summer PR curves (Eval) for representative feature sets (abbreviations: see Table 1).

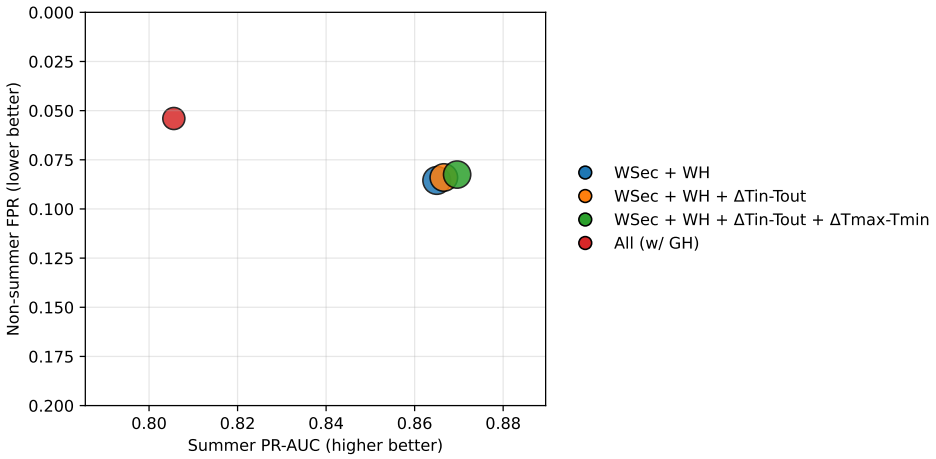


Figure 10. KPI scatter plot (Eval) (abbreviations: see Table 1). X-axis: non-summer FPR (lower is better). Y-axis: summer PR-AUC (higher is better). Point size: FA/day (smaller is better).

8.3. Discussion for RQ3

- **Robustness.** Proxy-driven subsets retained strong performance on Eval: micro PR-AUC stayed at 0.724–0.728 with minimal generalization gaps ($\pm$0.01), and macro F1 increased slightly ($+0.02$–$+0.03$). Calibration metrics were preserved or improved, underscoring reliability beyond raw accuracy.
- **Seasonal resilience.** Summer PR-AUC remained high (0.865–0.870) for compact subsets, while non-summer FPR was consistently moderate (0.082–0.086). The gas-including set reduced FPR further (0.054) but compromised summer PR-AUC (0.806).
- **Operational implications.** Compact subsets balanced detection and nuisance, suppressing daily false alarms to about 0.23 without sacrificing summer detection quality. In contrast, gas-heavy sets minimized false alarms (0.15/day) but traded off discrimination in peak seasons.

These findings highlight that compact proxy-driven subsets are not only accurate but also stable and well-calibrated, indicating their robustness for long-term household deployment. As each household was modeled independently, the present evaluation reflects within-household generalization across seasons rather than cross-household transfer. Nevertheless, the consistent performance trends observed across the two year-long datasets suggest potential applicability to new households, supporting future extension toward scalable deployment without retraining. While accuracy is often emphasized in model selection, large-scale studies show that calibration is equally critical [54]. Surveys further identify calibration as a key open challenge [46,55], and evidence from medical imaging demonstrates that calibration directly affects trustworthiness in practice [56]. In our dataset, shower-only events were more frequent during summer months, while bathtub soaking was relatively common in winter. This seasonal variation reflects users' comfort preferences and ambient temperature. The observed patterns are dataset-specific and intended to exemplify model behavior within the studied households.

Implication for RQ3: Proxy-driven subsets effective under controlled settings remained robust in real-world evaluation, showing that the dual-proxy framework scales to household deployment with both accuracy and trustworthiness.

9. Discussion

This study demonstrated that proxy can operate at two complementary levels: as a feature (`washing_seconds`) and as a scheme (proxy-driven training under label scarcity). RQ1 confirmed the effectiveness of feature-level proxy, RQ2 showed that scheme-level

proxy systematically identifies compact subsets, and RQ3 established that such subsets remain robust in real-world evaluation. Together, these findings validate the proposed dual-proxy framework for household-adaptive shower detection. Although LightGBM was adopted in this study, the dual-proxy concept itself is model-agnostic and can be transferred to other machine-learning architectures that operate under label-scarce or unlabeled conditions.

9.1. Integration of RQ Findings

- **RQ1 (feature-level proxy):** `washing_seconds` consistently improved discrimination, raising summer PR-AUC by about $+0.23$ (pooled median) and reducing non-summer FPR by -0.12. These gains held across both households.
- **RQ2 (scheme-level proxy):** Proxy-driven evaluation identified compact 2–3 feature sets anchored by `washing_seconds`, often combined with `water_hour` or `in_out_temp_diff`. Such sets achieved PR-AUC ≈ 0.81 with FPR ≈ 0.06, while gas-related sets traded higher FPR control (down to 0.05) for lower PR-AUC (≈ 0.75). Pooled analysis emphasized convergence toward generalizable sets.
- **RQ3 (dual-proxy robustness):** Subsets selected in Dev retained performance in Eval: micro PR-AUC remained 0.724–0.728 with generalization gaps within ± 0.01, summer PR-AUC stayed near 0.87, and non-summer FPR was 0.082–0.086. Calibration was stable or improved, and daily false alarms were suppressed to about 0.23.

9.2. Cross-Cutting Themes

Three themes emerge across RQs:

- **Proxy as unifying principle:** Proxy features and proxy-driven schemes jointly address ambiguity, label scarcity, and household variability.
- **Compactness and generalizability:** Compact 2–3 feature subsets balanced accuracy and robustness, reducing overfitting risk while supporting scalability.
- **Household adaptation:** Despite household-specific variation, pooled analyses consistently emphasized stable subsets. The proxy-based approach also showed resilience to minor sensor noise and household-specific fluctuations, maintaining consistent trends across environments.

9.3. Theoretical and Practical Implications

Theoretically, this work extends proxy from feature design to a methodological principle for label-free learning. The combination of lexicographic and Pareto analysis provides a reproducible strategy for feature selection under competing criteria [51–53]. Practically, household-adaptive models can be realized using only water-heater logs. Compact proxy-driven subsets reduced false alarms to 0.23/day while maintaining summer PR-AUC near 0.87, supporting cost-efficient, interpretable, and privacy-preserving deployment. In practical use, such models can be deployed in new households without full retraining, facilitating lightweight adaptation and reducing maintenance costs. Beyond shower detection, the dual-proxy framework could be extended to other smart-home and healthcare applications that require robust behavioral estimation under label scarcity. The proposed framework achieves a practical balance between accuracy, privacy, and energy efficiency by leveraging existing water-heater infrastructure. Because it requires no additional sensors or continuous high-frequency monitoring, it minimizes both sensing overhead and computational cost while maintaining reliable detection accuracy. This balance makes the approach suitable for sustainable, real-world deployment.

9.4. Limitations and Future Directions

This study involved two households only, with training restricted to winter data and evaluation limited to one year. Seasonal tendencies (e.g., more frequent showering in summer) were observed within our dataset, yet these trends may not generalize to different climates or cultural contexts. Because the year-long evaluation encompassed natural fluctuations in daily routines, including temporary variations such as guest visits or vacations, the consistent monthly PR-AUC and F1 trends suggest that the proxy-driven models were resilient to moderate household-dynamic changes. However, substantial structural or technological modifications (e.g., renovations or new appliances) remain beyond the current study scope and warrant further investigation in future work. In this study, model parameters were fixed after the winter training phase to isolate proxy-driven generalization effects. For long-term real-world deployment, however, periodic retraining or online adaptation will be required to accommodate gradual changes in household routines and maintain stable performance over time. Future work should expand to more diverse households, integrate additional proxy features, and extend proxy-driven schemes to other daily behaviors, including elderly care and energy management, where scalable, adaptive estimation is essential.

10. Conclusions

This study proposed a dual-proxy framework for household-adaptive shower detection, combining feature-level proxy design and a proxy-driven training scheme. Through three research questions, we showed that

- Feature-level proxy (`washing_seconds`) improves discrimination, increasing summer PR-AUC and reducing non-summer false alarms (RQ1).
- Proxy-driven evaluation identifies compact and generalizable subsets, typically requiring only 2–3 features anchored by `washing_seconds` (RQ2).
- These subsets remain robust in long-term evaluation, with small generalization gaps, preserved calibration, and favorable operational KPIs (RQ3).

Taken together, these findings demonstrate that proxy can function both as a feature and as a methodological principle. The dual-proxy framework enables accurate, stable, and cost-efficient behavioral sensing from existing infrastructure, offering a scalable path for real-world smart home applications. Future work will extend proxy-driven approaches to diverse households and broader daily activities, further advancing adaptive and privacy-preserving monitoring. Although the present implementation used LightGBM, the proposed framework is model-agnostic and applicable to other machine-learning architectures that operate under label-scarce or unlabeled conditions. In practical use, the framework can be adapted to new households or devices without extensive retraining, supporting lightweight and sustainable deployment. In future work, this concept could be extended beyond shower detection to other household and healthcare activities, supporting adaptive and privacy-preserving monitoring at scale.

Author Contributions: Conceptualization, T.N. and M.N.; methodology, T.N.; software, T.N.; validation, T.N.; formal analysis, T.N.; investigation, T.N., J.H. and A.Y.; resources, J.H., A.Y. and M.N.; data curation, J.H. and M.N.; writing—original draft preparation, T.N.; writing—review and editing, T.N., J.H. and M.N.; visualization, T.N.; supervision, A.Y. and M.N.; project administration, J.H.; funding acquisition, M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NORITZ Corporation under a collaborative research agreement with Kobe University. This research was partially supported by JSPS KAKENHI Grant Numbers JP25H01167, JP25K02946, JP25K24389, JP24K02765, JP24K02774, JP23K17006, JP23K28091, JP23K28383.

Institutional Review Board Statement: Ethical review and approval were waived for this study, as the data were obtained from research collaborators who provided consent for research use. The study did not involve patients, the general public, or animals.

Informed Consent Statement: Informed consent was obtained from all research collaborators involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author, subject to approval by NORITZ Corporation. The data are not publicly available due to privacy and contractual restrictions.

Acknowledgments: The authors thank NORITZ Corporation for providing the data used in this study and for their collaboration. Computational resources were provided by Kobe University. The authors appreciate anonymous subjects who participated in the experiments.

Conflicts of Interest: J.H. and A.Y. are employees of NORITZ Corporation. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The funders had no role in the design of the study; in the analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ECE	Expected Calibration Error
FPR	False-Positive Rate
FA/day	False Alarms Per Day
F1	F1-score (harmonic mean of precision and recall)
LightGBM	Light Gradient Boosting Machine
MCC	Matthews Correlation Coefficient
NILM	Non-Intrusive Load Monitoring
PR-AUC	Precision–Recall Area Under Curve
Prec@Rec	Precision at fixed Recall
Rec@Prec	Recall at fixed Precision
SVM	Support Vector Machine

References

1. Cook, D.J.; Crandall, A.S.; Thomas, B.L.; Krishnan, N.C. CASAS: A Smart Home in a Box. *Computer* **2013**, *46*, 62–69. [CrossRef]
2. Bouchabou, D.; Nguyen, S.M.; Lohr, C.; LeDuc, B.; Kanellos, I. A Survey of Human Activity Recognition in Smart Homes Based on IoT Sensors Algorithms: Taxonomies, Challenges, and Opportunities with Deep Learning. *Sensors* **2021**, *21*, 6037 . [CrossRef]
3. Tsanousa, A.; Meditskos, G.; Vrochidis, S.; Kompatsiaris, I. Multi-Sensors for Human Activity Recognition. *Sensors* **2023**, *23*, 4617. [CrossRef] [PubMed]
4. Majumder, S.; Aghayi, E.; Noferesti, M.; Memarzadeh-Tehran, H.; Mondal, T.; Pang, Z.; Deen, M.J. Smart Homes for Elderly Healthcare—Recent Advances and Research Challenges. *Sensors* **2017**, *17*, 2496. [CrossRef] [PubMed]
5. Kaur, H.; Rani, V.; Kumar, M. Human activity recognition: A comprehensive review. *Expert Syst.* **2024**, *41*, e13680. [CrossRef]
6. Dhekane, S.G.; Ploetz, T. Transfer Learning in Human Activity Recognition: A Survey. *arXiv* **2024**, arXiv.2401.10185. [CrossRef]
7. Thukral, M.; Dhekane, S.G.; Hiremath, S.K.; Haresamudram, H.; Ploetz, T. Layout-Agnostic Human Activity Recognition in Smart Homes through Textual Descriptions Of Sensor Triggers (TDOST). In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, Espoo, Finland, 14–16 October 2025; Volume 9. [CrossRef]
8. Bouchabou, D.; Nguyen, S.M.; Lohr, C.; LeDuc, B.; Kanellos, I. Using Language Model to Bootstrap Human Activity Recognition Ambient Sensors Based in Smart Homes. *Electronics* **2021**, *10*, 2498. [CrossRef]
9. Oluwalade, B.; Neela, S.; Wawira, J.; Adejumo, T.; Purkayastha, S. Human Activity Recognition using Deep Learning Models on Smartphones and Smartwatches Sensor Data. *arXiv* **2021**. [CrossRef]
10. Rypicz, L.; Witczak, I.; Supinova, M.; Salehi, H.P.; Jarabicova, O. Alarm fatigue and sleep quality in healthcare workers. *Eur. J. Public Health* **2024**, *34*, ckae144.890. [CrossRef]

11. Woo, M.; Bacon, O. Alarm Fatigue. In *Making Healthcare Safer III: A Critical Analysis of Existing and Emerging Patient Safety Practices*; Hall, K.K., Shoemaker-Hunt, S., Hoffman, L., Richard, S., Gall, E., Schoyer, E., Costar, D., Gale, B., Schiff, G., Miller, K., et al., Eds.; Agency for Healthcare Research and Quality (AHRQ): Rockville, MD, USA, 2020.
12. Fu, C.; Zeng, Q.; Du, X. HAWatcher: Semantics-Aware Anomaly Detection for Appified Smart Homes. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), Virtual, 11–13 August 2021; pp. 4223–4240.
13. Aguardio Shower Sensor. Available online: https://aguardio.com/products/aguardio-shower-sensor/ (accessed on 3 November 2025).
14. Wilhelm, S.; Kasbauer, J.; Jakob, D.; Elser, B.; Ahrens, D. Exploiting Smart Meter Water Consumption Measurements for Human Activity Event Recognition. *J. Sens. Actuator Netw.* **2023**, *12*, 46. [CrossRef]
15. Hyun, S.H. Sound-Event Detection of Water-Usage Activities Using Transfer Learning. *Sensors* **2024**, *24*, 22. [CrossRef]
16. Zhang, Y.; D'Haeseleer, I.; Coelho, J.; Vanden Abeele, V.; Vanrumste, B. Recognition of Bathroom Activities in Older Adults Using Wearable Sensors: A Systematic Review and Recommendations. *Sensors* **2021**, *21*, 2176. [CrossRef]
17. Marín-García, D.; Bienvenido-Huertas, D.; Moyano, J.; Rubio-Bellido, C.; Rodríguez-Jiménez, C.E. Detection of activities in bathrooms through deep learning and environmental data graphics images. *Heliyon* **2024**, *10*, e26942. [CrossRef]
18. Muliadi, M. Identification Water Level Monitoring System and Person Detection in Bathroom Using Iot Connet Application. *J. Media Elektr.* **2024**, *22*, 12–18. [CrossRef]
19. Yu, K.; Wu, C.Y.; Barnes, L.L.; Silbert, L.C.; Beattie, Z.; Croff, R.; Miller, L.; Dodge, H.H.; Kaye, J.A. Life-Space Mobility Is Related to Loneliness Among Living-Alone Older Adults: Longitudinal Analysis with Motion Sensor Data. *J. Am. Geriatr. Soc.* **2025**, *73*, 1125–1134. [CrossRef] [PubMed]
20. Froehlich, J.E.; Larson, E.; Campbell, T.; Haggerty, C.; Fogarty, J.; Patel, S.N. HydroSense: Infrastructure-mediated single-point sensing of whole-home water activity. In Proceedings of the 11th International Conference on Ubiquitous Computing, Orlando, FL, USA, 30 September–3 October 2009; pp. 235–244. [CrossRef]
21. Thomaz, E.; Bettadapura, V.; Reyes, G.; Sandesh, M.; Schindler, G.; Plötz, T.; Abowd, G.D.; Essa, I. Recognizing water-based activities in the home through infrastructure-mediated sensing. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 85–94. [CrossRef]
22. Klein Brinke, J.; Chiumento, A.; Havinga, P. Personal Hygiene Monitoring Under the Shower Using Wi-Fi Channel State Information. In Proceedings of the CHIIoT 2021, Delft, The Netherlands, 17 February 2021.
23. Amaxilatis, D.; Chatzigiannakis, I.; Tselios, C.; Tsironis, N.; Niakas, N.; Papadogeorgos, S. A Smart Water Metering Deployment Based on the Fog Computing Paradigm. *Appl. Sci.* **2020**, *10*, 1965. [CrossRef]
24. Hiremath, S.K.; Plötz, T. The Lifespan of Human Activity Recognition Systems for Smart Homes. *Sensors* **2023**, *23*, 7729. [CrossRef] [PubMed]
25. Kaselimi, M.; Protopapadakis, E.; Voulodimos, A.; Doulamis, N.; Doulamis, A. Towards Trustworthy Energy Disaggregation: A Review of Challenges, Methods, and Perspectives for Non-Intrusive Load Monitoring. *Sensors* **2022**, *22*, 5872. [CrossRef]
26. Cruz-Rangel, D.; Ocampo-Martinez, C.; Diaz-Rozo, J. Online non-intrusive load monitoring: A review. *Energy Nexus* **2025**, *17*, 100348. [CrossRef]
27. Gillman, M.D.; Donnal, J.S.; Paris, J.; Leeb, S.B.; Sayed, M.A.H.E.; Wertz, K.; Schertz, S. Energy Accountability Using Nonintrusive Load Monitoring. *IEEE Sens. J.* **2014**, *14*, 1923–1931. [CrossRef]
28. Athey, S.; Chetty, R.; Imbens, G.W.; Kang, H. *The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely*; Working Paper 26463; National Bureau of Economic Research: Cambridge, MA, USA, 2019. [CrossRef]
29. Tripuraneni, N.; Richardson, L.; D'Amour, A.; Soriano, J.; Yadlowsky, S. Choosing a Proxy Metric from Past Experiments. *arXiv* **2024**, arXiv.2309.07893. [CrossRef]
30. Pinnow, J.; Masoud, M.; Elhenawy, M.; Glaser, S. A review of naturalistic driving study surrogates and surrogate indicator viability within the context of different road geometries. *Accid. Anal. Prev.* **2021**, *157*, 106185. [CrossRef] [PubMed]
31. Addula, S.R.; Ramaswamy, Y.; Dawadi, D.; Khan, Z.; Veeramachaneni, P.; Pamidi Venkata, A.K. Blockchain-Enabled Healthcare Optimization: Enhancing Security and Decision-Making Using the Mother Optimization Algorithm. In Proceedings of the 2025 International Conference on Intelligent and Cloud Computing (ICoICC), Bhubaneswar, India, 2–3 May 2025; pp. 1–8.
32. Al-Marridi, A.Z.; Mohamed, A.; Erbad, A. Optimized blockchain-based healthcare framework empowered by mixed multi-agent reinforcement learning. *J. Netw. Comput. Appl.* **2024**, *224*, 103834. [CrossRef]
33. Arpitha, T.; Chouhan, D.; Shreyas, J. A Hybrid Optimization Approach to Enhance Source Location Privacy for IoT Healthcare. *IEEE Access* **2024**, *12*, 132801–132816. [CrossRef]
34. Freitas, A.A. The Case for Hybrid Multi-Objective Optimisation in High-Stakes Machine Learning Applications. *SIGKDD Explor. Newsl.* **2024**, *26*, 24–33. [CrossRef]
35. Peitz, S.; Hotegni, S.S. Multi-objective deep learning: Taxonomy and survey of the state of the art. *Mach. Learn. Appl.* **2025**, *21*, 100700. [CrossRef]

36. Liu, Z.; Chang, B.; Cheng, F. An interactive filter-wrapper multi-objective evolutionary algorithm for feature selection. *Swarm Evol. Comput.* **2021**, *65*, 100925. [CrossRef]
37. Hashemi, A.; Bagher Dowlatshahi, M.; Nezamabadi-pour, H. An efficient Pareto-based feature selection algorithm for multi-label classification. *Inf. Sci.* **2021**, *581*, 428–447. [CrossRef]
38. Pang, Y.; Wang, A.; Lian, Y.; Li, J.; Sun, G. A Multi-objective Optimization Method for Joint Feature Selection and Classifier Parameter Tuning. In Proceedings of the Knowledge Science, Engineering and Management: 15th International Conference, KSEM 2022, Singapore, 6–8 August 2022; pp. 237–248. [CrossRef]
39. Demir, K.; Nguyen, B.H.; Xue, B.; Zhang, M. Sparsity-based evolutionary multi-objective feature selection for multi-label classification. In Proceedings of the Genetic and Evolutionary Computation Conference Companion, Lille, France, 10–14 July 2021; pp. 147–148. [CrossRef]
40. Namakin, M.; Rouhani, M.; Sabzekar, M. A metaheuristic multi-objective interaction-aware feature selection method. *arXiv* **2022**, arXiv.2211.05423. [CrossRef]
41. Al-Tashi, Q.; Abdulkadir, S.J.; Rais, H.M.; Mirjalili, S.; Alhussian, H. Approaches to Multi-Objective Feature Selection: A Systematic Literature Review. *IEEE Access* **2020**, *8*, 125076–125096. [CrossRef]
42. Ruan, J.; Wang, M.; Liu, D.; Chen, M.; Gao, X. Multi-Label Feature Selection with Feature-Label Subgraph Association and Graph Representation Learning. *Entropy* **2024**, *26*, 992. [CrossRef]
43. Zhang, P.; Yin, H.; Tian, Y.; Zhang, X. An adjoint feature-selection-based evolutionary algorithm for sparse large-scale multiobjective optimization. *Complex Intell. Syst.* **2025**, *11*, 127. [CrossRef]
44. Dawadi, P.; Cook, D.J.; Schmitter-Edgecombe, M. Smart home-based longitudinal functional assessment. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, Seattle, WA, USA, 13–17 September 2014; pp. 1217–1224. [CrossRef]
45. Silva Filho, T.; Song, H.; Perello-Nieto, M.; Santos-Rodriguez, R.; Kull, M.; Flach, P. Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Mach. Learn.* **2023**, *112*, 3211–3260. [CrossRef]
46. Wang, C. Calibration in Deep Learning: A Survey of the State-of-the-Art. *arXiv* **2025**, arXiv.2308.01222. [CrossRef]
47. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. *arXiv* **2017**, arXiv.1706.04599. [CrossRef]
48. Brabec, J.; Komárek, T.; Franc, V.; Machlica, L. On Model Evaluation Under Non-constant Class Imbalance. In Proceedings of the Computational Science—ICCS 2020, Amsterdam, The Netherlands, 3–5 June 2020; pp. 74–87.
49. Hancock, J.T.; Khoshgoftaar, T.M.; Johnson, J.M. Evaluating classifier performance with highly imbalanced Big Data. *J. Big Data* **2023**, *10*, 42. [CrossRef]
50. Cao, C.; Chicco, D.; Hoffman, M.M. The MCC-F1 curve: A performance evaluation technique for binary classification. *arXiv* **2020**. [CrossRef]
51. Zhang, K.; Liu, Y.; Wang, X.; Mei, F.; Sun, G.; Zhang, J. Enhancing IoT (Internet of Things) feature selection: A two-stage approach via an improved whale optimization algorithm. *Expert Syst. Appl.* **2024**, *256*, 124936. [CrossRef]
52. Gao, Z.; Mo, H.; Yan, Z.; Fan, Q. A Multimodal Multi-Objective Feature Selection Method for Intelligent Rating Models of Unmanned Highway Toll Stations. *Biomimetics* **2024**, *9*, 613. [CrossRef]
53. Sharma, S.; Kumar, V.; Dutta, K. Multi-objective optimization algorithms for intrusion detection in IoT networks: A systematic review. *Internet Things-Cyber-Phys. Syst.* **2024**, *4*, 258–267. [CrossRef]
54. Dheur, V.; Ben Taieb, S. A Large-Scale Study of Probabilistic Calibration in Neural Network Regression. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; Volume 202, pp. 7813–7836.
55. Vasilev, R.; D'yakonov, A. Calibration of Neural Networks. *arXiv* **2023**, arXiv.2303.10761. [CrossRef] [PubMed]
56. Sambyal, A.S.; Niyaz, U.; Krishnan, N.C.; Bathula, D.R. Understanding calibration of deep neural networks for medical image classification. *Comput. Methods Programs Biomed.* **2023**, *242*, 107816. [CrossRef] [PubMed]