사물인터넷융복합논문지 Vol. 12, No. 2, pp. 79-85, 2026 https://doi.org/10.20465/KIOTS.2026.12.2.079

An Empirical Study on a Domain Adaptation Framework for Personalized Exercise Intensity Prediction Using Smartwatch Sensing

Jae-Hyuk Lee*
Research Professor, Institute of IT Convergence Technology, Seoul National University of Science and Technology

스마트워치 센싱 기반 개인화된 운동 강도 예측을 위한 도메인 적응 프레임워크 실증 연구

이재혁*
서울과학기술대학교 IT융합기술연구소 연구교수

Abstract This study investigated cross-subject generalization in smartwatch-based exercise intensity classification and proposed a lightweight domain adaptation framework under constrained sampling conditions. Exercise intensity was defined relative to individual peak heart rate (%HR_peak) using wrist photoplethysmography and inertial signals. A lightweight 1D-CNN was trained and evaluated using a subject-wise hold-out design to assess cross-subject domain transfer. An unsupervised test-time batch normalization (BN) recalibration strategy was applied to mitigate distribution mismatch. The model achieved strong validation performance (macro-F1 = 0.88, accuracy = 0.87). However, macro-F1 declined to 0.31 under cross-subject evaluation, indicating substantial degradation in detecting moderate- and high-intensity segments. BN adaptation improved macro-F1 to 0.35 and enhanced moderate-intensity detection. These findings demonstrate that cross-subject domain shift fundamentally constrains model generalization and that normalization-based lightweight adaptation can partially mitigate distribution mismatch. The proposed framework provides a scalable foundation for robust model development in wearable-based exercise monitoring.

Key Words : Cross-subject generalization, Domain adaptation, Exercise intensity prediction, Wearable sensing, 1D-CNN

요 약 본 연구는 스마트워치 기반 운동 강도 분류에서 피험자 간 일반화 성능을 분석하고, 제한된 표본 조건에서 적용 가능한 경량 도메인 적응 프레임워크를 제안하였다. 손목 광용적맥파 및 관성 센서 신호를 활용하여 운동 강도를 개인별 최대 심박수 대비 상대 강도(%HR_peak)로 정의하였으며, 경량 1D-CNN을 학습하였다. 피험자 간 도메인 전이 성능은 subject-wise hold-out 설계를 통해 평가하였고, 분포 불일치를 완화하기 위해 비지도 테스트 시점 배치 정규화(BN) 재보정 기법을 적용하였다. 검증 세트에서 모델은 macro-F1 0.88, 정확도 0.87의 우수한 성능을 보였다. 그러나 피험자 간 평가에서는 macro-F1이 0.31로 감소하여 중·고강도 구간 탐지에서 현저한 성능 저하가 나타났다. BN 적응 적용 후 macro-F1은 0.35로 향상되었으며, 중강도 구간의 탐지 성능이 개선되었다. 이러한 결과는 피험자 간 도메인 전이가 모델 일반화 성능을 구조적으로 제약함을 보여주며, 정규화 기반 도메인 적응 전략이 분포 불일치 완화에 기여할 수 있음을 시사하며, 웨어러블 센싱 기반 운동 모니터링에서 강건한 모델 설계를 위한 기술적 근거를 제공한다.

주제어 : 도메인 일반화, 도메인 적응 학습, 운동 강도 예측, 착용형 센싱, 합성곱 신경망

본 논문은 2023년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. RS-2023-00275579).
*교신저자 : 이재혁(jhl9405@seoultech.ac.kr)
접수일 2026년 03월 03일 수정일 2026년 04월 06일 심사완료일 2026년 04월 11일

1. Introduction

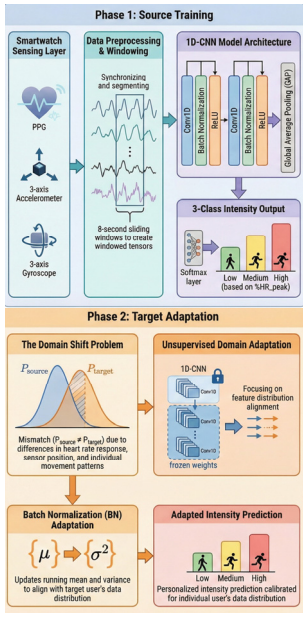
Recent advances in wearable technology and AI-driven analytics have transformed sports science and exercise prescription [1]. Consumer smartwatches now operate as portable sensing platforms capable of continuously acquiring real-time physiological and motion data, including heart rate, accelerometry, and gyroscopic signals. Integration of these multimodal data streams with deep learning enables objective quantification of exercise responses and automated estimation of individualized exercise status [2,3]. Importantly, such approaches extend monitoring beyond laboratory assessments and subjective indicators such as ratings of perceived exertion (RPE), facilitating continuous evaluation in free-living environments [4].

Exercise intensity is a key determinant of training adaptation and clinical outcomes, influencing cardiorespiratory fitness, metabolic health, and rehabilitation efficacy [5]. In aerobic exercise prescription, relative intensity expressed as a percentage of maximal or peak heart rate (%HR_max or %HR_peak) is widely used [6]. However, accurate real-world assessment and timely feedback remain challenging, motivating the development of automated smartwatch-based intensity classification systems with applications in athletic training, preventive health, cardiac rehabilitation, and older adult exercise guidance.

Deep neural networks are effective in modeling nonlinear patterns in time-series physiological signals [7]. In particular, one-dimensional convolutional neural networks (1D-CNNs) efficiently extract temporal features from heart rate and inertial sensor data [8]. Although strong intra-subject performance has been reported, performance often degrades substantially under inter-subject evaluation. This limitation stems from inter-individual variability in cardiovascular dynamics, movement patterns, and signal morphology [9,10], which constrains model generalization.

From a machine learning perspective, this performance gap reflects domain shift, where discrepancies between training and target data distributions impair predictive accuracy [11, 12]. In wearable-based exercise monitoring, such shifts arise from variability in physiological responses, sensor placement, skin properties, and movement mechanics. Addressing these discrepancies is critical for reliable real-world deployment.

Accordingly, this study proposes a 1D-CNN–based domain adaptation framework for real-time exercise intensity classification using smartwatch-derived multimodal signals and systematically evaluates cross-subject generalization. A subject-wise hold-out design quantifies inter-individual domain shift, and a lightweight test-time batch normalization (BN) recalibration strategy mitigates distribution mismatch. The findings delineate practical limitations of smartwatch-based intensity prediction (Figure 1) and emphasize the necessity of adaptive modeling strategies for sports performance monitoring and rehabilitation applications.



[Fig. 1] Proposed framework for cross-subject exercise intensity prediction and domain adaptation.

2. Methods

2.1 Dataset

This study utilized the publicly available "Wrist PPG during Exercise" dataset, originally described by Jarchi and Casson (2017) [13]. The dataset includes synchronized wrist photoplethysmography (PPG), chest electrocardiography (ECG), and inertial sensor recordings collected during aerobic exercise.

Eight healthy adults (3 males, 5 females; age range 22–32 years) performed treadmill walking, treadmill running, low-resistance cycling, and high-resistance cycling tasks. Each activity lasted up to 10 minutes, resulting in a total of 19 exercise recordings. All sessions began from rest, allowing heart rate to increase progressively during exercise.

Wrist PPG and inertial signals were recorded at 256 Hz using a wearable sensor system. Simultaneously, single-channel chest ECG was recorded to provide a gold-standard reference for heart rate. Manually annotated ECG R-peak timings are included in the dataset to enable accurate heart rate derivation.

2.2 Preprocessing

Raw wrist-worn multimodal signals were segmented into fixed-length samples using an 8-second sliding window with a 2-second stride at 256 Hz. Model inputs consisted of wrist PPG and three-axis inertial signals (gyroscope and accelerometer), whereas ECG and auxiliary timing signals were excluded from the input features.

Windows containing missing or non-finite values in any selected channel were discarded to ensure signal integrity. Reference heart rate was derived from ECG R–R intervals, and exercise intensity for each window was defined using the heart rate at the window midpoint. Intensity was expressed as a percentage of subject-specific peak heart rate (%HR_peak), defined as the maximum ECG-derived heart rate observed across all available records for each participant. Three intensity categories were defined as <60%, 60–79%, and ≥80%HR_peak.

All input channels were standardized using z-score normalization based exclusively on training-set statistics. The same normalization parameters were subsequently applied to validation and test sets to prevent data leakage. A subject-wise hold-out strategy was implemented during preprocessing, reserving one participant as an independent test domain while the remaining participants were used for model development.

2.3 1D-CNN

A lightweight 1D-CNN was developed to classify exercise intensity from smartwatch-derived multichannel time-series signals. Each input segment was structured as channels × time.

The network comprised three sequential convolutional blocks with 32, 64, and 128 filters, respectively. Each block included 1D convolution, BN, and ReLU activation. Strided convolutions (stride = 2) were applied to progressively reduce temporal resolution while preserving discriminative temporal features associated with cardiovascular and movement responses.

To enhance channel-wise feature recalibration under motion-contaminated wearable signals, squeeze-and-excitation (SE) attention modules were optionally incorporated after each convolutional block. These modules applied global temporal pooling followed by a gating mechanism to adaptively weight informative channels. This design is particularly relevant in wearable-based exercise monitoring, where signal quality may vary across physiological and inertial modalities.

A global average pooling layer aggregated the final feature maps into a compact representation, which was passed through a fully connected layer to generate class logits for the three exercise intensity categories.

2.4 Model training and evaluation

To assess real-world applicability, a subject-wise hold-out strategy was employed. Training and validation sets consisted of data from a subset of participants, whereas the test set contained data from entirely unseen individuals. No subject overlap occurred between training/validation and test sets.

This cross-subject design was intended to evaluate inter-individual generalization, reflecting practical deployment scenarios in sports and rehabilitation settings where models must be applied to new users without retraining.

Performance was evaluated using (1) macro-averaged F1 score, (2) balanced accuracy, and (3) per-class precision, recall, and F1-score. Macro-F1 was selected as the primary evaluation metric because it assigns equal weight to each class, making it appropriate for imbalanced exercise intensity distributions. Balanced accuracy was additionally reported to provide a prevalence-insensitive performance estimate. Validation metrics were used for model selection, and final performance was reported on the held-out test set to quantify cross-subject generalization.

2.5 Domain Adaptation

To further investigate robustness under inter-individual domain shift, an unsupervised test-time BN adaptation procedure was applied.

After loading the best validation model, all network parameters were frozen. The model was then exposed to unlabeled test data in training mode solely to update BN running mean and variance statistics. No gradient-based weight updates were performed, and convolutional parameters remained unchanged.

This lightweight adaptation recalibrates internal feature normalization to the target subject distribution, thereby compensating for inter-individual differences in signal amplitude, cardiovascular response patterns, and motion characteristics. Following BN statistic updates, the adapted model was re-evaluated on the same held-out test set using the predefined evaluation metrics.

2.6 Implementation Details

All experiments were implemented in Python 3.10.9 using PyTorch 2.0.1+cu117 on a Linux-based system (Linux 5.4.0-216-generic, x86_64). Training was conducted on NVIDIA RTX A5000 GPUs with CUDA 11.7 and cuDNN 8.5.0.

The input multimodal signals were sampled at 256 Hz and segmented into fixed-length windows of 8 seconds with a stride of 2 seconds. Each segment was reshaped into a channel-first format to match the input requirements of 1D convolutional operations. A lightweight 1D convolutional neural network was employed for exercise intensity classification. To address class imbalance, both a weighted sampling strategy and a class-weighted loss function were applied during training. Model selection was performed based on validation macro-F1 score, and the best-performing model was retained for final evaluation.

For cross-subject robustness, an unsupervised test-time BN adaptation strategy was applied [14]. During this process, only BN running statistics were updated using unlabeled test data, while all trainable parameters remained fixed.

The detailed model architecture and training configuration are summarized in Table 1.

〈Table 1〉 Model Architecture and Training Configuration

Category	Item	Setting
Model	Architecture	1D-CNN (3 blocks)
	Filters	32, 64, 128
	Kernel sizes	7, 5, 5
	Stride	2
	Normalization	BatchNorm1d
	Activation	ReLU
	Attention	SE attention
Training	Optimizer	AdamW
	Learning rate	1×10^{-3}
	Weight decay	1×10^{-4}

	Batch size	32
	Epochs	300
	Loss function	Focal loss (γ = 1.5)
Imbalance handling	Sampling	WeightedRandomSampler
	Loss weighting	Class-weighted focal loss
Model selection	Criterion	Validation macro-F1
Domain adaptation	Method	Test-time BN adaptation
	Batch size	64
	Adaptation passes	1

3. Results

3.1 Model Performance

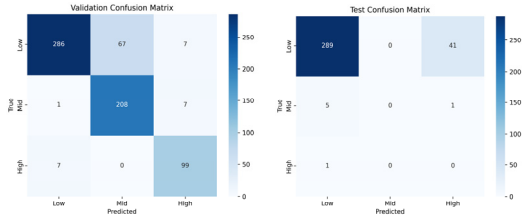
On the validation set, the model achieved a macro-F1 of 0.88 and an overall accuracy of 0.87. As shown in Table 2, recall was highest for the moderate- and high-intensity categories (0.96 and 0.93, respectively), whereas the low-intensity class showed comparatively lower recall (0.79), primarily due to misclassification into the moderate-intensity category.

〈Table 2〉 Validation Set Classification Performance

Class	Precision	Recall	F1-score
〈60%HR_peak	0.97	0.79	0.87
60~79%HR_peak	0.76	0.96	0.85
≥80%HR_peak	0.88	0.93	0.90
Macro Average	0.87	0.90	0.88

However, when evaluated on the subject-wise hold-out dataset, model performance decreased markedly. The macro-F1 score dropped to 0.31 despite an overall accuracy of 0.86. Class-wise analysis revealed that the model predominantly predicted the low-intensity category, resulting in near-zero recall and F1-scores for the moderate- and high-intensity classes. Although classification of low-intensity windows remained relatively preserved (recall = 0.88), the model failed to generalize to higher intensity levels in the unseen participant. This pronounced degradation indicates limited cross-subject generalization and suggests that inter-individual variability in physiological response patterns substantially affects model robustness under domain shift conditions.

Figure 2 illustrates the performance discrepancy between the validation set and the baseline cross-subject hold-out test set prior to domain adaptation.



[Fig. 2] Confusion matrices illustrating the classification performance of the proposed 1D-CNN model on the validation set (left) and the subject-wise hold-out test set (right).

3.2 Performance of Domain Adaptation

To address the marked performance degradation observed under cross-subject evaluation, a test-time BN adaptation strategy was applied. As shown in Table 3, application of BN adaptation increased the macro-F1 score from 0.31 to 0.35 in the cross-subject hold-out setting.

Class-wise performance indicated partial recovery of the moderate-intensity category, with recall improving from 0.00 to 0.33. Although overall accuracy decreased slightly, the increase in macro-F1 reflects a more balanced distribution of classification performance across intensity categories. These findings suggest that recalibration of BN statistics partially mitigates inter-individual domain discrepancies, although substantial variability across subjects remains.

〈Table 3〉 Effect of Domain Adaptation on Cross-Subject Hold-Out Performance

Metric	Baseline	BN-Adaptation
Macro-F1	0.31	0.35
〈60%HR_peak	0.92	0.90
60~79%HR_peak	0.00	0.14
≥80%HR_peak	0.00	0.00

4. Discussion

The present study investigated real-time exercise intensity classification using smartwatch-derived multimodal signals and evaluated cross-subject generalization under a subject-wise hold-out design. While the proposed 1D-CNN achieved strong validation performance (macro-F1 = 0.88), performance declined markedly when applied to an unseen participant (macro-F1 = 0.31). This discrepancy highlights the gap between within-cohort evaluation and real-world deployment, where models must operate across individuals with heterogeneous physiological and biomechanical characteristics.

The degradation under cross-subject evaluation indicates that inter-individual variability critically affects wearable-based intensity prediction. Variations in heart rate dynamics, workload tolerance, sensor placement, skin properties, and movement patterns may shift signal distributions beyond those observed during training [15,16]. Although overall accuracy remained high, macro-F1 revealed severe class imbalance, with predominant prediction of low-intensity segments. This underscores the importance of class-balanced metrics in clinical and sports applications, where accurate detection of moderate and high intensities is essential.

To mitigate this cross-subject performance degradation, an unsupervised test-time BN adaptation strategy was applied. BN recalibration improved macro-F1 from 0.31 to 0.35 and enhanced moderate-intensity detection, indicating that distributional misalignment contributes substantially to performance degradation. However, the modest improvement suggests that more advanced domain adaptation or personalization strategies are needed for robust cross-subject generalization.

It should be noted that the dataset used in this study consists of a relatively small number of participants (n = 8) with a restricted age range (22–32 years), which limits the diversity of physiological and behavioral patterns represented in the data. Therefore, the findings of this study should be interpreted as evidence of structural challenges in cross-subject generalization rather than as population-level generalizable conclusions. Despite this limitation, the substantial performance gap observed between validation and cross-subject evaluation suggests that domain shift is not merely a dataset-specific artifact, but a fundamental issue inherent to wearable-based exercise intensity prediction.

These findings emphasize that wearable-based exercise monitoring systems must account for individual physiological variability. Future work should investigate subject-aware normalization, meta-learning, or hybrid calibration approaches, and validate the proposed framework on larger and more diverse populations to enhance scalability and robustness.

5. Conclusion

In summary, this study demonstrates that although smartwatch-based 1D-CNN models can accurately classify exercise intensity within a development cohort, substantial performance degradation occurs under cross-subject evaluation. This decline reflects inter-individual domain shift in physiological and motion signal characteristics. While test-time BN adaptation partially mitigated this effect, robust deployment in real-world sports and rehabilitation settings will require adaptive or personalized modeling strategies. These findings emphasize the necessity of evaluating wearable-based exercise intensity models under subject-wise hold-out conditions to ensure practical generalizability.

REFERENCES

[1] D.A.J.MD and D.E.T.Kolli, "The future of sports training: Integrating artificial intelligence and wearable technology in performance enhancement," TPM-Testing, Psychometrics, Methodology in Applied Psychology, Vol.32, No.S2, pp.2145-2153, 2025.

[2] A.K.Chowdhury, D.Tjondronegoro, V.Chandran, J.Zhang, and S.G.Trost, "Prediction of relative physical activity intensity using multimodal sensing of physiological data," Sensors, Vol.19, No.20, pp.4509, 2019.

[3] D.Stromback, S.Huang, and V.Radu, "MM-Fit: Multimodal deep learning for automatic exercise logging across sensing devices," Proc. ACM Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), Vol.4, No.4, pp.1-22, 2020.

[4] H.Zhao, Y.Xu, Y.Wu, Z.Ma, Z.Ding, and Y.Sun, "Modeling of the rating of perceived exertion based on heart rate using machine learning methods," Anais da Academia Brasileira de Ciencias, Vol.95, No.2, pp.e20201723, 2023.

[5] J.L.Taylor, A.R.Bonikowske, and T.P.Olson, "Optimizing outcomes in cardiac rehabilitation: The importance of exercise intensity," Frontiers in Cardiovascular Medicine, Vol.8, pp.734278, 2021.

[6] D.S.Couto, I.Lopes, M.I.Oliveira, C.Schmidt, S.Magalhaes, H.Dores, et al., "Exercise intensity prescription in heart failure: A comparison of different physiological parameters," Revista Portuguesa de Cardiologia, Vol.44, No.6, pp.361-371, 2025.

[7] S.Mao and E.Sejdić, "A review of recurrent neural network-based methods in computational physiology," IEEE Transactions on Neural Networks and Learning Systems, Vol.34, No.10, pp.6983-7003, 2022.

[8] A.O.Ige and M.Sibiya, "State-of-the-art in 1D convolutional neural networks: A survey," IEEE Access, Vol.12, pp.144082-144105, 2024.

[9] D.R.Seshadri et al., "Wearable Sensors for Monitoring the Internal and External Workload of the Athlete," NPJ Digital Medicine, Vol.2, No.1, pp.71, 2019.

[10] M.D.Stephenson et al., "Applying Heart Rate Variability to Monitor Health and Performance in Tactical Personnel," IJERPH, Vol.18, No.15, pp.8143, 2021.

[11] Z.Wang et al., "Generalizing to Unseen Domains: A Survey on Domain Generalization," IEEE TPAMI, 2022.

[12] T.Bao, S.A.R.Zaidi, S.Xie, P.Yang, and Z.Q.Zhang, "Inter-subject domain adaptation for CNN-based wrist kinematics estimation using sEMG," IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol.29, pp.1068-1078, 2021.

[13] D.Jarchi and A.J.Casson, "Description of a database containing wrist PPG signals recorded during physical exercise with both accelerometer and gyroscope measures of motion," Data, Vol.2, No.1, pp.1, 2016.

[14] D.Wang et al., "Tent: Fully Test-Time Adaptation by Entropy Minimization," Proc. ICLR, 2021.

[15] Seshadri, D. R., Li, R. T., Voos, J. E., Rowbottom, J. R., Alfes, C. M., Zorman, C. A., & Drummond, C. K. (2019). Wearable sensors for monitoring the internal and external workload of the athlete. NPJ Digital Medicine, 2(1), 71.

[16] Stephenson, M. D., Thompson, A. G., Merrigan, J. J., Stone, J. D., & Hagen, J. A. (2021). Applying heart rate variability to monitor health and performance in tactical personnel: A narrative review. International Journal of Environmental Research and Public Health, 18(15), 8143.

이 재 혁(Jae-Hyuk Lee) [정회원]

- 2020년 2월 : 고려대학교 보건환경융합과학부 (이학박사)
- 2023년 9월 ~ 2025년 2월 : 공주대학교 스마트기술연구소 연구교수
- 2025년 3월 ~ 현재 : 서울과학기술대학교 IT융합기술연구소 연구교수

〈관심분야〉
디지털 헬스, 신호처리, 빅데이터, 인공지능