SoundTrack: A Contactless Mobile Solution for Real-time Running Metric Estimation for Treadmill Running in the Wild

JINGPING NIE, Columbia University, United States
YUANG FAN, Columbia University, United States
ZIYI XUAN, Columbia University, United States
MINGHUI ZHAO, Columbia University, United States
RUNXI WAN, Tenafly High School, United States
MATTHIAS PREINDL, Columbia University, United States
XIAOFAN JIANG, Columbia University, United States

Running metrics like cadence and ground contact time (GCT) are crucial for both novice and experienced runners to optimize performance and prevent injuries. We present SoundTrack, a contactless mobile solution that estimates these metrics by analyzing treadmill running sounds using on-device machine learning. Our main contributions are: (*i*) SoundTrackDB – a comprehensive 40-hour dataset of treadmill running sounds collected from 61 subjects across 363 sessions in 13 public gyms, created in collaboration with a licensed running coach; and (*ii*) SoundTrack – an on-device mobile system capturing treadmill running sounds, mitigating noise, estimating cadence and GCT with a custom multi-layer perceptron (MLP) model, and providing real-time feedback. Microbenchmarks and evaluations show that SoundTrack effectively mitigates real-world noise challenges in public gyms and adapts to individual variations among runners and treadmill models. It achieves mean absolute percentage errors (MAPEs) of 1.62% for cadence and 6.05% for GCT on the test set of unseen running sessions, yielding results that are superior or comparable to commercial sports wearables. SoundTrack offers an accessible solution for treadmill metrics on mobile platforms, reducing reliance on specialized wearables and broadening accessibility. SoundTrackDB, SoundTrack, and the demonstration video are available at: https://github.com/Columbia-ICSL/SoundTrackDB.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → *Health informatics*; • **Computing methodologies** → **Machine learning approaches**.

Additional Key Words and Phrases: Acoustic Sensing, Ubiquitous Computing, Fitness Tracking, Mobile Health, Edge AI, Running Metrics, Sports Science

ACM Reference Format:
Jingping Nie, Yuang Fan, Ziyi Xuan, Minghui Zhao, Runxi Wan, Matthias Preindl, and Xiaofan Jiang. 2025. SoundTrack: A Contactless Mobile Solution for Real-time Running Metric Estimation for Treadmill Running in the Wild. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 2, Article 42 (June 2025), 30 pages. https://doi.org/10.1145/3729486

Authors' Contact Information: Jingping Nie, Columbia University, New York, United States, jn2551@columbia.edu; Yuang Fan, Columbia University, New York, United States, yf2676@columbia.edu; Ziyi Xuan, Columbia University, New York, United States, zx2420@columbia.edu; Minghui Zhao, Columbia University, New York, United States, mz2866@columbia.edu; Runxi Wan, Tenafly High School, New Jersey, United States, wanrunxi838@gmail.com; Matthias Preindl, Columbia University, New York, United States, matthias.preindl@columbia.edu; Xiaofan Jiang, Columbia University, New York, United States, jiang@ee.columbia.edu.

ACM 2474-9567/2025/6-ART42
https://doi.org/10.1145/3729486

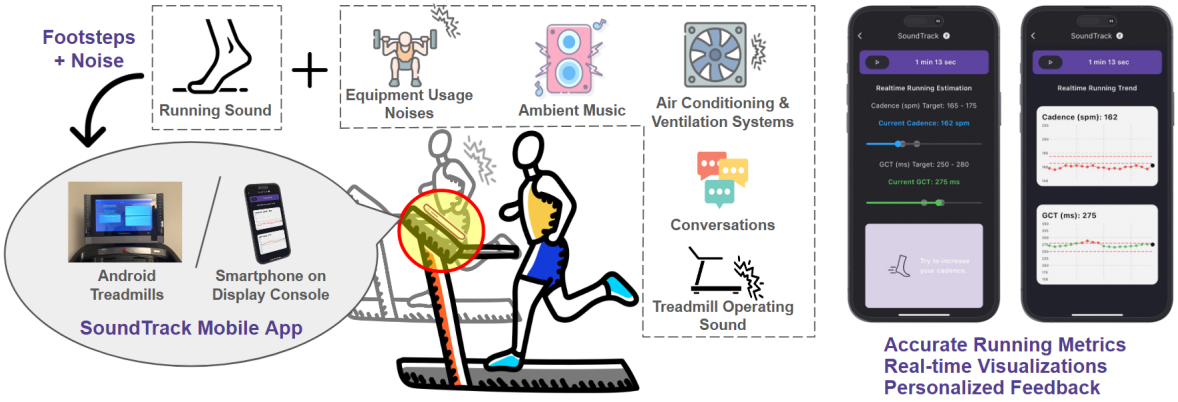


Fig. 1. SoundTrack: A contactless solution that uses mobile platforms to accurately estimate cadence and ground contact time (GCT) during treadmill running with on-device machine learning, suitable for runners of all levels and compatible with various treadmill models in gyms with different environmental noise, without needing user-specific calibration. It provides real-time visualizations of running metrics and personalized feedback and suggestions.

1 Introduction

Running is a widely popular exercise, with approximately 620 million runners globally and a growing number of distance runners and race participants [31, 36, 92]. Despite its health benefits, approximately 50% of runners experience injuries each year [85]. Measuring running metrics is essential for improving performance and preventing injuries across all runner levels. Studies show that elite distance runners with shorter ground contact time (GCT) have lower metabolic costs and better efficiency [17, 27], and increasing cadence (steps per minute) can correct overstriding and other injury-prone habits [8, 40].

Treadmill running is popular for its convenience and ability to be done in any weather. While most treadmills provide basic data on distance, incline, and speed, they rarely offer detailed insights into individual running-form-related running metrics such as cadence and ground contact time (GCT) [91]. Advanced facilities, such as professional laboratories equipped with motion capture systems and wearable sensor networks, are often required to measure these metrics precisely. Additionally, recent methods utilizing novel sensing technologies—such as millimeter-wave (mmWave) and Ultra-Wideband (UWB) radar, inertial measurement units (IMUs), and pressure sensors—have been proposed to assess running form. However, these setups remain largely inaccessible and impractical for everyday use due to their cost, complexity, and the specialized environment they demand, making them unrealistic to be used on a day-to-day basis [20, 23, 30, 45, 52, 66].

On the other hand, more accessible metrics tracking products used by general runners typically include wearable devices like sports watches, foot pods, and smart rings, with prices ranging from approximately $200 to over $1,000 [55, 112]. However, sports watches have significant limitations indoors and often fail to estimate advanced metrics such as GCT due to their reliance on GPS and localization technologies [42, 83, 104]. IMU-based foot pods can provide more reliable and extensive metrics (e.g., cadence, GCT, footstrike types, and impact force) during treadmill running but usually require extensive calibration and are costly [83, 104]. Additionally, variations in running form, experience, and treadmill models complicate accurate metric estimation [111].

While only about 30% of Americans use a wearable device like a smartwatch or fitness tracker [71], nearly 90% own a smartphone [74]. There are several smartphone-based running apps on the market, including Strava, Nike Run Club, and Runkeeper. These apps require users to hold or wear the phone, which can affect running

form, and typically don't provide advanced metrics like GCT standalone [48, 89]. Although treadmill-runners usually bring their phones for workouts, the phones tend to be put on the treadmill's display console or in the cup holder [7].

While current running apps treat this placement as a limitation, it actually creates an ideal setup for capturing an often-overlooked source of running information: sound. Researchers have demonstrated the versatility of smartphone microphones in various sensing applications, from acoustic thermometers [14] to touch interfaces on solid surfaces [49]. In the running context, licensed running coaches note that different sounds are produced when running on treadmills due to variations in height, weight, footwear, speeds, and running forms. Indeed, different running forms can create distinct sounds due to variations in footstrike patterns, cadence, and running mechanics. For example, a higher cadence usually results in lighter sounds, while runners with a longer stride may produce louder sounds due to increased impact force [97].

These acoustic signatures suggest an exciting opportunity: leveraging the smartphones that runners already bring to their workouts to capture and analyze running metrics through sound. Our vision is to create a contactless mobile platform that can understand these treadmill running sounds, *accurately estimate running metrics in real time*—in particular, cadence and GCT—and provide *real-time visualization and feedback*. This system should require *no additional devices* other than the smartphone or tablets that users already have, be *easily accessible* to runners with different levels of expertise, and be *privacy-aware*. In addition, this system is *versatile* enough to be used on different treadmills in various gym or indoor environments with diverse background noises and environmental conditions.

Realizing such a system poses several challenges and requirements:

(1) Data Availability: There exists no public dataset containing treadmill running sounds with associated running metrics across different expertise levels.

(2) Ground Truth Collection: Existing devices are suboptimal for ground truth collection—smartwatches suffer from low indoor accuracy and coarse granularity, while Bluetooth-based running pods face connectivity issues and data loss. This necessitates designing a portable and robust data collection setup.

(3) Temporal Precision: Ground Contact Time occurs at millisecond-level precision, requiring precise synchronization between our platform and ground-truth collection devices.

(4) Experiment Design: The data collection procedure must maximize valid data across different speeds and participant backgrounds while ensuring participant safety and managing fatigue.

(5) Data Processing: The system must process and analyze massive amounts of multi-device data to generate precise ground truth labels.

(6) Mobile and Privacy-aware Implementation: The final system must ensure privacy by running locally on resource-constrained mobile platforms while maintaining real-time processing capabilities and achieving performance comparable to specialized devices.

In this work, we present SoundTrack, a mobile-platform-based contactless solution that accurately estimates runners' cadence and GCT through sound, as shown in Figure 1. Working with a licensed running coach, we address the aforementioned challenges through several key developments. First, we design a robust data collection setup using wearable devices and mobile platforms to create SoundTrackDB—an *open-source* comprehensive treadmill-running audio dataset from 61 valid subjects (out of 70 recruited). This dataset includes ground-truth cadence and GCT labels generated through our custom computer vision-based labeling method, baseline metrics from professional wearable devices, and detailed subject background information. Building on this foundation, we develop SoundTrack's on-device machine learning pipeline that processes audio in real time *without requiring user calibration*, providing immediate running metrics and actionable feedback. Compared to traditional wearable technologies, our solution offers a more accessible, comfortable, and less intrusive approach for runners seeking to enhance their performance and mitigate injury risks.

We summarize our contributions as follows:

- **System and Algorithm Design:** We design and develop SoundTrack, a privacy-aware mobile application that leverages acoustic sensing to accurately estimate running metrics. The system processes treadmill running sounds locally, mitigates real-world noise, extracts acoustic features, and uses machine learning to estimate cadence and GCT, delivering in-situ instantaneous running metrics and real-time feedback.
- **Novel Dataset with Efficient and Accurate Ground Truth Annotation**: We create SoundTrackDB, a first-of-its-kind treadmill running dataset collected from 61 subjects with diverse demographics. We design a portable and robust data collection setup and experiment procedure in collaboration with a licensed running coach. The dataset includes synchronized audio recordings, ground-truth running metrics, and comparative data produced by smart watches (Apple/Garmin/Coros), and professional running wearables (RunScribe). We develop a semi-automated annotation method using YOLO-V8 to efficiently and accurately label running cadence and GCT ground truth by detecting foot-treadmill contact from video and validating it against human-labeled samples. We open-source SoundTrackDB to the public.
- **Real-world Prototype and Validation:** We implement the complete SoundTrack pipeline to run locally on mobile platforms, achieving low computational complexity while being privacy-aware without requiring user-specific calibration. Our system achieves a mean absolute percentage error (MAPE) of 1.62% for cadence and 6.05% for GCT, delivering results superior to those of sports watches for cadence estimation and comparable to or better than professional devices, such as the RunScribe running pods (priced around $500), in cadence and GCT estimation. Through case studies across 3 new locations with 11 previously unseen subjects, we demonstrate SoundTrack's robustness with MAPE 1.74% and 7.21% for cadence and GCT, respectively.

To the best of our knowledge, SoundTrackDB is the *first* audio dataset of treadmill running metrics collected from 61 subjects (22 females, 39 males, ages 18–55) running at six different paces (5–10 mph), across 363 treadmill sessions (150s to 50 minutes each), totaling 40 hours of data. Data was collected from 24 treadmills (12 models) in 13 public gyms in real-world settings. We **open-source SoundTrackDB and release SoundTrack as a cross-platform mobile app supporting iPhones, Android Phones, and Android Treadmills** to democratize access to professional-grade running metrics[1]. We envision this work can (*i*) advance research in related fields, such as sports science and ML-centered fitness and health; (*ii*) enable high-accuracy GCT and cadence estimations to everyday runners without additional hardware investment while offering a level of precision typically available only through specialized devices costing $500 or more; (*iii*) foster the development of accessible tools for improving running form and preventing injuries across diverse communities.

2 Background and Related Work

2.1 Running Metrics Background

Runners' preferences and experience levels can affect running metrics like cadence, GCT, strike types, pronation, and stride length [57]. Cadence, measured in steps per minute (spm), is the number of steps per minute. GCT, usually measured in milliseconds (ms), is the duration a runner's foot stays on the ground. Foot strike types include heel, midfoot, and forefoot strike. Running metrics are essential for evaluating efficiency and performance [22, 68]. Beginners focus on cadence to establish a step count, while advanced runners optimize these metrics for training goals [47]. Proper attention to running metrics also helps prevent injuries [41, 67, 68, 86]. For example, a higher cadence can reduce impact forces, vertical loading rates, and injury risk [41, 67, 86], and a shorter GCT allows for quicker energy transfer through muscles and tendons, reducing energy expenditure and muscular effort, helping maintain higher speeds and improves running economy, key for long-distance performance [17, 27]. According to our collaborating running coach, changing foot strike type while running is possible but requires a gradual

[1] https://github.com/Columbia-ICSL/SoundTrackDB

approach and proper technique to avoid injuries. As such, SoundTrack focuses on estimating cadence and GCT based on the foot strike sound during treadmill running.

2.2 Running-Form-based Running Metrics Estimation Methods

2.2.1 Fixed-System-Based Setup. Combinations of image processing techniques with cameras, floor-based pressure and force sensors, floor-based vibration sensors, or/and wearable/human-carried motion sensors are usually used for gait analysis or exercise monitoring, including strike type, cadence, and GCT, in lab settings [20, 23, 30, 45, 52, 66, 77, 115]. For example, [75] and [73] use Vicon 3D motion capture system, based on the use of multiple high-speed cameras and reflective markers attached to key points on the subject's body, to measure joint kinematics of treadmill running at 3.0, 4.5, and 5.5 *mph* and around 6.55 ± 0.78 *mph* [73, 75]. A few works use computer-vision-based approaches, such as YOLOv8 and MediaPipe, for pose estimation [79, 113]. Floor-based pressure sensors are used to detect foot strike type and GCT during walking [53, 99]. Combinations of these approaches are also used, such as the M3D gait analysis system with motion sensors and force plates [62], and IDIoT, which fuses motion sensor data with camera footage [6]. While these systems provide comprehensive and precise results, they are less portable and more expensive than wearable alternatives, and camera-based methods can be privacy intrusive.

mmWave and UWB radar are increasingly explored sensing modalities for human activity recognition (HAR), gait analysis, and exercise monitoring [15, 61, 64, 100, 101]. While much of the existing research on exercise metrics focuses on anaerobic activities like strength training, there is limited attention on continuous posture and performance monitoring for aerobic exercises, especially running [18, 114]. These aforementioned fixed-system-based methods are not portable, are inconvenient, and have excessive overhead for everyday runners or gym goers to use in public gyms.

2.2.2 Portable Personal Wearable Device and Sensor Networks. Wearable devices provide a portable alternative for monitoring running metrics and have consistently been among the top three fitness trends since 2016 [5, 26]. Key commodity commercial-off-the-shelf (COTS) options include smart sports watches (e.g., Garmin [34], Coros [21], Apple [3]), foot pods (e.g., RunScribe [84], Stryd [95]), smart rings, and smart insoles based on flexible thin film pressure sensors [63]. Sports watches, commonly used daily and in races, use IMU and GPS technologies to track metrics like pace, distance, and elevation [78]. However, their accuracy is limited indoors. In the literature, new sensing form factors on wearables are investigated for exercise monitoring or posture tracking [50, 93, 116]. For example, proximity magnetic sensing is utilized for exercise monitoring via a single wrist-worn wearable or wall-mounted smartphone, while full-body poses are inferred from body silhouettes using a miniature camera on a wristband [50, 59].

Manufacturers and researchers are actively researching combining wearable devices or using sensor networks for better results in exercise monitoring [38, 94, 110]. [38] leveraged wrist-worn wearables and arm-mounted smartphones to assess dynamic postures in workouts. ER-rhythm was proposed to monitor exercise and respiration rhythm for Locomotor Respiratory Coupling (LRC) estimation using COTS RFIDs attached to the human body [110]. Numerous research works also use single or multiple IMUs from smartphones, mounted on shoes, or attached to the body to assess gait quality and running movement during running [10, 25, 32, 90]. However, wearables might be intrusive and uncomfortable, cause imbalance, influence running form, and increase the risk of injury. Additionally, there may be a need to purchase extra devices.

2.2.3 Audio-based. Audio-based approaches are widely used for fitness and exercise monitoring [54, 102, 107]. The on-device solution that allows user data and models to be stored and used locally is one of the most popular ways to protect users' privacy [33, 87, 96]. Hou et al. proposed a passive acoustic sensing method with smartphones and headphones to detect rope-jumping and breathing sounds [43]. Smartphones and earables are employed for

monitoring exercise intensity through captured breathing sounds, breathing modes during running, and running rhythm for LRC estimation [39, 44, 81]. Smart speakers are used for activity detection and fitness type monitoring via acoustic sensing [54, 107].

While the majority of acoustics-enabled fitness and workout trackers focus on activity type classification and vital sign monitoring, a few of them explore treadmill running sounds, where the treadmill operation noises are nontrivial. The step sounds captured through audio tracks provide detailed information on running variations [58, 60], offering insights into a runner's gait, including foot strike patterns and the loading intensity of footfalls. Heel strikers tend to produce a louder, more forceful impact sound compared to midfoot and forefoot strikers due to the significant collision with the treadmill belt [4, 97]. Although intuitive, the sound of treadmill running is highly correlated with the user's running form and metrics. However, decoding it is challenging due to high variations among runners, treadmill operation sounds, and environmental noise. [108] estimates running cadence using treadmill sounds with an LSTM-based model. However, as benchmarked in Section 6, its generalizability in public gyms with diverse environmental noise is limited, as data was collected from a single treadmill model in one gym with a small number of subjects. [72] presents preliminary results for estimating running cadence and GCT but does not demonstrate consistent performance in real-world scenarios.

In the aforementioned approaches, most works only include minimal running speed ranges with a restricted number of subjects with limited running experience. These approaches also have issues in portability, price, and comfort. SoundTrack overcomes these limitations with a non-contact, portable, and user-friendly design that leverages mobile platforms treadmill runners may already own. It supports a broad range of speeds suitable for both novice and experienced runners, whether for recreational or training purposes.

2.3 Treadmill Running Dataset

Variations in treadmill design (e.g., cushioning, motor power, incline, sensors) and runner characteristics (e.g., age, body metrics, running style, experience) can significantly affect both the running experience and collected data [13, 19, 76, 103]. These factors introduce challenges in building comprehensive treadmill running datasets and affect accuracy. Existing datasets on treadmill running, such as those incorporating IMUs and strain sensors, focus predominantly on biomechanical metrics [35]. Other datasets like "TRIPOD" offer insights into walking dynamics but do not cover running speeds, only three different walking speeds [98]. SoundTrackDB addresses these limitations by aligning treadmill running sounds with running metrics.

3 Data Collection

We designed a portable, reliable data collection setup in collaboration with a licensed running coach, ensuring reproducibility and adaptability across different treadmills and runner types. Figure 2 outlines the structured three-stage methodology used for data collection in this study, encompassing `Pre-Run`, `In-The-Run`, and `Post-Run` phases. This research protocol received approval from the Institutional Review Board (IRB). We recruit 70 adult subjects from diverse demographic distributions. Participants were recruited through a call posted on the lab website, flyers distributed on campus, and outreach within personal networks to ensure a diverse sample. Informed consent was obtained from all participants before their involvement in the study.

Ground Truth: We opted to use a **`Video Phone`** instead of delicate motion-capture systems or high-speed cameras for its portability. Since the data was collected in public gyms rather than a controlled lab environment, we prioritized low-cost, portable equipment to reduce setup time and minimize disruptions to other gymgoers during the experiments.

Data Collection Procedures: All participants first completed the `Pre-Run` informed consent process, during which they were informed of the data collection, study purpose, and their right to withdraw at any time. `Pre-Run` Questionnaire collects general information and details about their running experience as illustrated in Section 4.2

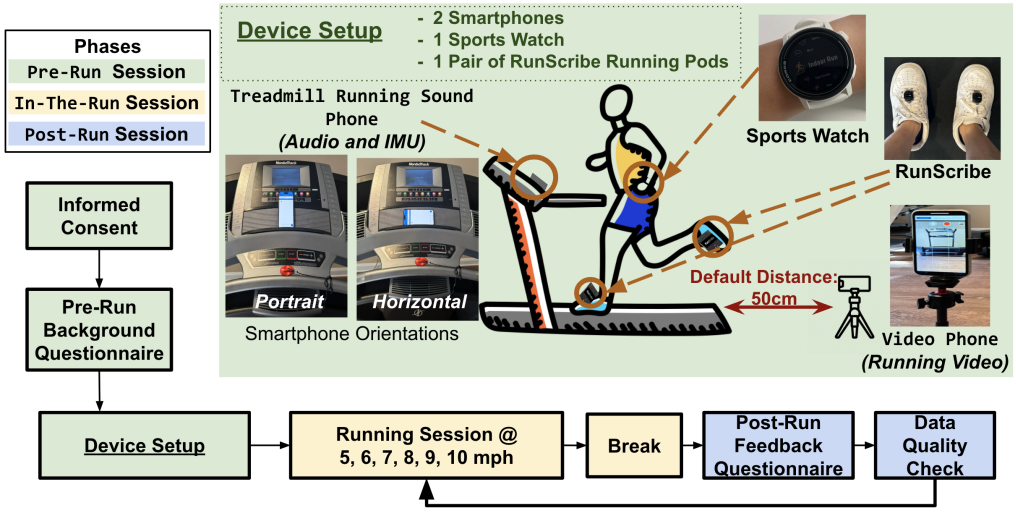


Fig. 2. The data collection process is divided into three phases: Pre-Run, In-the-Run, and Post-Run sessions, capturing treadmill running acoustics (audio and IMU), audio data.

and in Figure 6. The study investigators will set up the devices, including (*i*) a **Treadmill Running Sound Phone**, positioned on the treadmill's display console in either portrait or landscape orientation, used to capture treadmill running acoustics via a custom mobile app (audio sampled at 16,000 Hz; IMU data sampled at 100 Hz); (*ii*) a **Video Phone**, placed 50 centimeters behind the treadmill belt, capturing running video (sampled at 60 fps) to generate ground truth for cadence and GCT. The camera's center of view is aligned with the treadmill belt, ensuring that each foot contact occurs at a consistent vertical height in the video; and (*iii*) a Garmin/Coros/Apple sports watch (sampled at 1 Hz, models listed in Appendix A) and a pair of RunScribe running pods (sampled at 500 Hz, calibrated for each subject before their running session) for baseline comparison. In our experiment, the custom mobile app was implemented on various **Treadmill Running Sound Phone** models, including four different iPhone models, a Pixel 7, and an iPad mini.

When In-The-Run, participants select their running speed from {5, 6, 7, 8, 9, 10} miles per hour (*mph*) based on their personal comfort and ability. The duration of the session is self-regulated, with mandatory rest periods to ensure recovery. A licensed coach monitors the session to maintain safety protocols and ensure the appropriate wear of baseline devices. During Post-Run, participants completed a questionnaire, provided feedback, and reported their perceived exertion level. After data quality checks, they could start a new session or conclude the data collection. As shown in Figure 3 and illustrated in Appendix A, the *Manual Data Syncing* step precisely aligns the video, audio, IMU, and RunScribe data streams using Unix timestamps, an audible cue, and a distinct toe-tapping foot pattern despite inherent timestamp discrepancies.

4 SoundTrackDB Running Metrics Annotation Generation and Validation

As noted in [108], manually labeling video frames for treadmill running is precise but labor-intensive. While [108] used MediaPipe Pose for ground truth generation, it struggled with overlapping feet and lacked testing across diverse runners, treadmills, and environments. To address this, we developed a semi-automated method using YOLO-V8 [80] for accurate and efficient cadence and GCT annotation in SoundTrackDB. Figure 3 outlines the workflow, including *Manual Data Syncing*, *YOLO-based Processing*, and *Manual Thresholding*.

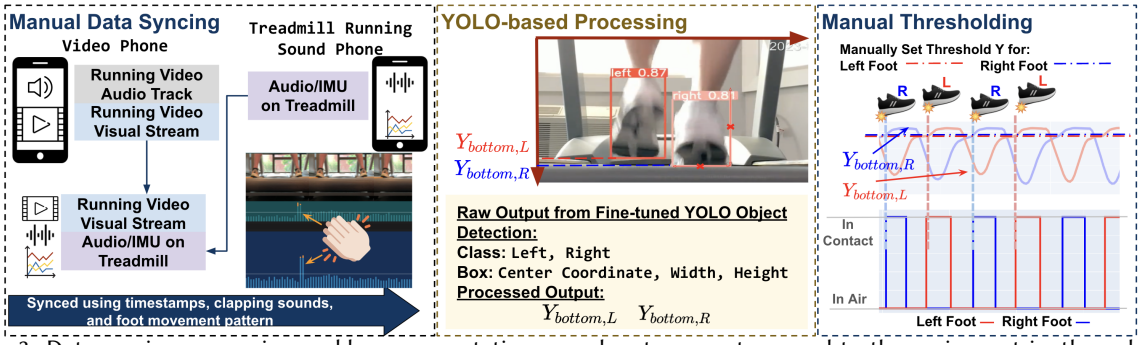


Fig. 3. Data syncing, processing, and human annotation procedure to generate ground-truth running metrics through CV method for SoundTrackDB.

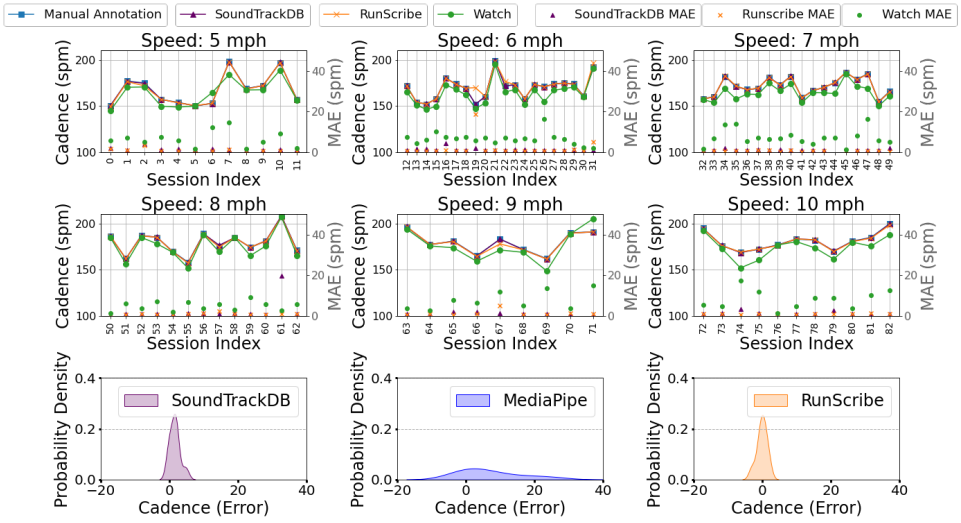


Fig. 4. *Top two rows:* Performance comparison of SoundTrackDB, RunScribe, and a sports watch for cadence estimation relative to manual annotations. *Bottom row:* Distribution of cadence estimation errors for SoundTrackDB, the MediaPipe-based method from [108], and RunScribe.

4.1 YOLO-based Ground-truth Generation Method and Evaluation

As shown in Figure 3, in the *YOLO-based Processing* step, we fine-tuned YOLO-V8 Medium on locally labeled bounding boxes to process synchronized videos, extracting the Y-coordinates of the lowest edges of the left and right shoes ($Y_{bottom,L}$ and $Y_{bottom,R}$). These coordinates, consistent across footstrike types due to the arrangements of **`Video Phone`**, are used to determine ground contact time (GCT) and cadence. GCT is calculated based on threshold lines for each foot in the *Manual Thresholding* step, while cadence is derived from complete contact cycles. This semi-automated process is applied to all running sessions to construct SoundTrackDB (see Section 4.2 for details).

Evaluation: We randomly selected videos from 83 running sessions at varying speeds, recorded from 15 runners. The $Y_{bottom,L,manual}$ and $Y_{bottom,R,manual}$ values were manually labeled by three human annotators, and GCT_{manual} and $Cadence_{manual}$ were derived using the thresholding method from Section 4.1, serving as "reference data" in this evaluation. Figures 4 and 5 (top rows) compare the CV-based method used to create SoundTrackDB,

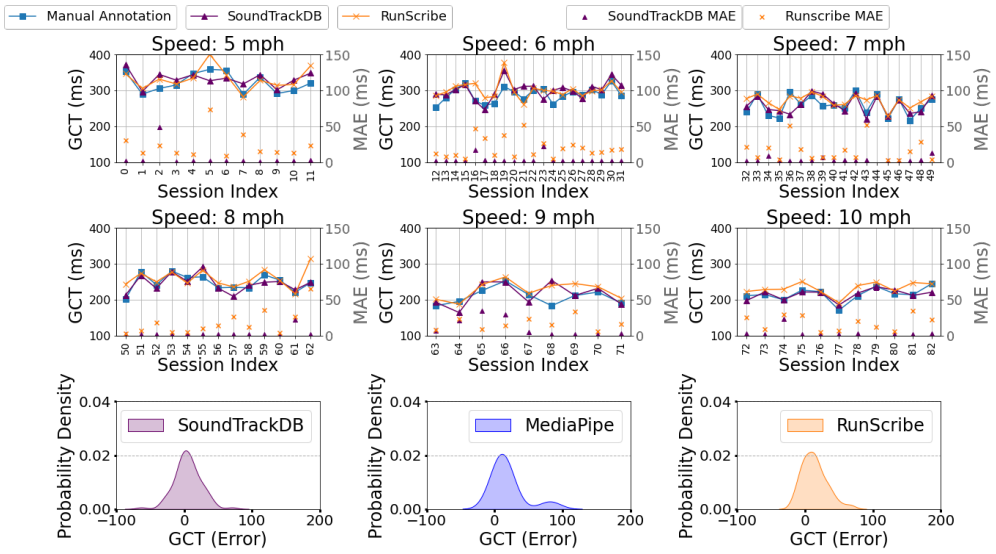


Fig. 5. *Top two rows:* Comparison of GCT estimation performance among SoundTrackDB and RunScribe, benchmarked against manual annotations. *Bottom row:* Distribution of GCT estimation errors for SoundTrackDB, the MediaPipe-based method from [108], and RunScribe.

	Cadence Estimation — Average MAE (rpm)			GCT Estimation — Average MAE (ms)	
Speed	**SoundTrackDB**	**RunScribe**	**Watch**	**SoundTrackDB**	**RunScribe**
5 mph	1.51	1.06	6.47	5.83	23.24
6 mph	1.30	1.77	6.55	3.55	19.72
7 mph	1.11	0.64	7.23	3.38	16.08
8 mph	2.44	0.92	4.81	3.54	17.81
9 mph	1.20	1.29	7.72	11.87	16.48
10 mph	1.39	0.92	8.11	10.00	18.15
Overall	1.47	1.13	5.62	5.62	18.61

Table 1. Average MAE Across All Sessions for Cadence and GCT Estimation at Different Speeds Using Various Approaches

RunScribe, and a sports watch for cadence and GCT estimation across speeds (note: the sports watch does not provide GCT estimation). The x-axis represents the session indices of the 83 sessions. Table 1 presents the Mean Absolute Error (MAE) for cadence and GCT estimation. SoundTrackDB achieves an average MAE of 1.47 *spm* for cadence, outperforming the sports watch (6.75 *spm*) and performing comparably to RunScribe (1.13 *spm*), with slightly higher errors at 8 *mph*. For GCT, SoundTrackDB achieves an MAE of 5.62 *ms*, significantly lower than RunScribe's 18.61 *ms*, particularly at speeds of 6, 7, and 8 *mph*. The bottom rows of Figures 4 and 5 show the error distributions for cadence and GCT using SoundTrackDB, the MediaPipe-based method from [108], and RunScribe. The results indicate that SoundTrackDB consistently achieves narrower error margins and higher accuracy, whereas both RunScribe and the MediaPipe-based method tend to overestimate GCT.

4.2 Description of SoundTrackDB

Figure 6 illustrates SoundTrackDB, which we curated to serve as the basis for ML-model development and evaluation for SoundTrack. Although 70 subjects participated in this study, data from 9 subjects were excluded from SoundTrackDB. 4 subjects quit the study, and there was data loss from 5 subjects. As such, SoundTrackDB includes 40-hour data collected from 61 subjects (22 females, 39 males, ages 18-55) in 363 running sessions across

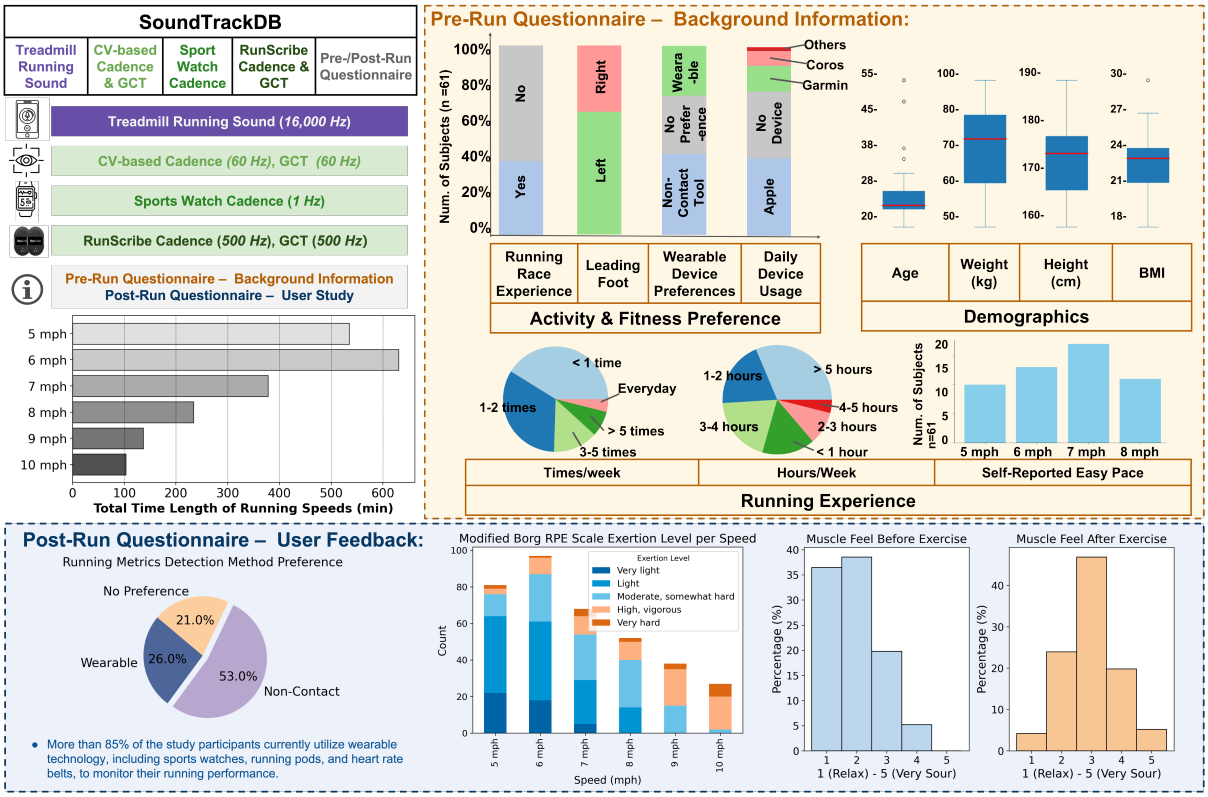


Fig. 6. Overview of SoundTrackDB to be open sourced: treadmill running sound, CV-based, sports watch and RunScribe collected running metrics, background information, and user study.

24 treadmills of 12 different models situated in 13 different public gyms. Demographics (age, weight, height, BMI), running habits (activity frequency, running duration, self-reported easy pace), wearable device preferences (use of Apple, Garmin, Coros, or other devices) and `Pre-Run` preferences for non-contact tools for fitness tracking are collected through the `Pre-Run` questionnaire.

The 13 public gyms have a variety of environmental noises, including background music, sounds from air conditioning systems, noise from other people using gym equipment (such as running or walking on neighboring treadmills, or using strength training machines), and conversations among gym-goers. Different treadmills produce varying levels of operational noise. To assess the noise levels, we randomly selected one session from each public gym and measured the decibel levels using 10-second sound snippets of three conditions: environment sound only (treadmill off), operating sound of empty treadmill (without anyone running on it), and running sound of treadmill with runner running on it. These measurements were collected using the **`Treadmill Running Sound Phone`**, which was placed on the display console following the setup method described in Section 3, as illustrated in Figure 7. The environmental noise levels varied from 38 dB to 69 dB, depending on how busy the gym was at the time of data collection.

However, as shown in Figure 7, once the treadmill operates, noise levels converge to 65–74 dB across all locations, regardless of initial ambient noise. For example, despite higher baseline noise at Location AD, its noise level aligns with quieter sites like NP_A during treadmill use, indicating the treadmill's mechanical noise masks ambient variations and creates a consistent acoustic environment. These facts indicate that the treadmill's

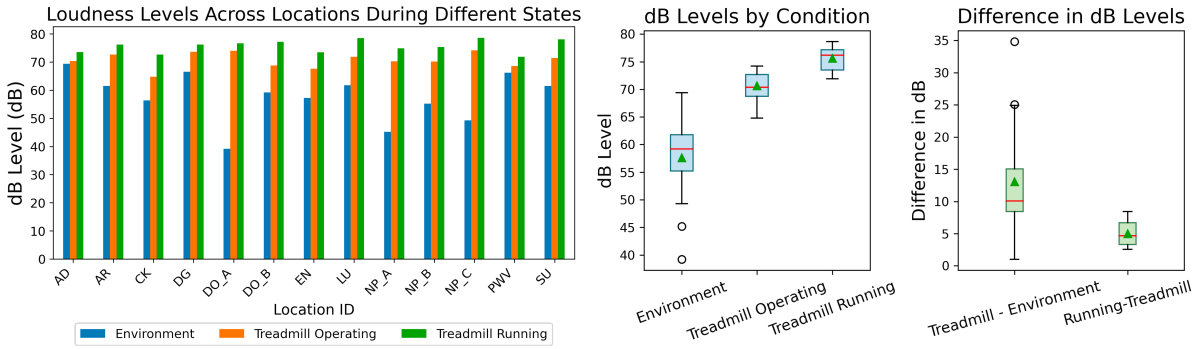


Fig. 7. The example average decibel (dB) levels for the environment (treadmill off), treadmill operating (without people running), and treadmill running sound (a runner running on the operating treadmill) in different locations from randomly selected sessions.

operating sound reduces the influence of ambient noise in the collected sound files. On average, people running on the treadmill add only about 5 dB to the treadmill's operating noise (forming the treadmill running sounds). We observed that the additional acoustic energy from running heavily depends on each subject's weight, running shoe types (e.g., extra-cushioning or carbon-plated), and foot strike force.

SoundTrackDB captures a wide array of running metrics collected by various devices at varying paces from 5 mph to 10 mph, through a rigorous data cleaning and verification process. In particular, SoundTrackDB includes five categories of data: (*i*) the treadmill running sounds captured by the mobile platform placed on the treadmill consoles through the custom smartphone App; (*ii*) cadence and GCT generated by the CV-based semi-automatic ground truth generation method mentioned in Section 4.1; (*iii*) cadence estimated by sports watches (Garmin/Coros/Apple Watch); (*iv*) cadence and GCT captured by RunScribe; and (*v*) background information collected in the `Pre-run` Background Questionnaire, including demographics, personal activity and fitness preference, as well as running experience as shown in Figure 6.

In the `Post-Run` questionnaire, participants were asked to report their perceived exertion level for each running session using a modified five-level Borg RPE (Rating of Perceived Exertion) scale [11]. As shown in Figure 6, the majority of participants reported exertion levels ranging from very light to moderate. After completing the experiment, participants also rated their muscle soreness on a scale of 1 to 5, both before and after the running sessions, with most reporting moderate soreness. Additionally, participants were asked to express their preferences for a *Running Metrics Detection Method* during treadmill running again, choosing between: (*i*) wearable devices, (*ii*) a non-contact method, or (*iii*) no preference. At this phase, participants had the opportunity to try wearable options (e.g., sports watches, RunScribe running pods) and learn about our non-contact system. While over 65% of participants currently use wearable devices to monitor their exercise and running performance, more than 53% expressed a preference for a non-contact alternative and found SoundTrack to be useful, whereas only 26% preferred to continue using wearable devices. Notably, compared to `Pre-Run` preferences for non-contact tools for fitness tracking (39%), there was a 14% increase in the number of participants who preferred non-contact methods, as shown in the `Post-Run` Questionnaire. Some participants commented that "*it's good not to wear anything to get my metrics measured*" (P32) and "it's great to have the smartphone system as an alternative" because "*sometimes I forget to wear my watch, but I always have my phone with me*" (P45).

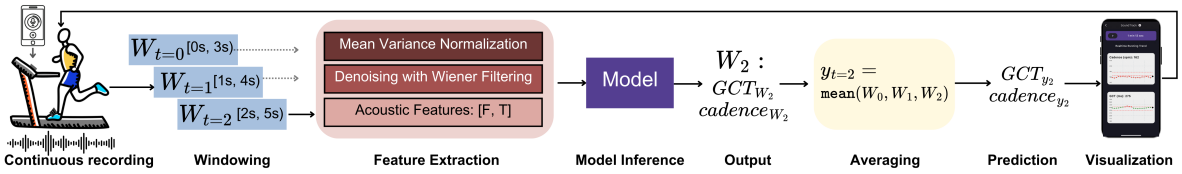


Fig. 8. SoundTrack system architecture.

5 SoundTrack System Architecture

Figure 8 shows the architecture of SoundTrack, designed to operate on a mobile platform placed on a treadmill's control panel, which records, processes, estimates, and visualizes cadence and ground contact time (GCT) from treadmill running sounds on-device. Inspired by [108], SoundTrack uses a sliding window approach with a 3-second window length and a 1-second stride.

To account for changes in environmental noise and variations in treadmill operation sounds, mean-variance normalization (MVN) is applied to each window, ensuring consistent scaling and centering of the data within each segment. In addition, a Wiener filter is applied to further suppress the background noise and improve the quality of the signal of interest (Section 6.1). Each window segment undergoes acoustic feature generation, converting audio data from the time domain into frequency-domain feature matrices, with dimensions $[F, T]$ where F and T are the number of features and time frames, tailored to the implemented model (Section 6.2). The features are input into a multi-task learning (MTL) `Model` designed to estimate $cadence_{W_i}$ and GCT_{W_i} simultaneously for the current window, $W_{t=i}$. This `Model` is chosen for efficiency, enabling real-time processing on mobile devices with less delay and enhanced user experience. By leveraging shared learning from related audio attributes, it achieves results comparable to single-task models (see Section 6.4). The predicted $cadence_{y_i}$ and GCT_{y_i} at $t = i$ are calculated as the average of the current window and the two preceding windows.

SoundTrack dynamically visualizes the latest $cadence_{y_i}$ and GCT_{y_i} estimates after processing each audio segment. As shown in Figure 9, the app's interface starts on the *Dashboard Page*, displaying a daily summary and historical trends of running metrics. Users can set specific target ranges for GCT and cadence on the *Set Target Page* using adjustable sliders. During a running session, users can switch between the *Focus Page* and the *Trend Page*. The *Focus Page* displays real-time metrics and target ranges through a dynamic bar with markers, while the *Trend Page* shows a graphical representation of performance trends over time. If the user fails to meet the target range, the app provides immediate feedback: the *Focus Page* prompts adjustments, and the *Trend Page* marks underperforming segments in red.

6 Training and Microbenchmarks

This section presents a micro-benchmark of (*i*) denoising methods, (*ii*) various feature combinations, and (*iii*) machine learning (ML) models to optimize the SoundTrack mobile platform's performance. We aim to balance model complexity and accuracy for real-time inference on mobile devices.

6.1 Denoising Methods

All data were collected exclusively from public gyms, as detailed in Section 4.2. In addition to the treadmill running sounds of interest, the dataset SoundTrackDB includes various background noises. Different phone models were used as the **`Treadmill Running Sound Phone`**, placed either horizontally or vertically on the display consoles, resulting in variations in sound energy due to model and placement differences. While mean-variance normalization (MVN) ensures consistent sound energy across different phones, effective noise suppression

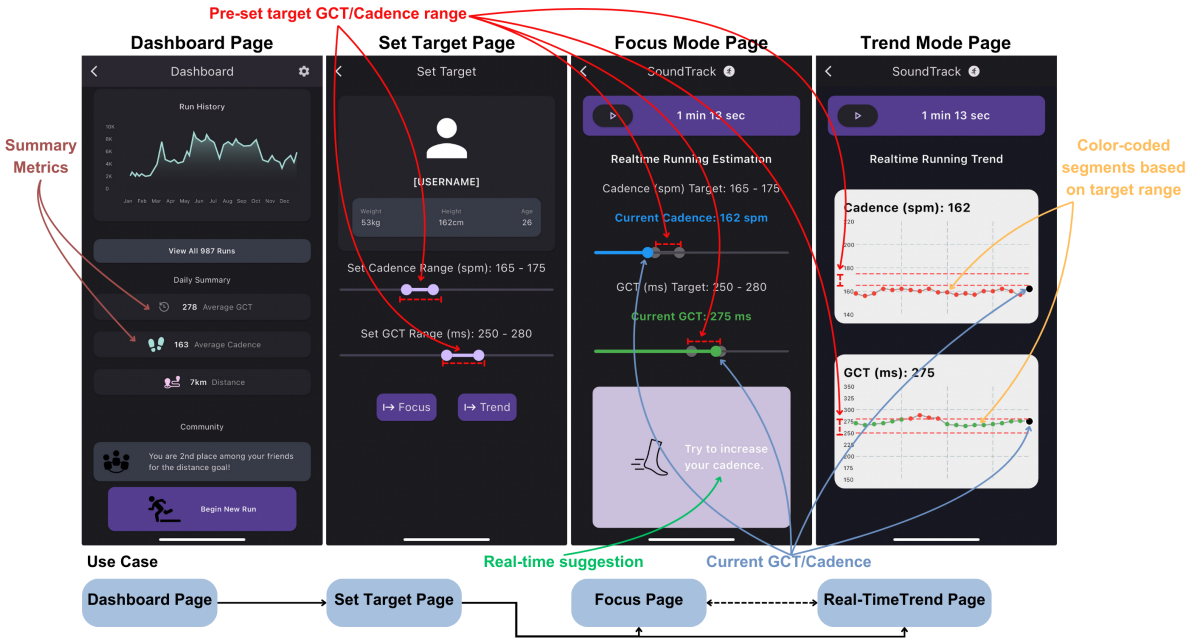


Fig. 9. SoundTrack mobile App demonstration, which contains: (1) *Dashboard Page* that displays summary metrics; (2) *Set Target Page* that allows the user to set specific targets for cadence and GCT as well as choose the visualization modes, *focus* or *trend*, during the run; (3) *Focus Mode Page* that provides real-time running metrics and suggestion; and (4) *Trend Mode Page* that offers a trend analysis of cadence and GCT during the run in addition to the real-time metrics.

methods are crucial for maintaining the performance of SoundTrack. We assessed spectral subtraction and filtering techniques, including Wiener filtering, commonly used in speech and acoustic processing for effective noise reduction.

6.2 Features

Acoustic Features: We evaluated widely recognized acoustic features used in audio signal processing: Mel spectrogram (Mel), Mel-frequency cepstral coefficients (MFCC), power spectral density (PSD), and root mean square energy (RMS) [51, 82, 106]. Audio was sampled at 16,000 Hz. For Mel and MFCC, we used 40 Mel bands, set the maximum frequency to 8,000 Hz, and applied a 1,024 window size for the short-time Fourier transform (STFT) with a hop length of 160. A 3-second audio segment produced feature matrices of shape (40, 300) for Mel and MFCC and (1, 300) for PSD and RMS, ensuring comprehensive audio data representation for our model.
IMU Features: We collected IMU data at 100 Hz from smartphones placed on the treadmill's control panel, which, unlike wearable devices, do not capture motion patterns as clearly due to their static position. To normalize orientation variations, we averaged and combined readings from the three IMU axes using the root-mean-square method, resulting in a unified feature matrix of shape (1, 300), reducing noise and enhancing IMU data usability for model training.

6.3 Models and Training

We investigate three prediction methods for estimating cadence and GCT from treadmill running audio: *Seq2Seq Detection*, *Classification*, and *Regression*, as detailed in Appendix B and shown in Figure 18. Both treadmill and IMU

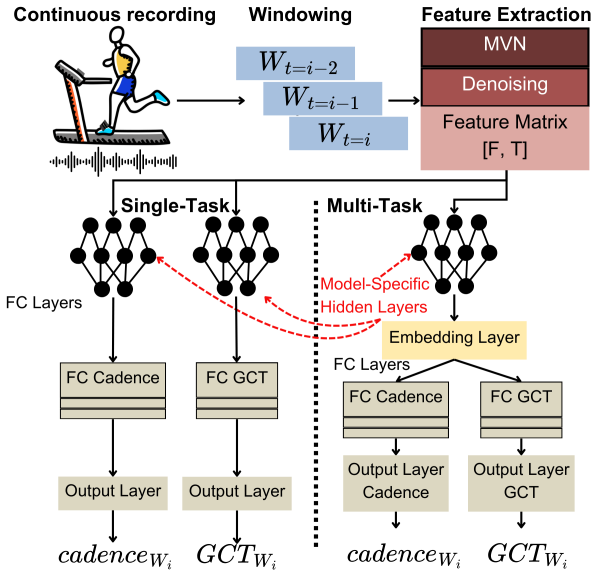


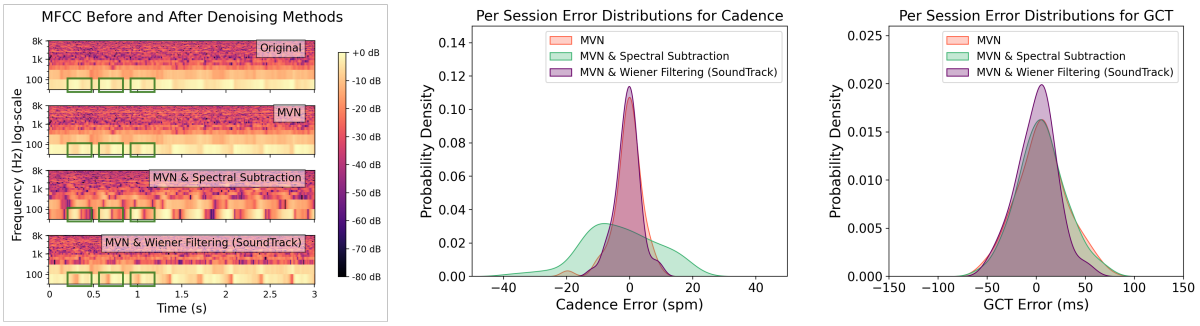
Fig. 10. Three types of models trained for each model architecture: (1) single-task GCT, (2) single-task cadence, and (3) multi-task model leveraging shared acoustic attributes for GCT and cadence estimation.

data were segmented into 3-second windows and transformed into feature matrices tailored to fit specific neural network input layer requirements. Given privacy concerns and potential network latency, we aim to implement all processing directly on-device. *As such, we focused on compact model architectures suitable for smartphones, excluding resource-intensive models like foundation and transformer-based models.*

6.3.1 Model Training. The training procedure for the models outlined in Section 6.3 involves several steps. We divided the SoundTrackDB dataset into 80% training, 10% validation, and 10% test sets, ensuring splits at the running session level. We chose not to split the data at the subject level because we empirically observed that running sounds exhibit significant variability even for the same subject under different conditions, such as wearing different shoes, running in different gyms, on different treadmills, or at varying speeds. Splitting at the session level increases data variability and prevents overly similar training, validation, and test sets.Sessions were segmented into 3-second windows, extracting *MFCC*, *Mel*, *RMS*, *PSD*, and IMU features as detailed in Section 6.2. To evaluate different feature combinations, we vertically stacked the feature matrices for each model configuration.

Each combination of features and model architectures is used to train three types of models: one for GCT, one for cadence, and one for multi-task learning—provided the model's structure supports it, depicted in Figure 10. *Seq2Seq* models, predicting foot-treadmill contact per frame, derive GCT and cadence during post-processing, thus requiring one model. *Classification* and *Regression* models were trained as single-task and multi-task models. Multi-task learning was achieved by extracting high-level embeddings with initial hidden layers and then feeding them into separate branches of fully connected layers for GCT and cadence. This shared structure leverages common acoustic attributes, creating a model with fewer parameters without sacrificing accuracy.

Seq2Seq models used Mean Squared Error (MSE) as the loss function. Classification and Regression models used Binary Cross Entropy (BCE) and MSE for both GCT and cadence, and a combined BCE/MSE for the multi-task model, weighted by an empirically derived hyperparameter $\lambda = 0.5$ (refer to Eq 1, Eq 2, and Eq 3 respectively). Although Mean Absolute Error (MAE) might provide a more direct interpretation of model performance, MSE



(a) (b) (c)

Fig. 11. (a): the effect of various denoising methods on the MFCC of an example 3-second window, with the `MVN & Wiener Filtering` approach yielding the most effective features for GCT-related sounds (boxed in green). (b) and (c): Comparison of error distribution for Cadence and GCT after different denoising methods for the testing set.

is preferred during training to more heavily penalize larger deviations from the ground truth. Models were trained using PyTorch for 100 epochs with a batch size of 128, AdamOptimizer, and a threshold learning rate scheduler to ensure convergence. We reserved 10% of the 363 sessions of SoundTrackDB as the test set. Each model was selected through a 9-fold cross-validation process at the session level, with each fold consisting of 80% training sessions and 10% validation sessions relative to the full dataset of 363 sessions in SoundTrackDB. The best-performing fold of each model is selected, and their effectiveness and comparative performance on the test set are analyzed in Section 6.4.

$$\text{Loss}_{\text{GCT}} = \text{MSE}_{\text{GCT}} = \frac{1}{n_{\text{GCT}}} \sum_{i=1}^{n_{\text{GCT}}} (\hat{w}_{i,\text{GCT}} - w_{i,\text{GCT}})^2 \tag{1}$$

$$\text{Loss}_{\text{Cadence}} = \text{MSE}_{\text{Cadence}} = \frac{1}{n_{\text{Cadence}}} \sum_{i=1}^{n_{\text{Cadence}}} (\hat{w}_{i,\text{Cadence}} - w_{i,\text{Cadence}})^2 \tag{2}$$

$$\text{Loss}_{\text{Multi-task}} = \lambda(\text{MSE}_{\text{GCT}}) + (1-\lambda)(\text{MSE}_{\text{Cadence}}) \tag{3}$$

6.4 Microbenchmarks

We benchmark the denoise approaches and models from Section 6.3 for cadence and GCT estimation performance, feature processing time, and model inference time on a smartphone.

Evaluation Metrics and Ground Truth: SoundTrack predicts average $cadence_y$ (spm) and GCT_y (ms) for each 3-second window W_i. The ground-truth cadence and GCT are averaged based on the per-step GCT and cadence labels in SoundTrackDB. While MSE is used during training, MAE evaluates model accuracy, providing a more intuitive performance measure.

Effectiveness of Denoising: As illustrated in green boxed areas in Figure 11a, Wiener Filter not only reduces noise but also preserves more acoustic signatures related to footstep sounds during treadmill running. As shown in Figure 11, using SoundTrack's MTL MLP-based `model` with MFCC and RMS as the feature, applying the Wiener Filter enhances the `model` performance compared to not using a filter and employing spectral subtraction as a denoising method.

Model Estimation Performance: Table 2 shows the best performance metrics for each model and features selected from the 9-Fold process in Section 6.3.1, evaluated on the test set. Single-task models generally do not outperform the multi-task model significantly, supporting the efficacy of multi-task learning in achieving

			Features	Mean Absolute Error (MAE)				Power Consumption (mW)
				Single-Task		Multi-Task		
				GCT* (ms)	Cadence (spm)	GCT* (ms)	Cadence (spm)	
Models	Seq2Seq	**LSTM-Seq2Seq**	MFCC+Mel+PSD+RMS	N/A	N/A	42.1	10.5	748.14
		1D-UNet	MFCC+Mel+PSD+RMS	N/A	N/A	44.2	18.9	618.78
		MLP-Seq2Seq	MFCC+Mel+PSD+RMS	N/A	N/A	46.6	14.4	627.54
	Classification	**AlexNet**	MFCC+Mel+PSD+RMS	17.9	1.9	21.5	1.9	674.19
		LSTM-Class	MFCC+Mel+PSD+RMS	17.4	3.0	18.1	3.1	777.45
	Regression	**GRU**	MFCC	16.2	6.3	17.7	6.4	727.08
		MLP	MFCC+RMS	16.6	2.4	**15.8**	**2.9**	**612.60**

Table 2. Micro-benchmarks of model estimation performances selected from the 9-Fold cross-validation, reporting the accuracy achieved by each type of model (single-task & multi-task) on the test set, the features they used, and power consumption on a smartphone (iPhone 15 Pro Max). * Note that GCT in SoundTrackDB are derived from 60 fps video, each missed frame equates to a 16 ms error.

competitive results with simpler models, thus optimizing mobile implementation latency. *Seq2Seq Detection* models showed suboptimal performance. The LSTM-Seq2Seq model, effective in prior studies [108], had high error rates for both cadence and GCT with the SoundTrackDB dataset, indicating challenges in generalizing across diverse conditions. In the *Classification* approach, AlexNet, in its cadence-only configuration, achieved the best cadence estimation accuracy with a 1.9 spm MAE but had limited GCT estimation capability, possibly due to the 2D convolution's proficiency in identifying distinct patterns like cadence but not the details with finer time resolution required for GCT. For the *Regression* approach, the multi-task MLP model not only performed well in GCT estimation with the best accuracy of 15.8 ms in MAE, surpassing its single-task counterpart but also maintained a respectable performance in cadence estimation at 2.9 spm in MAE. This suggests that multi-task MLP learns to estimate both cadence and GCT simultaneously, outperforming estimating each of them individually.

Model Power Consumption: Table 2 shows the power consumption of each model and its feature preprocessing on an iPhone 15 Pro Max, averaged over a 10-minute run based on battery drain. The multi-task MLP model had the lowest power consumption at 612.6 mW, making it the most energy-efficient among all models while maintaining top performance. In contrast, the LSTM-Class model, despite its competitive GCT estimation, consumed approximately 27% more than the MLP model. The Seq2Seq models showed moderate power consumption ranging from 627.54 mW to 748.14 mW, but their suboptimal accuracy makes them less attractive for practical deployment. AlexNet's power consumption of 674.19 mW, while moderate, doesn't justify its use given its limited GCT estimation capability despite having the best cadence estimation.

Feature Processing Latency: SoundTrack uses the Flutter framework [28], which lacks native acoustic feature extraction support. We developed feature extraction functions (MVN, Wiener Filtering, and acoustic features) from scratch, following established mathematical procedures detailed in Section 7.1. Table 3 presents average processing times for a 3-second window on an iPhone 15 Pro Max. The comprehensive feature set, MVN + Wiener Filtering + (MFCC+Mel+PSD+RMS+IMU), requires about 182 ms to process. MVN + Wiener Filtering, with individual MFCC, Mel, or PSD processing, takes about 130 ms each due to FFT and power spectrum calculations.

Model Inference Latency: For an accurate assessment of real-time performance capabilities, all models originally developed in PyTorch were converted to TensorFlowLite (TFLite) format using `ai_edge_torch` [37] and their inference latency was evaluated on an iPhone 15 Pro Max. To ensure consistency in our comparisons, each model was tested using the same input features, specifically a feature matrix of shape (41, 300) consisting of MFCC and RMS. Table 4 summarizes the findings, including the number of parameters, total multiply-add operations (mult-adds) for a forward pass, and the average inference time measured across 1000 samples. The MLP-Seq2Seq model was the most lightweight and rapid, with 0.02 million (M) parameters and an inference time of 2.52 ms.

Features MVN + Wiener Filter +	Output Dimension (F, T)	Processing Time (*ms*)
MFCC	(40, 300)	137
Mel	(40, 300)	133
PSD	(1, 300)	130
RMS	(1, 300)	65
IMU	(1, 300)	51
MFCC+Mel+PSD+RMS+IMU	(83, 300)	182
MFCC+PSD+RMS	(42, 300)	140
Mel+PSD+RMS	(42, 300)	134
MFCC+RMS	(41, 300)	**139**

Table 3. Micro-benchmarks of custom feature processing time implemented on a smartphone (iPhone 15 Pro Max).

Model	# Parameters (*M*)	Total mult-adds (*M*)	Inference Time (*ms*)
LSTM-Seq2Seq	1.27	381.24	21.36
1D-UNet	0.35	56.64	3.59
MLP-Seq2Seq	0.02	6.61	2.52
AlexNet	4.01	71.87	8.82
LSTM-Class	1.51	440.26	23.78
GRU	1.78	493.56	25.52
MLP	13.31	13.31	**3.35**

Table 4. Micro-benchmarks of TFLite model inference time on a smartphone (iPhone 15 Pro Max).

Acoustic Features (F)	MFCC (40)	Mel (40)	RMS (1)	PSD (1)	IMU (1)	MFCC (40) Mel (40) PSD (1) RMS (1) IMU (1)	MFCC (10-30)	MFCC (40) PSD (1) RMS (1)	Mel (40) PSD (40) RMS (1)	MFCC (40) IMU (1)	MFCC (40) RMS (1)
GCT MAE (*ms*)	17.1	18.1	28.1	34.5	38.0	28.3	19.0	22.3	19.7	16.8	**15.8**
Cadence MAE (*spm*)	4.1	3.7	2.4	8.5	3.5	2.2	5.9	2.3	2.4	3.2	**2.9**

Table 5. Micro-benchmarks for the feature selection process of the multi-task MLP model. 11 acoustic feature combinations were tested, with MFCC+RMS producing the best result.

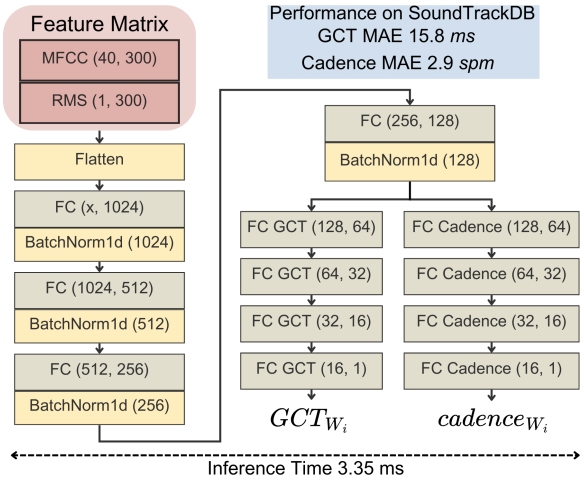


Fig. 12. Neural network structure of regression-based multi-task MLP model with MFCC+RMS as input features.

Fold	Cadence MAE (spm)			GCT MAE (ms)	
	SoundTrack	RunScribe	Watch	SoundTrack	RunScribe
1	3.5	0.6	8.4	14.3	15.2
2	3.4	0.2	7.4	18.5	20.1
3	3.7	0.9	6.0	18.6	19.0
4	2.9	0.6	7.6	18.1	14.7
7	3.3	0.9	7.3	16.5	18.3
5	3.8	0.7	8.4	17.5	19.7
6	4.2	0.7	6.0	19.1	20.1
8	3.5	0.9	8.1	16.4	19.6
9	**2.7**	**0.8**	**7.8**	**14.8**	**18.5**

Table 6. 9-Fold Cross-Validation results of the Multitask-MLP model with MFCC+RMS features.

RNN-based models, including LSTM-Seq2Seq, LSTM-Classification, and GRU, had higher complexity and required over 20 ms for inference.

MLP Acoustic Features Selection: The multi-task MLP-based model demonstrated the highest performance in our analysis, warranting a closer look at the feature selection process. As shown in Table 5, we evaluated 11 acoustic feature combinations. Among the higher-dimensional MFCC and Mel features, MFCC proved more effective for Ground Contact Time (GCT), while Mel features excelled in cadence estimation. However, combining MFCC and Mel did not improve performance for both metrics. IMU features alone showed some capability in estimating cadence (MAE of 3.5 *spm*) but was inadequate for GCT estimation (MAE of 38.0 *ms*). Additionally,

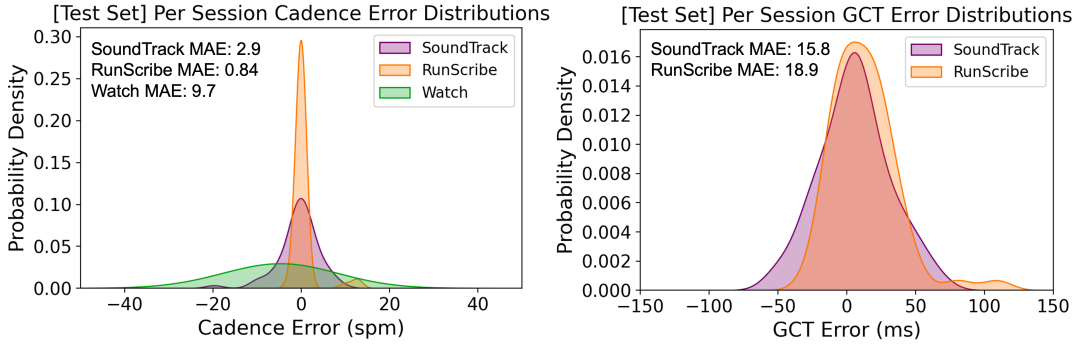


Fig. 13. Session-wise error distribution comparison for Cadence and GCT estimation in Test Set.

combining MFCC with IMU did not yield better results than combining MFCC with RMS, indicating redundancy in the information captured by IMU and RMS features. Previous studies suggested that isolating mid-frequency bands in MFCC or Mel could optimize performance [108], but our broader dataset indicates that maintaining the full frequency range is more effective. Lower-dimensional features like PSD, RMS, and IMU are sufficient for capturing cadence but inadequate for accurately estimating GCT, highlighting the need for higher-dimensional features like MFCC/Mel.

Ultimately, the combination of MFCC and RMS was the most effective, achieving an MAE of 15.8 *ms* for GCT and 2.9 *spm* for cadence. This is likely due to MFCC's frequency-domain information aiding GCT estimation and RMS's clear pattern recognition aiding cadence estimation. Based on these findings, MFCC and RMS are used as the features in SoundTrack.

6.5 SoundTrack Evaluation

The multi-task learning MLP model achieved consistently high accuracy during the 9-fold cross-validation process, shown in Table 6 for illustration purposes. Taking into account the trade-off between computational complexity, inference latency, and the accuracy of cadence and GCT estimation for real-time mobile applications – as demonstrated by the microbenchmarks in Section 6.4 – we selected this regression-based multi-task learning MLP model with MVN and Wiener Filter for noise handling, and MFCC and RMS for acoustic features in SoundTrack (with the model architecture shown in Figure 12). With SoundTrackDB as the ground truth, we compare the per-session cadence and GCT estimation performance for the test set between SoundTrack, RunScribe, and sports watch, with the error distribution shown in Figure 13. For cadence estimation, SoundTrack significantly outperforms the sports watch. While SoundTrack has a slightly higher overall mean absolute error (MAE) compared to RunScribe, it is important to note that RunScribe exhibits significant overestimations of cadence in certain sessions, with errors reaching around 10 *spm* (as indicated by the side lobe). For GCT estimation, SoundTrack outperforms RunScribe, which tends to overestimate GCT.

In addition to session-wise errors, the per-window errors for the 3-second windows in the test set are also analyzed. Figure 14 shows the comparison of Cadence and GCT Errors per window across speed, location, and subject in test set, the numbers of windows for each category are annotated by the red bars. We can see that SoundTrack has stable performance across all speeds for both cadence and GCT estimation. For different locations, the results indicate that SoundTrack overall consistently outperforms the sports watch and RunScribe across various locations, although some locations may present slightly higher errors, suggesting environmental factors might influence the estimations. For different subjects, the analysis reveals variability in performance, where some subjects experience higher errors than others. This discrepancy may be attributed to individual differences in running style, body mechanics, or familiarity with the running conditions. Despite this variability, SoundTrack

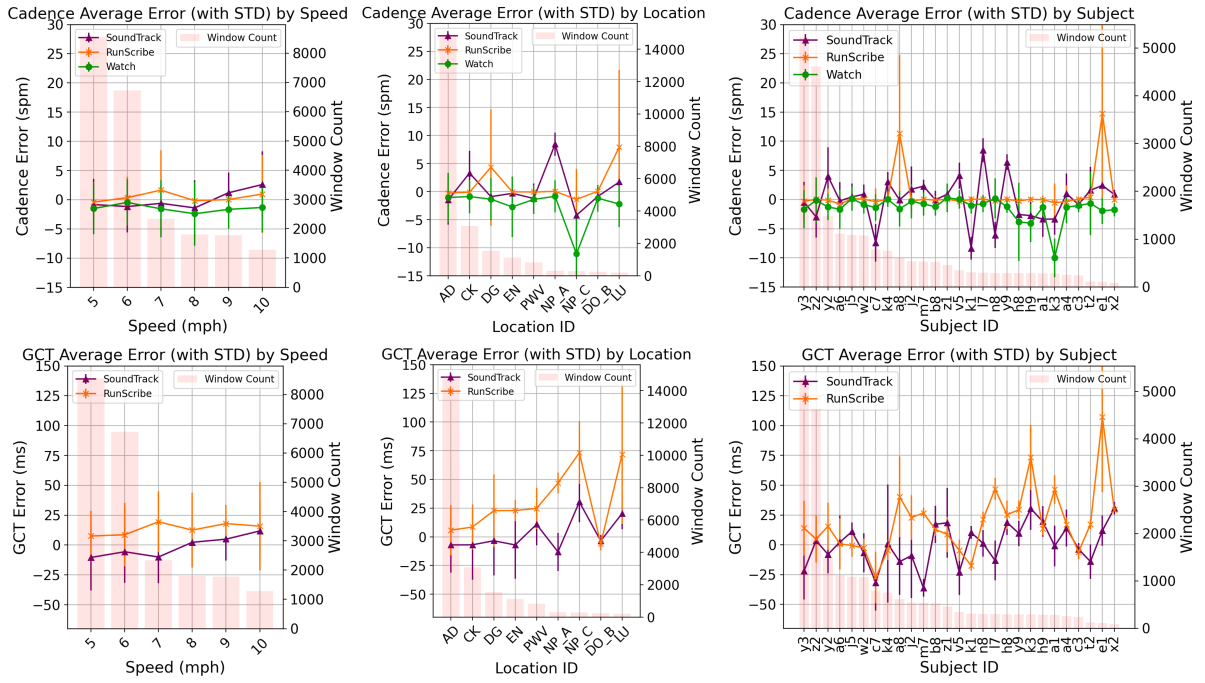


Fig. 14. Comparison of Cadence and GCT Errors per window across speed, location, and subject in test set. (Note that sport watches do not provide GCT for treadmill running.)

maintains a competitive edge in overall accuracy compared to RunScribe across most subjects, especially in GCT estimation. These factors suggest that while SoundTrack maintains an overall competitive advantage, further refinement could be achieved by creating and investigating a more balanced dataset that ensures an equal number of sessions across all categories.

7 SoundTrack Implementation and Case Study

7.1 SoundTrack Implementation

We implemented SoundTrack using the Flutter framework, taking advantage of its cross-platform compatibility with Android and IOS and a wide range of mobile platforms.

Custom Feature Extraction: The implementation of SoundTrack on mobile devices presented unique challenges due to the absence of a native audio processing package and an efficient matrix operation library in Flutter necessitates reliance on sequential calculations in Flutter, as discussed in Section 6.4. To address these challenges, we custom-develop essential signal processing functions such as framing, windowing, Short-Time Fourier Transform (STFT), Discrete Cosine Transform (DCT), Power Spectral Density (PSD), and Root Mean Square (RMS) in Flutter and ensure these functions produce identical outputs on mobile platforms and during training.

Implemented Model Structure: Figure 12 illustrates the neural network structure of the regression-based multi-task MLP model with MFCC and RMS as features. The model achieves GCT MAE of 15.8 *ms*, cadence MAE of 2.9 *spm*, and inference time of 3.35 *ms*. It flattens the input matrix (41, 300) and processes it through four fully connected layers. The output is then split into two branches, each with four fully connected layers ending in a single neuron to estimate GCT and cadence.

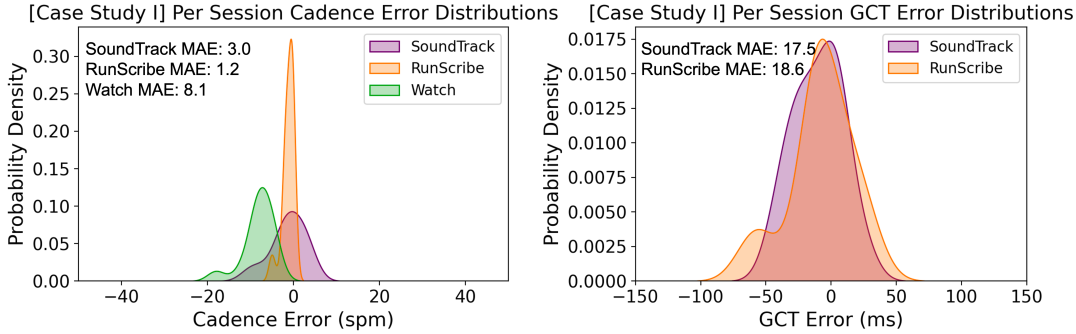


Fig. 15. Case study session-wise error distribution comparison for Cadence and GCT estimation.

Data	MAE Cadence (spm)			MAE GCT (ms)		SoundTrack Implementation Platform
	SoundTrack	RunScribe	Watch	SoundTrack	RunScribe	
SoundTrackDB	2.9	0.8	9.7	15.8	18.9	Smartphone
Case Study I	3.0	1.2	8.1	17.5	18.6	Smartphone
Case Study II	1.4	0.7	10.7	14.7	13.2	Android Treadmill
Case Study III	1.3	0.4	2.1	12.5	16.3	Smartphone

Table 7. Results comparing SoundTrack's performance on SoundTrackDB, 26 regular sessions in Case Study I, 2 varying-speed fartlek running sessions with SoundTrack implemented on NordicTrack 2450 treadmill in Case Study II, and additional tests in Case Study III.

Model Conversion: To facilitate mobile deployment, we converted the PyTorch-based MLP model to TensorFlowLite (TFLite) using Google's `ai_edge_torch` package [37]. We validated the TFLite model with a random sample, ensuring it retained the original model's accuracy within 0.01%. This step was crucial to ensure the mobile implementation matched the desktop prototype's performance.

7.2 Case Study

We implemented SoundTrack on smartphones (including an iPhone 15 Pro Max, an iPhone 14, and a Google Pixel 7) and conducted a user study with 11 *new* subjects across 3 *new* environments (public gyms) to evaluate the system's performance, versatility, and real-world applicability. Qualitative feedback from the subjects on the SoundTrack is also collected. This evaluation is crucial because the SoundTrack `model` was trained by randomly splitting SoundTrackDB at the session level, meaning that the same subjects, treadmills, and environments could appear in both the training, validation, and test sets. In contrast, for the case study, we ensured that the subjects, treadmills, and environments were entirely unseen during the `model` training and evaluation. Our case study involved 3 parts: (*i*) 26 regular running sessions with identical setup as SoundTrackDB with unseen subjects and gym environments, (*ii*) 2 running sessions with varying speed, and (*iii*) 7 running sessions with a 2-degree incline.

7.2.1 Case Study I - Regular Running. **Procedure:** All 11 new participants completed one or more 5-minute running sessions at self-selected speeds (6, 7, 8, or 9 *mph*), following the same experiment setup as SoundTrackDB described in Section 3 for fair comparison. **Results:** Table 7 summarizes the results for SoundTrack estimating the 26 regular case study sessions compared to the SoundTrackDB test set. SoundTrack recorded a MAE of 17.5 *ms* for GCT_y, with a MAPE of 7.21%, and a $cadence_y$ MAE of 3.0 *spm* with a MAPE of 1.74%, closely aligning with its performance on the SoundTrackDB test set. Figure 15 further illustrates the similarity in error distributions

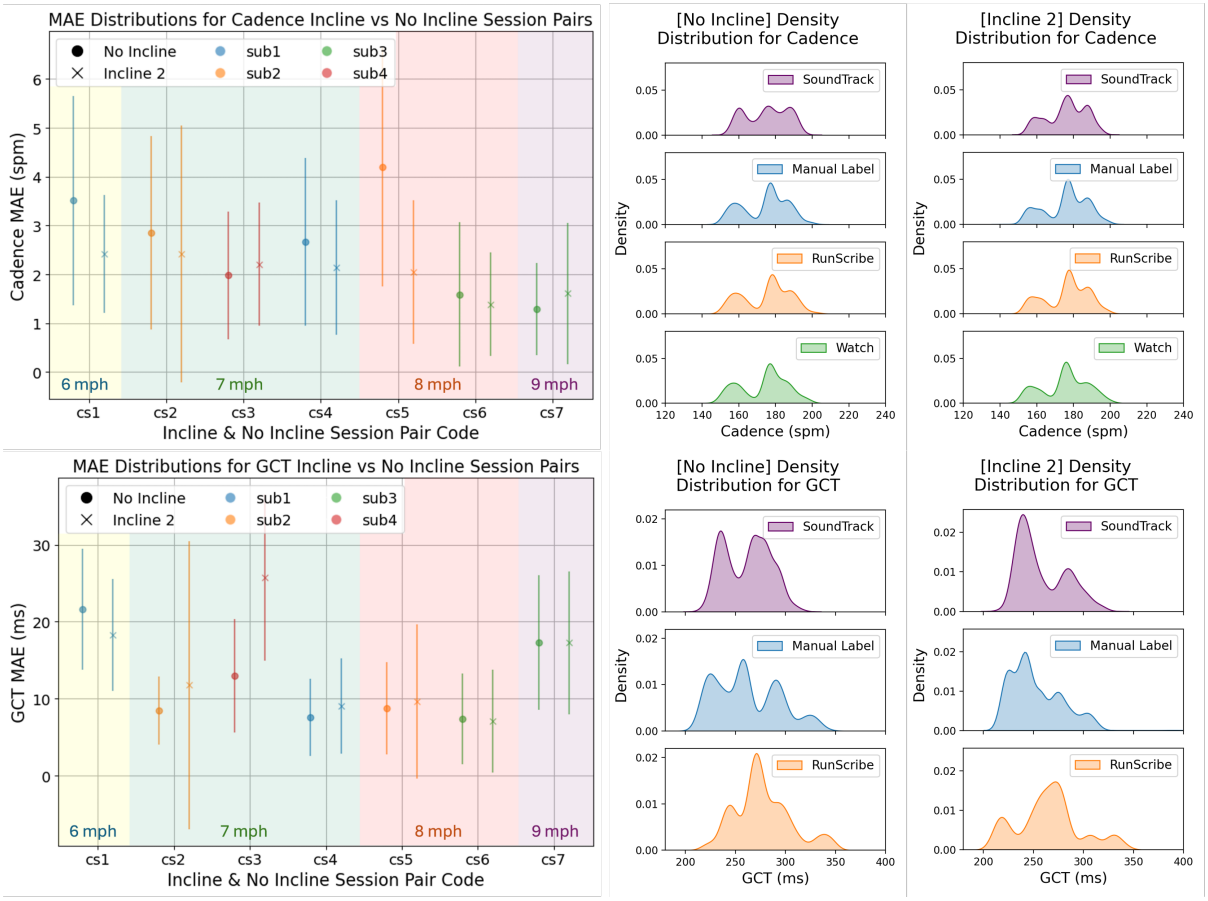


Fig. 16. Error and distribution comparison of Cadence and GCT for running sessions with/without incline in Case Study III.

between SoundTrack's estimation of the case study and the SoundTrackDB test set, confirming its generalization capability across new runners and environments.

7.2.2 Case Study II - SoundTrack Implemented on Treadmill & Running at Varying Speeds. **Procedure:** To simulate real-world treadmill running conditions more closely, we implemented SoundTrack on a NordicTrack 2450 treadmill [1] (a recent premium treadmill model with Android System, priced at 2, 999, as shown in Figure 1). From the 11 new participants, 2 runners were invited to complete a fartlek session, during which they varied their running speed (e.g., 30 seconds each at 7, 8, and 9 mph consecutively). This speed variation likely induced subconscious changes in the runners' cadence and ground contact time (GCT). The processing time for each 3-second window, including noise mitigation, acoustic feature extraction, and model inference, averaged around 1, 900 *ms* on the NordicTrack 2450. **Results:** Table 7 also shows the results for SoundTrack estimating the 2 fartlek sessions at varying speeds, obtaining an MAE of 14.7 *ms* for *GCT*, with a MAPE of 6.41%, and a *cadence* MAE of 1.4 *spm* with a MAPE of 1.47%, slightly outperforming SoundTrack's estimations on both the test set of SoundTrackDB and the 26 regular running sessions of Case Study I, showing that SoundTrack is able to adapt to real-world treadmill running scenarios.

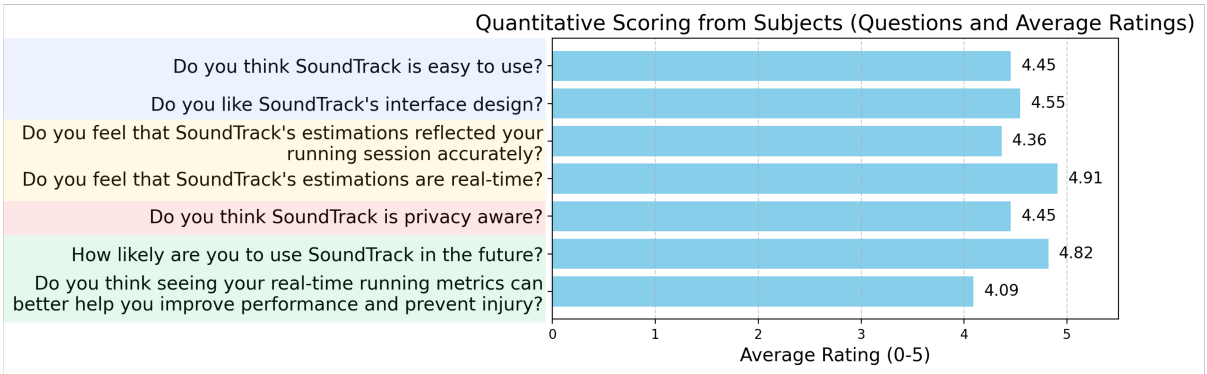


Fig. 17. Quantitative scoring of SoundTrack by the 11 new subjects in the case study in the areas of (*i*) Usability, (*ii*) Performance, (*iii*) Privacy, and (*iv*) Future Use.

7.2.3 Case Study III - Inclined Running. **Procedure:** We tested inclined running sessions in our case study to understand the SoundTrack's performance for treadmill running with incline, as many runners incorporate moderate inclines in their regular workouts. 4 of the 11 new subjects also performed 7 inclined running sessions (2-degree incline, a common incline for runners to run on treadmills) at the same speeds as their non-inclined sessions, enabling a comparative analysis. Since the YOLO-based ground-truth generation method described in Section 4.1 is not designed for inclined running, we manually labeled the 7 inclined and non-inclined session pairs to produce the ground-truth labels. **Results:** Figure 16 shows the error bars (standard deviations of MAE) for the 7 inclined sessions compared to the 7 non-inclined sessions at the same speeds. We observe similar MAE error ranges of both cadence and GCT for most inclined and non-inclined sessions. Additionally, Figure 16 also illustrates the distribution of cadence and GCT estimations and manual labels for both inclined and non-inclined sessions with the comparison to baseline commodity devices. SoundTrack's cadence estimation distributions for inclined and non-inclined sessions closely match the shape of the manually labeled ground-truths. SoundTrack's GCT estimates show greater variability in inclined and non-inclined sessions compared to manual labels, though they outperform RunScribe's estimates. Notably, this incline case study results do not reveal any significant reduction in SoundTrack's accuracy for estimating cadence or GCT in inclined sessions compared to non-inclined sessions.

7.2.4 Quantitative and Qualitative Evaluation from Subjects. To evaluate the usability, accuracy, and potential impact of SoundTrack, we conducted a user study with the 11 new participants in our case study who tested the system in real-world treadmill running scenarios. The study focused on four key aspects: (*i*) Usability, (*ii*) Performance, (*iii*) Security and Privacy, and (*iv*) Satisfaction and Future Use. Participants provided feedback through Likert-scale ratings and open-ended responses, offering insights into their experience with SoundTrack. Figure 17 illustrates the specific questions asked and the average scoring of the participants, reflecting participants' overall satisfaction.

Participants found SoundTrack intuitive and easy to use, with minimal configuration required and "*no need for calibration*" (Sub2). Many appreciated the ability to track running metrics without additional hardware, as it improved comfort and convenience, without "*wearables causing skin irritation during long runs*" (Sub7). The user interface was rated as highly intuitive, and real-time feedback was seen as "*particularly useful*," as mentioned by a subject: "*Seeing my cadence and GCT trend live makes adjusting my pace much easier than reviewing data afterward*" (Sub6). Since SoundTrack processes all data locally, privacy concerns were minimal. Participants preferred this over cloud-based alternatives, with one stating, "*knowing SoundTrack keeps everything on my phone*

instead of sending recordings elsewhere makes me feel much more comfortable using it" (Sub3). Most participants expressed interest in continuing to use SoundTrack, appreciating its convenience, privacy, and real-time feedback.

8 Discussion, Limitation, and Future Work

Although much of the literature has focused on gait metric analysis for walking, with applications in areas like person recognition and biomechanics analysis [9, 15, 20, 23, 24, 32, 65, 69, 88, 100, 109], walking generally presents a lower risk of injury compared to running and often does not require as detailed a focus on metrics. Running and walking differ significantly in biomechanics, with running characterized by higher intensity, a longer stride length, and a distinct aerial phase where both feet are momentarily off the ground. Consequently, SoundTrack was designed specifically for treadmill running rather than walking, aiming to provide a robust, accessible, affordable, portable, and pervasive solution for real-time monitoring of running metrics. Providing high-accuracy GCT and cadence estimations at a level of precision typically reserved for experienced or elite runners using professional devices priced at $500, SoundTrack is intended to support a broader audience in tracking exercise performance, enhancing running techniques, and potentially reducing injury risks, rather than a replacement for professional-grade tools.

We did not evaluate SoundTrack on outdoor running, as surfaces like concrete, asphalt, and dirt vary widely in texture and firmness, presenting a stark contrast to treadmill belts. Outdoor running also introduces different environmental acoustics: in open spaces, sound waves dissipate quickly, while in enclosed urban environments, they may echo. On treadmills, however, sound is more contained, influenced by the treadmill's mechanical noise and indoor acoustics, resulting in a distinct auditory profile. Given that many newer treadmill models now include built-in speakers and microphones and based on our Case Study II described in Section 7.2.2, we see potential for treadmill manufacturers to integrate SoundTrack directly into their systems. Although the current processing time on the NordicTrack 2450 is longer than on smartphones, manufacturers could enhance performance by incorporating more advanced processing systems. This integration would allow users to access detailed running metrics directly from their treadmill, enriching the workout experience.

In SoundTrack, we utilized a Wiener filter to mitigate the impact of pervasive environmental noises, such as those commonly found in public gyms and variable treadmill operating sounds. While Section 6 demonstrates the effectiveness of the Wiener filter, we are also investigating alternative methods for noise reduction, sound separation, and audio enhancement. Potential approaches include unsupervised sound separation using mixture invariant training [105], self-supervised audio enhancement through pretraining with synthetic data [46, 117], and real-time noise suppression [12]. Additionally, integrating a sound quality assessment module could enhance the accuracy of running metric estimations by focusing on confident audio segments [16, 29]. Consequently, we are exploring deep-learning-based audio quality classifiers specifically tailored for treadmill running scenarios.

In addition to cadence and ground contact time (GCT), there are several other running metrics that are essential for comprehensive exercise performance monitoring and injury prevention. Metrics such as footstrike type, which can provide insights into running form and biomechanics, and exertion levels, which reflect the intensity of the workout, are particularly valuable. We plan to enhance SoundTrack to incorporate the detection of these additional running metrics.

We aim to continuously expand our participant pool, treadmill models, running speeds and inclinations, and gym locations to enrich SoundTrackDB, thereby validating our findings with a broader user base and enhancing SoundTrack. As indicated in Section 6.5 and illustrated in Figure 14, environmental factors and individual differences can significantly influence SoundTrack's performance. To address this, we plan to create a more balanced dataset that accounts for variations among subjects, treadmill models, locations, and running speeds. Additionally, we will investigate whether calibration can enhance performance for specific participants and assess the necessity of incorporating per-user calibration. Furthermore, we will evaluate the system's effectiveness

across the most common treadmill types and identify which models yield the best results. As mentioned in Section 7.2.3, our current CV-based ground truth annotation method has limitations when applied to running on an incline. To address this, we plan to explore improved methods for generating accurate ground truth data specifically for treadmill running with incline and further validate SoundTrack's performance.

Moreover, we see the potential in further exploring the Seq2Seq approach investigated in Section 6.3, experimenting with more signal processing techniques and refining the downstream post-processing step. By enabling runners to input personal attributes such as expertise levels and body metrics, we anticipate creating a more personalized and accurate SoundTrack implementation. These adaptations will enable us to tailor feedback and insights to individual profiles, enhancing user engagement through customized recommendations and training plans that cater to each runner's unique needs and goals.

9 Conclusion

In conclusion, we designed, developed, and evaluated SoundTrack, a contactless mobile-platform-based solution that accurately estimates cadence and ground contact time (GCT) using on-device machine learning, providing feedback and visualization in real time. We created SoundTrackDB, the *first* extensive audio dataset of treadmill running metrics, collected from 61 subjects running at 6 different paces (5, 6, 7, 8, 9, and 10 *mph*), from 24 treadmills in 13 public gyms, covering 363 treadmill sessions and totaling 40 hours of data. *We open-source SoundTrackDB to advance research in ubiquitous computing, audio signal processing, sports science, and related fields.* Our evaluations show that SoundTrack outperforms the commercial RunScribe running pods in GCT estimation, achieving a mean absolute percentage error (MAPE) of 6.05%. Additionally, SoundTrack demonstrates improved cadence estimation accuracy over commercial sports watches and achieves performance comparable to RunScribe running pods, with a MAPE of 1.62%. Case studies further demonstrate SoundTrack's robustness and generalizability across new users, different environments, and incline running, validating its suitability for real-world applications. Collectively, SoundTrack and SoundTrackDB offer a non-intrusive alternative to traditional wearable devices, democratizing access to advanced running metrics and supporting runners in enhancing performance and mitigating injury risks. SoundTrack delivers high-accuracy GCT and cadence estimations to everyday runners, providing a level of precision typically reserved for experienced or elite athletes using devices priced at $500 or more.

Acknowledgments

This research was partially supported by the National Science Foundation under Grant Number 2133516 and Apple Inc. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Columbia University, NSF, or the U.S. Government or any of its agencies.

References

[1] 2024. NordicTrack Commercial 2450 Treadmill. https://www.nordictrack.com/en-US/product/c-2450-ntl19124 Accessed: 2024-10-29.

[2] Haya Alaskar, Nada Alzhrani, Abir Hussain, and Fatma Almarshed. 2019. The implementation of pretrained AlexNet on PCG classification. In *Intelligent Computing Methodologies: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part III 15*. Springer, 784–794.

[3] Apple Inc. 2024. Apple Watch. https://www.apple.com/watch/?afid=p238%7CsNZgeoZeS-dc_mtid_1870765e38482_pcrid_692396963514_pgrid_99322576784_pntwk_g_pchan__pexid__ptid_kwd-52218226_&cid=aos-us-kwgo-watch--slid---product- Accessed: 2024-06-29.

[4] Ivan Au, Leo Ng, Paul Davey, Marco So, Brian Chan, Pinky Li, Will Wong, Tania Althorpe, Sarah Stearne, and Roy Cheung. 2021. Comparison of foot strike sound between rearfoot, midfoot and forefoot strike runners. *Journal of Athletic Training* (2021).

[5] J Baltich, CA Emery, JL Whittaker, and BM Nigg. 2017. Running injuries in novice runners enrolled in different training interventions: a pilot randomized controlled trial. *Scandinavian Journal of Medicine & Science in Sports* 27, 11 (2017), 1372–1383.

[6] Adeola Bannis, Shijia Pan, Carlos Ruiz, John Shen, Hae Young Noh, and Pei Zhang. 2023. IDIoT: Multimodal framework for ubiquitous identification and assignment of human-carried wearable devices. *ACM Transactions on Internet of Things* 4, 2 (2023), 1–25.

[7] BarBend. 2022. So How Do Smartphones Affect Your Workout, Anyway? https://barbend.com/smartphones-affect-your-workout/ Accessed: 2024-06-28.

[8] Joanne Barker. 2020. What running mistakes lead to injury? https://answers.childrenshospital.org/running-mistakes-injury/ Accessed: 2024-06-28.

[9] Lauren C Benson, Anu M Räisänen, Christian A Clermont, and Reed Ferber. 2022. Is this the real life, or is this just laboratory? A scoping review of IMU-based running gait analysis. *Sensors* 22, 5 (2022), 1722.

[10] Lauren C Benson, Anu M Räisänen, Valeriya G Volkova, Kati Pasanen, and Carolyn A Emery. 2020. Workload a-WEAR-ness: monitoring workload in team sports with wearable technology. A scoping review. *Journal of Orthopaedic & Sports Physical Therapy* 50, 10 (2020), 549–563.

[11] Gunnar Borg. 1998. *Borg's perceived exertion and pain scales*. Human Kinetics.

[12] Sebastian Braun, Hannes Gamper, Chandan KA Reddy, and Ivan Tashev. 2021. Towards efficient models for real-time deep noise suppression. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 656–660.

[13] Sicco A Bus. 2003. Ground reaction forces and kinematics in distance running in older-aged men. *Medicine & Science in Sports & Exercise* 35, 7 (2003), 1167–1175.

[14] Chao Cai, Zhe Chen, Henglin Pu, Liyuan Ye, Menglan Hu, and Jun Luo. 2020. AcuTe: Acoustic thermometer empowered by a single smartphone. In *Proceedings of the 18th conference on embedded networked sensor systems*. 28–41.

[15] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. 2022. Cross vision-rf gait re-identification with low-cost rgb-d cameras and mmwave radars. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–25.

[16] Andrew A Catellier and Stephen D Voran. 2020. Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 331–335.

[17] Robert F Chapman, Abigail S Laymon, Daniel P Wilhite, James M Mckenzie, David A Tanner, and Joel M Stager. 2012. Ground contact time as an indicator of metabolic cost in elite distance runners. *Medicine and science in sports and exercise* 44, 5 (2012), 917–925.

[18] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust vital signs waveform recovery via deep interpreted RF sensing. In *Proceedings of the 27th annual international conference on mobile computing and networking*. 392–405.

[19] Enrique Colino, Jose Luis Felipe, Bas Van Hooren, Leonor Gallardo, Kenneth Meijer, Alejandro Lucia, Jorge Lopez-Fernandez, and Jorge Garcia-Unanue. 2020. Mechanical properties of treadmill surfaces compared to other overground sport surfaces. *Sensors* 20, 14 (2020), 3822.

[20] CA Collazos, HE Castellanos, JA Cardona, JC Lozano, A Gutiérrez, and MA Riveros. 2017. A simple physical model of human gait using principles of kinematics and BTS GAITLAB. In *VII Latin American Congress on Biomedical Engineering CLAIB 2016, Bucaramanga, Santander, Colombia, October 26th-28th, 2016*. Springer, 333–336.

[21] COROS. 2024. COROS: Performance Sports Technology. https://us.coros.com/ Accessed: 2024-06-28.

[22] Rocco Di Michele and Franco Merni. 2014. The concurrent effects of strike pattern and ground-contact time on running economy. *Journal of science and medicine in sport* 17, 4 (2014), 414–418.

[23] Yiwen Dong, Megan Iammarino, Jingxiao Liu, Jesse Codling, Jonathon Fagert, Mostafa Mirshekari, Linda Lowes, Pei Zhang, and Hae Young Noh. 2024. Ambient floor vibration sensing advances the accessibility of functional gait assessments for children with muscular dystrophies. *Scientific Reports* 14, 1 (2024), 10774.

[24] Jonathon Fagert, Mostafa Mirshekari, Shijia Pan, Linda Lowes, Megan Iammarino, Pei Zhang, and Hae Young Noh. 2021. Structure-and sampling-adaptive gait balance symmetry estimation using footstep-induced structural floor vibrations. *Journal of Engineering Mechanics* 147, 2 (2021), 04020151.

[25] Mathieu Falbriard, Frédéric Meyer, Benoît Mariani, Grégoire P Millet, and Kamiar Aminian. 2020. Drift-free foot orientation estimation in running using wearable IMU. *Frontiers in bioengineering and biotechnology* 8 (2020), 65.

[26] Fellrnr. n.d.. Running Sensors. https://fellrnr.com/wiki/Running_Sensors Accessed: 2024-06-23.

[27] Matt Fitzgerald. 2020. Ground Contact Time and Running Performance. https://www.trainingpeaks.com/blog/ground-contact-time-and-running-performance/ Accessed: 2024-06-28.

[28] Flutter. 2024. Flutter: Build beautiful native apps in record time. https://flutter.dev/ Accessed: 2024-06-28.

[29] Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang. 2018. Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. *arXiv preprint arXiv:1808.05344* (2018).

[30] Moshe Gabel, Ran Gilad-Bachrach, Erin Renshaw, and Assaf Schuster. 2012. Full body gait analysis with Kinect. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 1964–1967.

[31] Bojana Galic. 2023. 126 Running Statistics You Need to Know in 2024. https://www.livestrong.com/article/13730338-running-statistics/ Accessed: 2024-06-28.

[32] Francisco A Garcia, Juan C Pérez-Ibarra, Marco H Terra, and Adriano AG Siqueira. 2022. Adaptive algorithm for gait segmentation using a single IMU in the thigh pocket. *IEEE Sensors Journal* 22, 13 (2022), 13251–13261.
[33] Abhinav Garg, Gowtham P Vadisetti, Dhananjaya Gowda, Sichen Jin, Aditya Jayasimha, Youngho Han, Jiyeon Kim, Junmo Park, Kwangyoun Kim, Sooyeon Kim, et al. 2020. Streaming On-Device End-to-End ASR System for Privacy-Sensitive Voice-Typing.. In *Interspeech*. 3371–3375.
[34] Garmin. 2024. Garmin: Advanced GPS Technology. https://www.garmin.com/en-US/ Accessed: 2024-06-28.
[35] M. Gholami and C. Menon. 2024. Treadmill Running with IMUs and Custom Piezoresistive Strain Sensors on one Lower Limb. https://doi.org/10.20383/103.0871
[36] Mansueto Gomes Neto, Leonardo Fossati Metsavaht, Fabio Luciano Arcanjo, Janice de Souza Guimarães, Cristiano Sena Conceição, Eliane Celina Guadagnin, Vitor Oliveira Carvalho, and Gustavo Leporace de Oliveira Lomelino Soares. 2023. Epidemiology of Lower-extremity Musculoskeletal Injuries in Runners: An Overview of Systematic Reviews. *Current Emergency and Hospital Medicine Reports* 11, 2 (2023), 74–87.
[37] Google AI Edge. n.d.. ai edge torch. https://ai.google.dev/edge/lite/models/convert_pytorch Accessed: 2024-06-25.
[38] Xiaonan Guo, Jian Liu, and Yingying Chen. 2017. FitCoach: Virtual fitness coach empowered by wearable mobile devices. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE, 1–9.
[39] Tian Hao, Guoliang Xing, and Gang Zhou. 2015. RunBuddy: a smartphone system for running rhythm monitoring. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 133–144.
[40] Mahmoud Hassan, Florian Daiber, Frederik Wiehr, Felix Kosmalla, and Antonio Krüger. 2017. Footstriker: An EMS-based foot strike assistant for running. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 1 (2017), 1–18.
[41] Bryan C Heiderscheit, Elizabeth S Chumanov, Max P Michalski, Christa M Wille, and Michael B Ryan. 2011. Effects of step rate manipulation on joint mechanics during running. *Medicine and science in sports and exercise* 43, 2 (2011), 296.
[42] André Henriksen, Martin Haugen Mikalsen, Ashenafi Zebene Woldaregay, Miroslav Muzny, Gunnar Hartvigsen, Laila Arnesdatter Hopstock, and Sameline Grimsgaard. 2018. Using fitness trackers and smartwatches to measure physical activity in research: analysis of consumer wrist-worn wearables. *Journal of medical Internet research* 20, 3 (2018), e110.
[43] Xiaowen Hou and Chao Liu. 2022. Rope Jumping Strength Monitoring on Smart Devices via Passive Acoustic Sensing. *Sensors* 22, 24 (2022), 9739.
[44] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring Breathing Mode during Running with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25.
[45] Zhizhang Hu, Yue Zhang, and Shijia Pan. 2021. Footstep-Induced Floor Vibration Dataset: Reusability and Transferability Analysis. In *Proceedings of the 19th ACM conference on embedded networked sensor systems*. 546–551.
[46] Bryce Irvin, Marko Stamenovic, Mikolaj Kegler, and Li-Chia Yang. 2023. Self-supervised learning for speech enhancement through synthesis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.
[47] Armağan Karahanoğlu, Rúben Gouveia, Jasper Reenalda, and Geke Ludden. 2021. How are sports-trackers used by runners? Running-related data, personal goals, and self-tracking in running. *Sensors* 21, 11 (2021), 3687.
[48] Mia Kercher. 2022. Run With Your Phone: Options For Carrying Your Phone While Running. https://marathonhandbook.com/how-to-run-with-your-phone/ Accessed: 2024-06-28.
[49] Hyosu Kim, Anish Byanjankar, Yunxin Liu, Yuanchao Shu, and Insik Shin. 2018. UbiTap: Leveraging acoustic dispersion for ubiquitous touch interface on solid surfaces. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. 211–223.
[50] Jiha Kim, Younho Nam, Jungeun Lee, Young-Joo Suh, and Inseok Hwang. 2023. ProxiFit: Proximity Magnetic Sensing Using a Single Commodity Mobile toward Holistic Weight Exercise Monitoring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–32.
[51] Agni Kumar, Vikramjit Mitra, Carolyn Oliver, Adeeti Ullal, Matt Biddulph, and Irida Mance. 2021. Estimating respiratory rate from breath audio obtained through wearable microphones. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 7310–7315.
[52] Worapan Kusakunniran, Qiang Wu, Jian Zhang, and Hongdong Li. 2010. Support vector regression for multi-view gait recognition based on local motion feature selection. In *2010 IEEE Computer society conference on computer vision and pattern recognition*. IEEE, 974–981.
[53] Philipp Leusmann, Christian Mollering, Lars Klack, Kai Kasugai, Martina Ziefle, and Bernhard Rumpe. 2011. Your floor knows where you are: sensing and acquisition of movement data. In *2011 IEEE 12th International Conference on Mobile Data Management*, Vol. 2. IEEE, 61–66.
[54] Dong Li, Jialin Liu, Sunghoon Ivan Lee, and Jie Xiong. 2022. Lasense: Pushing the limits of fine-grained activity sensing using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.
[55] Ryan T Li, Scott R Kling, Michael J Salata, Sean A Cupp, Joseph Sheehan, and James E Voos. 2016. Wearable performance devices in sports medicine. *Sports health* 8, 1 (2016), 74–78.

[56] Shuheng Li, Ranak Roy Chowdhury, Jingbo Shang, Rajesh K Gupta, and Dezhi Hong. 2021. Units: Short-time fourier inspired neural networks for sensory time series classification. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 234–247.
[57] Daniel E Lieberman, Eric R Castillo, Erik Otarola-Castillo, Meshack K Sang, Timothy K Sigei, Robert Ojiambo, Paul Okutoyi, and Yannis Pitsiladis. 2015. Variation in foot strike patterns among habitually barefoot and shod runners in Kenya. *PloS one* 10, 7 (2015), e0131354.
[58] Boram Lim and Young Sub Kwon. 2021. The Effect of Stride Frequency on Running Economy and Running Distance During High Intensity Treadmill Running. *Journal of Health, Sports, and Kinesiology* 2, 1 (2021), 13–14.
[59] Hyunchul Lim, Yaxuan Li, Matthew Dressa, Fang Hu, Jae Hoon Kim, Ruidong Zhang, and Cheng Zhang. 2022. Bodytrak: Inferring full-body poses from body silhouettes using a miniature camera on a wristband. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–21.
[60] James Little and Jeffrey Boyd. 1998. Recognizing people by their gait: the shape of motion. *Videre: Journal of computer vision research* 1, 2 (1998), 1–32.
[61] Alan Liu, Yu-Tai Lin, and Karthikeyan Sundaresan. 2024. View-agnostic Human Exercise Cataloging with Single MmWave Radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–23.
[62] Tao Liu, Yoshio Inoue, Kyoko Shibata, and K Shiojima. 2011. A mobile force plate and three-dimensional motion analysis system for three-dimensional gait assessment. *IEEE Sensors Journal* 12, 5 (2011), 1461–1467.
[63] Cunguang Lou, Shuo Wang, Tie Liang, Chenyao Pang, Lei Huang, Mingtao Run, and Xiuling Liu. 2017. A graphene-based flexible pressure sensor with applications to plantar pressure measurement and gait analysis. *Materials* 10, 9 (2017), 1068.
[64] Saif Mahmud, Ke Li, Guilin Hu, Hao Chen, Richard Jin, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2023. Posesonic: 3d upper body pose estimation through egocentric acoustic sensing on smartglasses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–28.
[65] Samah AF Manssor, Shaoyuan Sun, and Mohammed AM Elhassan. 2021. Real-time human recognition at night via integrated face and gait recognition technologies. *Sensors* 21, 13 (2021), 4323.
[66] Luca Marotta, Jaap H Buurke, Bert-Jan F van Beijnum, and Jasper Reenalda. 2021. Towards machine learning-based detection of running-induced fatigue in real-world scenarios: Evaluation of IMU sensor configurations to reduce intrusiveness. *Sensors* 21, 10 (2021), 3451.
[67] John A Mercer, PAUL Devita, Tim R Derrick, and Barry T Bates. 2003. Individual effects of stride length and frequency on shock attenuation during running. *Medicine & Science in Sports & Exercise* 35, 2 (2003), 307–313.
[68] Isabel S Moore, Kelly J Ashford, Charlotte Cross, Jack Hope, Holly SR Jones, and Molly McCarthy-Ryan. 2019. Humans optimize ground contact time and leg stiffness to minimize the metabolic cost of running. *Frontiers in Sports and Active Living* 1 (2019), 53.
[69] Alvaro Muro-De-La-Herran, Begonya Garcia-Zapirain, and Amaia Mendez-Zorrilla. 2014. Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. *Sensors* 14, 2 (2014), 3362–3394.
[70] Arun Asokan Nair and Kazuhito Koishida. 2021. Cascaded time+ time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7153–7157.
[71] National Heart, Lung, and Blood Institute. 2023. Study reveals wearable device trends among U.S. adults. https://www.nhlbi.nih.gov/news/2023/study-reveals-wearable-device-trends-among-us-adults Accessed: 2024-06-28.
[72] Jingping Nie, Yuang Fan, Ziyi Xuan, Matthias Preindl, and Xiaofan Jiang. 2024. Real-Time Non-Contact Estimation of Running Metrics on Treadmills using Smartphones. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 1644–1646. https://doi.org/10.1145/3636534.3697446
[73] Corina Nüesch, Elena Roos, Geert Pagenstert, and Annegret Mündermann. 2017. Measuring joint kinematics of treadmill walking and running: Comparison between an inertial sensor based system and a camera-based system. *Journal of biomechanics* 57 (2017), 32–38.
[74] Pew Research Center. 2024. Demographics of Mobile Device Ownership and Adoption in the United States. https://www.pewresearch.org/internet/fact-sheet/mobile/ Accessed: 2024-06-28.
[75] Alexandra Pfister, Alexandre M West, Shaw Bronner, and Jack Adam Noah. 2014. Comparative abilities of Microsoft Kinect and Vicon 3D motion capture for gait analysis. *Journal of medical engineering & technology* 38, 5 (2014), 274–280.
[76] Angkoon Phinyomark, Blayne A Hettinga, Sean T Osis, and Reed Ferber. 2014. Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. *PloS one* 9, 8 (2014), e105246.
[77] Yogarajah Pratheepan, Joan V Condell, and Girijesh Prasad. 2009. The use of dynamic and static characteristics of gait for individual identification. In *2009 13th International Machine Vision and Image Processing Conference*. IEEE, 111–116.
[78] Manju Rana and Vikas Mittal. 2020. Wearable sensors for real-time kinematics analysis in sports: A review. *IEEE Sensors Journal* 21, 2 (2020), 1187–1207.

[79] Aman Rehman and Shailender Kumar. 2024. Leveraging MediaPipe and YOLO Keypoint Detection in Ensemble Approaches for Workout Pose Recognition. In *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. IEEE, 695–700.
[80] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. 2023. Real-time flying object detection with YOLOv8. *arXiv preprint arXiv:2305.09972* (2023).
[81] Yanzhi Ren, Zhourong Zheng, Hongbo Liu, Yingying Chen, Hongwei Li, and Chen Wang. 2021. Breathing sound-based exercise intensity monitoring via smartphones. In *2021 international conference on computer communications and networks (ICCCN)*. IEEE, 1–10.
[82] Francesco Renna, Jorge Oliveira, and Miguel T Coimbra. 2019. Deep convolutional neural networks for heart sound segmentation. *IEEE journal of biomedical and health informatics* 23, 6 (2019), 2435–2445.
[83] Mareike Roell, Hubert Mahler, Johannes Lienhard, Dominic Gehring, Albert Gollhofer, and Kai Roecker. 2019. Validation of wearable sensors during team sport-specific movements in indoor environments. *Sensors* 19, 16 (2019), 3458.
[84] RunScribe. 2024. RunScribe: Advanced Running Metrics. https://runscribe.com/ Accessed: 2024-06-28.
[85] Amber Sayer. 2024. How Many People Have Run A Marathon?: A Global Analysis. *Marathon Handbook* (2024). https://marathonhandbook.com/how-many-people-have-run-a-marathon/ Accessed: 2024-10-16.
[86] Amy G Schubert, Jenny Kempf, and Bryan C Heiderscheit. 2014. Influence of stride frequency and length on running mechanics: a systematic review. *Sports health* 6, 3 (2014), 210–217.
[87] Christian Seeger, Alejandro Buchmann, and Kristof Van Laerhoven. 2012. myHealthAssistant: a phone-based body sensor network that captures the wearer's exercises throughout the day. In *6th International ICST Conference on Body Area Networks*.
[88] Dimple Sethi, Sourabh Bharti, and Chandra Prakash. 2022. A comprehensive survey on gait analysis: History, parameters, approaches, pose estimation, and future work. *Artificial Intelligence in Medicine* 129 (2022), 102314.
[89] Matthias Seuter, Max Pfeiffer, Gernot Bauer, Karen Zentgraf, and Christian Kray. 2017. Running with technology: Evaluating the impact of interacting with wearable devices on running movement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–17.
[90] Matthias Seuter, Alexandra Pollock, Gernot Bauer, and Christian Kray. 2020. Recognizing running movement changes with quaternions on a sports watch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–18.
[91] Statista. 2024. Number of Treadmill Users in the United States from 2006 to 2024. https://www.statista.com/statistics/191605/users-of-treadmills-in-the-us-since-2006/ Accessed: 2024-06-27.
[92] Strava, Inc. 2022. Year in Sport 2022. https://www.strava.com/yis-community-2022#community Accessed: 2024-06-28.
[93] Paul Streli, Rayan Armani, Yi Fei Cheng, and Christian Holz. 2023. Hoov: Hand out-of-view tracking for proprioceptive interaction using inertial sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
[94] David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.
[95] Stryd. 2024. Stryd: Power Meter for Running. https://www.stryd.com/us/en Accessed: 2024-06-28.
[96] Ke Sun, Chen Chen, and Xinyu Zhang. 2020. " Alexa, stop spying on me!" speech privacy protection against voice assistants. In *Proceedings of the 18th conference on embedded networked sensor systems*. 298–311.
[97] Jeremiah J Tate and Clare E Milner. 2017. Sound-intensity feedback during running reduces loading rates and impact peak. *journal of orthopaedic & sports physical therapy* 47, 8 (2017), 565–569.
[98] Justin Trautmann, Lin Zhou, Clemens Markus Brahms, Can Tunca, Cem Ersoy, Urs Granacher, and Bert Arnrich. 2021. TRIPOD—A Treadmill Walking Dataset with IMU, Pressure-Distribution and Photoelectric Data for Gait Analysis. *Data* 6, 9 (2021), 95. https://doi.org/10.3390/data6090095
[99] Ruben Vera-Rodriguez, John SD Mason, Julian Fierrez, and Javier Ortega-Garcia. 2012. Comparative analysis and fusion of spatiotemporal information for footstep recognition. *IEEE transactions on pattern analysis and machine intelligence* 35, 4 (2012), 823–834.
[100] Dequan Wang, Xinran Zhang, Kai Wang, Lingyu Wang, Xiaoran Fan, and Yanyong Zhang. 2024. RDGait: A mmWave Based Gait User Recognition System for Complex Indoor Environments Using Single-chip Radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–31.
[101] Jianyang Wang, Dongheng Zhang, Binbin Zhang, Jinbo Chen, Yang Hu, and Yan Chen. 2024. RF-GymCare: Introducing Respiratory Prior for RF Sensing in Gym Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–28.
[102] Penghao Wang, Ruobing Jiang, Zhongwen Guo, and Chao Liu. 2024. Afitness: Fitness Monitoring on Smart Devices via Acoustic Motion Images. *ACM Transactions on Sensor Networks* 20, 4 (2024), 1–24.
[103] Zeng Wang, Xiancheng Wang, and Ruidong Li. 2023. Treadmill Deck Performance Optimization Design Based on Muscle Activity during Running. *Applied Sciences* 13, 18 (2023), 10457.
[104] Richard W Willy. 2018. Innovations and pitfalls in the use of wearable devices in the prevention and rehabilitation of running related injuries. *Physical Therapy in Sport* 29 (2018), 26–33.

[105] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron Weiss, Kevin Wilson, and John Hershey. 2020. Unsupervised sound separation using mixture invariant training. *Advances in neural information processing systems* 33 (2020), 3846–3857.
[106] Stephen Xia, Jingping Nie, and Xiaofan Jiang. 2021. Csafe: An intelligent audio wearable platform for improving construction worker safety in urban environments. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (Co-Located with CPS-IoT Week 2021)*. 207–221.
[107] Yadong Xie, Fan Li, Yue Wu, and Yu Wang. 2021. HearFit: Fitness monitoring on smart speakers via active acoustic sensing. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.
[108] Ziyi Xuan, Ming Liu, Jingping Nie, Minghui Zhao, Stephen Xia, and Xiaofan Jiang. 2023. CaNRun: Non-Contact, Acoustic-based Cadence Estimation on Treadmills using Smartphones. In *Proceedings of Cyber-Physical Systems and Internet of Things Week 2023* (San Antonio, TX, USA) *(CPS-IoT Week '23)*. Association for Computing Machinery, New York, NY, USA, 272–277. https://doi.org/10.1145/3576914.3589561
[109] Huanqi Yang, Mingda Han, Mingda Jia, Zehua Sun, Pengfei Hu, Yu Zhang, Tao Gu, and Weitao Xu. 2023. XGait: Cross-Modal Translation via Deep Generative Sensing for RF-based Gait Recognition. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 43–55.
[110] Yanni Yang, Jiannong Cao, and Xiulong Liu. 2019. ER-rhythm: Coupling exercise and respiration rhythm using lightweight COTS RFID. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.
[111] Fraser Young, Rachel Mason, Jason Moore, Samuel Stuart, Rosie Morris, and Alan Godfrey. 2022. A proposed computer vision model for running gait assessment. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 4773–4776. https://doi.org/10.1109/EMBC48229.2022.9871739
[112] Yingying Yuan, Bo Liu, Hui Li, Mo Li, Yingqiu Song, Runze Wang, Tianlu Wang, and Hangyu Zhang. 2022. Flexible wearable sensors in medical monitoring. *Biosensors* 12, 12 (2022), 1069.
[113] Jinrui Zhang, Deyu Zhang, Xiaohui Xu, Fucheng Jia, Yunxin Liu, Xuanzhe Liu, Ju Ren, and Yaoxue Zhang. 2020. MobiPose: Real-time multi-person pose estimation on mobile devices. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 136–149.
[114] Tianyue Zheng, Zhe Chen, Shujie Zhang, Chao Cai, and Jun Luo. 2021. MoRe-Fi: Motion-robust and fine-grained respiration monitoring via deep-learning UWB radar. In *Proceedings of the 19th ACM conference on embedded networked sensor systems*. 111–124.
[115] Bo Zhou, Sungho Suh, Vitor Fortes Rey, Carlos Andres Velez Altamirano, and Paul Lukowicz. 2022. Quali-mat: Evaluating the quality of execution in body-weight exercises with a pressure sensitive sports mat. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–45.
[116] Han Zhou, Yi Gao, Xinyi Song, Wenxin Liu, and Wei Dong. 2019. Limbmotion: Decimeter-level limb tracking for wearable-based human-computer interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.
[117] Qiu-Shi Zhu, Jie Zhang, Zi-Qiang Zhang, and Li-Rong Dai. 2023. A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 1927–1939.

A Data Collection Details

This section provides additional details on the specific models of sports watches used during data collection and the synchronization procedures described in Section 3.

Specific Watch Models: COROS Pace 3, Garmin Forerunner 255, Garmin Fenix 6 Sapphire, and Apple Watch Series 8.

Data Synchronization: Precise device synchronization is critical for accurate running metric estimation, as ground contact time (GCT) is measured at the millisecond level. While Unix timestamps are used, synchronization is imperfect due to sampling rate variations, network latency, NTP delays, and clock drift, introducing discrepancies of several hundred milliseconds. To address this, we introduced an audible cue (two clapping sounds) and a specific foot movement pattern to improve synchronization between the video from the **`Video Phone`**, the audio and IMU data from the **`Treadmill Running Sound Phone`**, and the RunScribe, as shown in Figure 3. For the foot movement pattern, participants were instructed to tap their toes five times on both the left and right feet, simulating extreme forefoot running. Afterward, participants started recording on the smartwatch and began the running session.

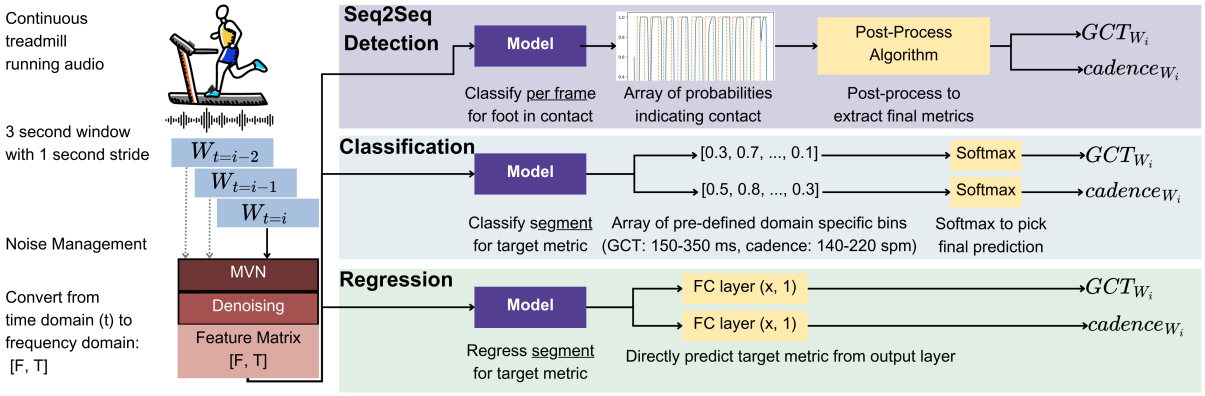


Fig. 18. Three prediction methods experimented: (1) *Seq2Seq Detection*, (2) *Classification*, and (3) *Regression*.

B Prediction Method Experimented

As described in Section 6.3, in this work, we experimented with three prediction methods for Cadence and GCT estimation based on treadmill running sound. In particular, *Seq2Seq Detection* uses a sequence-to-sequence (Seq2Seq) approach, where the model processes a sequence of frames and outputs a sequence of probabilities indicating foot-treadmill contact likelihood. A threshold-based post-processing step derives the final GCT and cadence predictions. We experimented with Seq2Seq LSTM networks [108] and 1dUNet models [70, 82]. *Classification* categorizes each input into predefined bins representing GCT and cadence ranges. Typical GCT values range from 150 to 350 milliseconds, and cadence values range from 140 to 220 steps per minute. The output layer is designed with 200 classes for GCT and 80 classes for cadence, using a softmax function for category determination. We investigated LSTM and CNN-based models, such as AlexNet, for time-series event classification [2, 56]. *Regression* directly predicts numeric values of cadence and GCT. This method uses a fully connected output layer with a single neuron as a regressor. We explored GRU-based and MLP-based models. GRUs effectively manage sequential data, addressing RNN issues like vanishing and exploding gradients, while MLPs, simpler yet robust, are ideal for on-device inference with minimal computational demand.