Design of a Web-Based Platform for Gamified Behavioral Data Collection with AI-Assisted Data Extraction

1st Asina Bopushova
Faculty of Engineering and Informatics
Ala-Too international university
Bishkek, Kyrgyzstan
https://orcid.org/0009-0002-1061-8675

2nd Ruslan Isaev
ICT Faculty
Paragon International University
Phnom Penh, Cambodia
risaev@paragoniu.edu.kh
https://orcid.org/0000-0003-4426-8837

***Abstract*—Sedentary behavior and fragmented health monitoring remain key challenges in modern digital environments, particularly among university students and working professionals. Existing mobile health (mHealth) systems often provide incomplete behavioral coverage and suffer from low long-term user engagement.**

This paper presents the design of a web-based platform for gamified behavioral data collection with AI-assisted data extraction. The system integrates structured daily check-ins, gamification mechanisms, and an AI-based pipeline for extracting health metrics from screenshots of third-party applications. The platform combines objective sensor-derived data and subjective self-reports into a unified longitudinal dataset.

A pilot deployment across eight countries involving 38 participants demonstrates the system's ability to collect diverse health metrics, including physical activity, sleep, hydration, and screen time, while maintaining user engagement through gamification mechanisms such as points, streaks, badges, and leaderboards. The resulting dataset is intended for further machine learning-based analysis and predictive modeling of health-related outcomes.

***Index Terms*—mobile health, gamification, AI-assisted data extraction, behavioral monitoring, longitudinal dataset.**

I. Introduction

Sedentary behavior has become a prevalent lifestyle pattern in modern digital society. Prolonged screen exposure, remote work, and intensive academic workloads contribute to extended sitting time and reduced physical activity, particularly in university environments where students and staff engage in long hours of studying, teaching, research, and administrative tasks.

Insufficient physical activity and prolonged sedentary behavior are strongly associated with increased risks of cardiovascular diseases, metabolic disorders, sleep disturbances, and reduced overall well-being. These risks highlight the need for continuous and reliable monitoring of lifestyle-related behavioral factors in real-world settings.

Mobile health (mHealth) systems are widely used for behavioral monitoring through smartphones and wearable devices, typically collecting data such as step count, heart rate, and sleep duration. However, existing systems often provide incomplete representations of daily behavior due to heterogeneous coverage of health indicators across applications. For instance, some applications track sleep metrics but omit stress or screen time, while others focus exclusively on physical activity. Furthermore, the use of multiple independent health and fitness applications leads to fragmented and inconsistent data across platforms, limiting the ability to construct unified longitudinal behavioral profiles.

Another key limitation is user engagement. In practice, users often demonstrate inconsistent long-term participation in health tracking systems. Although gamification techniques are commonly employed to improve engagement, they are typically implemented independently of the data collection process and do not fully address the challenge of sustained behavioral monitoring.

To address these limitations, this paper presents a web-based platform for integrated behavioral data collection. The system combines structured daily check-ins, gamification mechanisms, and an AI-assisted pipeline for extracting health metrics from screenshots of third-party applications. It integrates objective sensor-based data and subjective self-reports into a unified longitudinal dataset, enabling comprehensive analysis of behavioral patterns over time.

II. Related works / Literature review

According to the World Health Organization, physical inactivity is one of the leading global risk factors for mortality, contributing significantly to cardiovascular diseases, diabetes, and other chronic conditions [1]. Sedentary behavior, particularly prolonged sitting during study or office-based work, has been associated with increased all-cause mortality and metabolic risk [2]. These findings emphasize the importance of continuous monitoring of physical activity patterns in real-world environments.

A strong association has been reported between total sitting time and mortality risk, independent of overall physical activity levels [2]. This suggests that health outcomes are influenced not only by physical activity but also by extended periods of

uninterrupted sitting, highlighting the need to capture continuous behavioral patterns rather than isolated activity events.

To address this, mobile health (mHealth) systems have emerged as a key direction in digital health research. These systems integrate mobile sensing, wearable devices, and cloud-based analytics to collect behavioral and physiological data such as step count, sleep duration, heart rate variability, and energy expenditure [3], [4]. Evidence from large-scale wearable studies suggests that these signals can provide meaningful insights into health status and behavioral change [4].

However, most existing mHealth systems rely on single-ecosystem data collection (e.g., Apple HealthKit, Google Fit), which limits interoperability and cross-platform integration. At the same time, large-scale wearable datasets are often affected by inconsistencies, missing data, and a lack of standardization across devices and applications [4]. Together, these limitations highlight a critical challenge in achieving reliable and scalable data integration in real-world mHealth systems.

From a behavioral intervention perspective, mHealth platforms increasingly integrate gamification techniques to improve user engagement and long-term adherence. Systematic reviews show that gamification elements such as progress tracking, rewards, and feedback loops can significantly improve health-related behavior change [5]. However, maintaining long-term engagement remains a major challenge in real-world deployments. User experience and interface usability are also important factors affecting user engagement and long-term adoption of digital platforms [6].

More recent research has also explored the role of artificial intelligence in health-related gamified systems, including adaptive feedback generation and personalized interventions [7]. Nevertheless, most of these systems still rely on manually entered or sensor-generated structured data.

A key limitation in existing mHealth ecosystems is the lack of support for heterogeneous and unstructured data sources. In practice, users interact with multiple independent health applications (fitness trackers, sleep apps, nutrition apps), which produce fragmented data representations that are not easily unified.

Recent advances in computer vision and document intelligence have enabled automatic extraction of structured information from visual inputs such as screenshots, mobile dashboards, and application interfaces. Classical OCR systems such as Tesseract provide foundational text recognition capabilities [8], while modern transformer-based document understanding models significantly improve robustness in complex and noisy visual environments. For example, TrOCR introduces an end-to-end transformer architecture for optical character recognition [9], while LayoutLM enables joint modeling of text and layout information [10]. Recent progress in multimodal document understanding has improved the alignment between visual content and textual semantics, enabling more accurate interpretation of complex graphical interfaces and supporting structured information extraction from heterogeneous screen-based data sources [11].

These approaches enable transformation of unstructured visual inputs into structured datasets suitable for downstream analytics [12]. In addition, studies on gamified learning systems have demonstrated that game-based mechanisms improve user engagement, motivation, and task completion efficiency. These findings further support the use of gamification as an effective strategy for sustaining long-term participation in interactive data collection systems [13], [14]. However, their application in mobile health (mHealth) systems remains limited and underexplored.

In parallel, research in ubiquitous computing and digital health highlights the importance of integrating multi-source behavioral data streams into unified longitudinal datasets, including the combination of sensor data, user interaction data, and contextual signals for comprehensive health analytics.

III. Methodology

A. Analysis of Existing Health Tracking Systems

Existing mobile and web-based health tracking systems, including wearable platforms (e.g., Apple Health, Google Fit) and lifestyle applications, primarily rely on sensor-based data such as steps, heart rate, and sleep duration. They provide limited support for contextual behavioral factors (e.g., workload, stress, screen time, mood), and data is often fragmented across multiple applications. Most systems also lack effective mechanisms for long-term user engagement. These limitations motivated the development of a more integrated approach.

B. System Design and Proposed Solution

The proposed web-based platform integrates behavioral data collection, gamification, and AI-assisted data extraction. Built using Lovable, a prompt-driven, low-code AI-assisted environment, it enables rapid UI and backend development with built-in database functionality for structured and unstructured data. To reduce the friction of manual data entry, the platform combines gamification mechanisms (points, streaks, badges, leaderboards) with an optional AI-assisted screenshot extraction pipeline, allowing users to either enter metrics manually or upload a screenshot from a third-party health app. Sensor-based and self-reported data are fused in a client–backend–database–AI architecture. The AI module processes unstructured inputs, such as screenshots from third-party health apps, converting them into structured health metrics. The full dataset can also be exported as a CSV file for further analysis.The system workflow is shown in Fig. 1.

C. Data Collection and Study Population

The dataset was collected through a pilot deployment of the platform in several countries, including Kyrgyzstan, Cambodia, Russia, Canada, Hungary, Kazakhstan, and Italy. It contains daily behavioral check-ins with repeated observations over time. Each record includes demographic, behavioral, physiological, and AI-extracted data. In addition to raw behavioral and demographic data, derived health indicators were computed, including Body Mass Index (BMI, kg/m²), which was calculated from self-reported height and weight

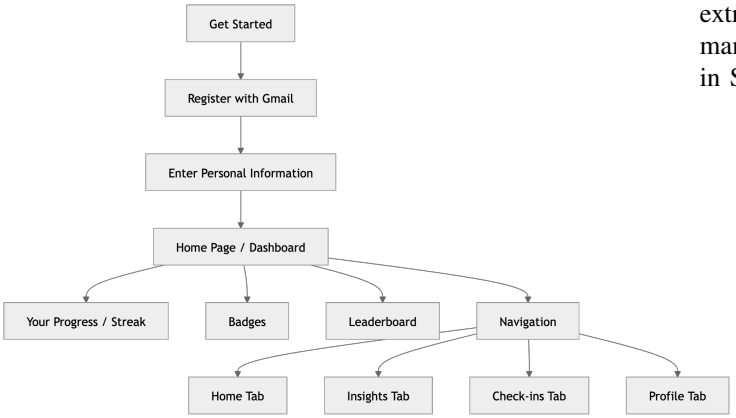


Fig. 1. User flow diagram of the proposed platform

Participants included university students (computer science, humanities, and engineering) and working professionals from fields such as healthcare, education, IT, and administration. Ages ranged from 17 to 57, with both male and female participants represented. Each user provided multiple daily entries, forming a longitudinal dataset suitable for analysis over time.

D. Gamification Mechanisms and Engagement Modeling

To improve long-term participation and reduce user dropout, gamification mechanisms were integrated into the platform design. These mechanisms serve both motivational and behavioral reinforcement functions.

To encourage sustained engagement, the platform integrates gamification elements such as points awarded for completing daily check-ins, streak tracking for continuous participation, and milestone-based badges reflecting consistency and activity levels. Leaderboards and progress dashboards introduce social comparison and competitive motivation. In addition, gamification metrics (e.g., points, streak length, and engagement frequency) are used as quantitative variables in the dataset to assess user adherence and behavioral consistency over time.

E. AI-Assisted Screenshot-Based Data Extraction

The system features an AI-assisted pipeline for extracting structured health data from screenshots of third-party applications. Screenshot processing is implemented as a server function that calls the Lovable AI Gateway, routed to Google's Gemini 2.5 Flash multimodal model, which performs OCR and metric-field extraction directly from the uploaded image. Users upload images through the check-in interface, the image is normalized and then will be sent, AI-based field extraction — Gemini 2.5 Flash interprets the screenshot and returns structured candidate values for relevant fields (e.g., step count, distance, calories burned, sleep duration), extracted values are mapped to the platform's predefined metric schema. This reduces manual data entry, integrates varied sources, and enhances data completeness. No formal accuracy evaluation of the extraction pipeline was performed in this pilot phase; extracted values were not cross-validated against ground-truth manual measurements. This is discussed further as a limitation in Section V.

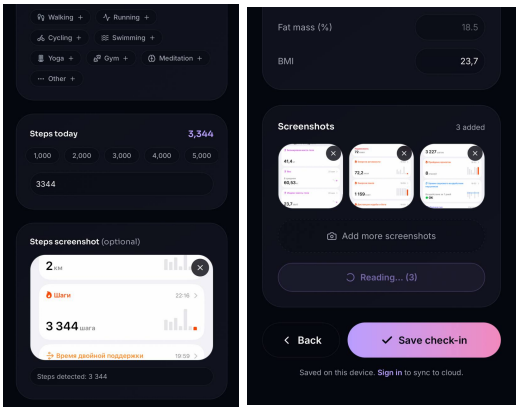


Fig. 2. Example of structured health data generated from a user screenshot

F. Ethics Statement

All participants provided informed consent prior to participation in the study. The collected data was anonymized and stored in compliance with basic data protection principles

IV. Results

The pilot deployment of the proposed platform involved 38 participants aged 17–57 years across Europe, Asia, and North America Fig. 3. Among them, 55.3% were male and 44.7% were female. The average participant age was 28.4 ± 8.7 years. BMI values, available for 34 participants, indicated that 15% were below the normal range, 65% within normal range, and 20% above normal range.

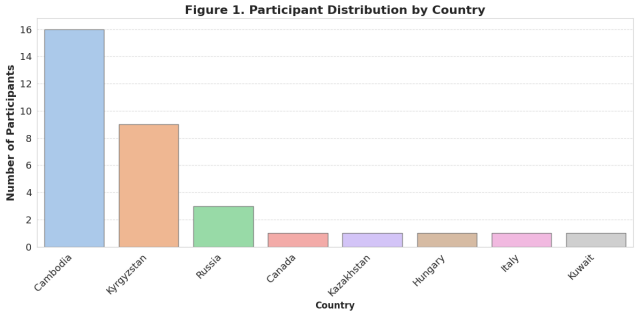


Fig. 3. Participant distribution by country.

Among active users, 70% maintained a streak of at least two consecutive days. The most frequently entered metrics were sleep duration (entered by 90% of users), steps (85%), and water intake (80%), whereas diet type and fasting information were reported less frequently (45% and 30%, respectively).

Health and behavioral metrics showed considerable variation: sleep duration ranged from 5 to 9 hours (mean 7.1 ± 1.2 h), daily steps from 1,200 to 18,000 (mean 7,450 ± 4,300), and active minutes from 20 to 120 (mean 65 ± 25). Calories burned ranged from 1,500 to 3,500 kcal (mean 2,400 ± 500). These

distributions illustrate the heterogeneity of the dataset, which is suitable for modeling behavioral patterns and engagement.

TABLE I
HEALTH AND BEHAVIORAL METRICS COLLECTED IN THE STUDY

Metric	Units / Scale
Sleep Duration	hours
Sleep Quality	1-5
Steps	steps
Distance Walked	km
Active Minutes	minutes
Calories Burned	kcal
Activity Type	categorical
Water Intake	liters
Work/Study Hours	hours
Screen Time	hours
Diet Type	categorical
Intermittent Fasting	yes/no
Religious Fasting	yes/no

The summary of collected health and behavioral metrics is presented in Table I, including sleep, physical activity, hydration, screen time, workload, mood, stress, and dietary habits. The platform also supports uploading screenshots from external health and fitness applications. An AI-assisted pipeline converts these screenshots into structured records containing metrics such as step count, distance, and calories burned when available Fig. 2.

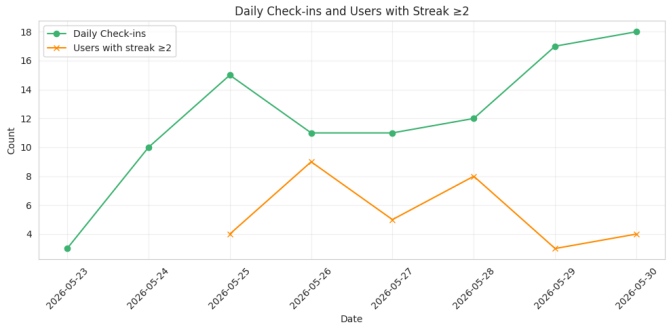


Fig. 4. Daily Check-ins and Users with Streak ≥ 2.

The system operated continuously throughout the pilot study via a web-based interface. Integration of self-reported and AI-extracted data enabled a unified longitudinal dataset from multiple sources. The platform also incorporated gamification elements (points, streaks, achievements, and progress tracking), which supported repeated user engagement and consistent completion of daily check-ins Fig. 5.

Overall, the results demonstrate the feasibility of the proposed platform for collecting and integrating behavioral health data while maintaining user engagement through gamified interaction mechanisms.

V. DISCUSSION

Results show that traditional mHealth systems often provide fragmented behavioral data and fail to maintain long-term user engagement. In contrast, the proposed platform integrates

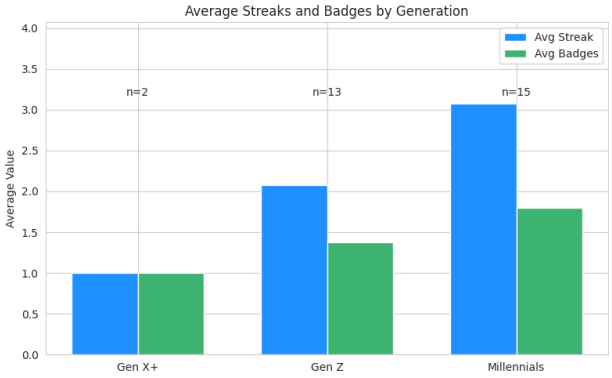


Fig. 5. User dashboard showing gamification elements (points, streaks, badges).

self-reports, gamification, and AI-based screenshot extraction into a unified system, enabling more complete and continuous behavioral monitoring.

Gamification elements (points, streaks, badges, leaderboards) significantly improved user participation during the pilot study, as reflected in repeated daily check-ins. This suggests that embedding motivation mechanisms directly into data collection can help reduce user dropout in longitudinal studies.

The AI-assisted screenshot processing module reduces manual data entry and improves data completeness by extracting structured metrics from third-party health applications. However, accuracy depends on interface variability across apps.

A direct quantitative comparison against established document-understanding baselines, such as Tesseract [8], TrOCR [9], or LayoutLM [10], as well as against manual data entry accuracy, was outside the scope of this pilot deployment and remains an important direction for future benchmarking.

The collected dataset is suitable for machine learning tasks, including behavioral pattern analysis and predictive modeling. It also includes engagement-related features that can be useful for modeling user adherence.

Limitations include small sample size, potential self-report bias, and limited generalizability. Despite this, the results suggest the feasibility of the proposed approach for building integrated behavioral datasets for future predictive health applications.

VI. CONCLUSION

This paper presented a web-based system for gamified behavioral data collection with AI-assisted extraction of health metrics from screenshots. The platform integrates multiple data sources, including self-reported check-ins and external health applications, into a unified longitudinal dataset.

The pilot deployment demonstrated the feasibility of collecting multi-dimensional behavioral data across different countries while maintaining consistent user engagement through gamification mechanisms such as points, streaks, and badges. Additionally, the use of an AI-based processing pipeline reduced manual data entry and enabled structured extraction

of information from unstructured visual inputs. Future work will include benchmarking the AI-assisted extraction pipeline against established OCR and document-understanding methods, as well as expanding the pilot to a larger and more balanced participant sample.

Overall, the proposed system provides a scalable approach for continuous behavioral monitoring and supports future machine learning applications in health analytics and predictive modeling.

REFERENCES

[1] World Health Organization. (2022). Global status report on physical activity 2022. Geneva: World Health Organization.

[2] Wilmot, E., Edwardson, C., Achana, F., Davies, M., Gorely, T., Gray, L., Khunti, K., Yates, T., Biddle, S. (2012). Sedentary time in adults and the association with diabetes, cardiovascular disease and death: systematic review and meta-analysis. Diabetologia, 55, 2895-2905. https://doi.org/10.1007/s00125-012-2677-z.

[3] Muzny, M., Henriksen, A., Giordanengo, A., Muzik, J., Grøttland, A., Blixgård, H., ... Årsand, E. (2020). Wearable sensors with possibilities for data exchange: Analyzing status and needs of different actors in mobile health monitoring systems. International journal of medical informatics, 133, 104017.

[4] Bent, B., Wang, K., Grzesiak, E., Jiang, C., Qi, Y., Jiang, Y., . Dunn, J. (2021). The digital biomarker discovery pipeline: An open-source software platform for the development of digital biomarkers using mHealth and wearables data. Journal of clinical and translational science, 5(1), e19.

[5] Mazeas, A., Duclos, M., Pereira, B., Chalabaev, A. (2022). Evaluating the effectiveness of gamification on physical activity: systematic review and meta-analysis of randomized controlled trials. Journal of medical Internet research, 24(1), e26779.

[6] A. Ermakov, E. Doszhanov, E. Ismailova, and R. Amanov, "Analysis of user experience data and methodology of application to improve the development of user interface," Alatoo Academic Studies, no. 1, pp. 542–553, 2025, doi: 10.17015/aas.2025.251.44.

[7] Johnson, D., Deterding, S., Kuhn, K. A., Staneva, A., Stoyanov, S., Hides, L. (2016). Gamification for health and wellbeing: A systematic review of the literature. Internet interventions, 6, 89-106.

[8] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.

[9] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., . Wei, F. (2023, June). Trocr: Transformer-based optical character recognition with pre-trained models. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 13094-13102).

[10] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M. (2020, August). Layoutlm: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery data mining (pp. 1192-1200).

[11] Xu, Y., Liu, B., Chen, Z., Li, M., Wei, F., Zhou, M. (2021). LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. Proceedings of the 29th ACM International Conference on Multimedia (MM '21), 2484–2493. https://doi.org/10.1145/3474085.3475292

[12] R. Gumerov and R. Isaev, "Serverless Visual Voice Assistant Widget for Website Integration," 2026 18th International Conference on Electronics, Computer, and Computation (ICECCO), Kaskelen, Kazakhstan, 2026, pp. 1-5, doi: 10.1109/ICECCO67619.2026.11488758.

[13] S. Shaiakhmetov, R. R. Mekuria, G. Gimaletdinova, M. I. Gulbarga, Z. Orozakhunov, M. T. Khan, and A. Davletshin, "Criteria for Assessing the Quality of Educational Games and Their Impact on Student Learning Outcomes," Proc. Int. Conf. Comput. Syst. Technol. (CompSysTech), pp. 200–206, 2024, doi: 10.1145/3674912.3674925.

[14] S. G. Zaydullina, E. V. Druzhinskaya, and R. R. Isaev, "Training school students in developing mobile Android applications within the Samsung IT School course," Pedagogical Journal of Bashkortostan, no. 3(64), pp. 46–50, 2016.