Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia, 2026, Iss. 103, pp. 78–84

UDC 004.9:616-7

Infrastructure for the Real-Time Biosignal Datasets Collection

Mosiichuk V. S., Sharpan O.B., Yalosovetskyi V. I.

National Technical University of Ukraine "Ihor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

E-mail: *mvs@ros.kpi.ua*

The effectiveness of modern methods of biosignal analysis largely depends on the availability of structured data sets containing both primary signals and related metadata. At the same time, most existing biosignal registration systems do not provide automated accumulation of such data in centralized databases, which complicates the formation of training samples and the further application of machine learning algorithms. The article considers an automated system for forming representative sets of biomedical data to increase the efficiency of machine learning tasks in the analysis of the human physiological state. The infrastructure of the information system for automated collection of biosignals and metadata and formation of data sets in real time has been developed. The system has a multi-level architecture that includes a sensor level for registering biosignals, a communication level for data transmission, and a server level for their processing and storage. Signals were registered during system testing using a photoplethysmographic sensor integrated with a wireless module based on a microcontroller with Wi-Fi support. The server part is implemented in a virtualized environment using open source software, a web server, a database management system, and signal processing software modules. For visualization of biosignals and interaction with users, a Web-API and a web interface have been developed, which provide access to measurements, metadata management, and signal visualization. The system implements a data streaming pipeline that includes query verification, signal storage, and biomedical parameter calculation to assess the user's functional state. An experimental study of the system's performance was conducted in a load testing mode that simulates the simultaneous operation of a significant number of sensor devices. The results showed that the developed infrastructure is capable of handling more than a hundred simultaneous connections with an average query processing time of less than 100 ms. The results obtained confirm the possibility of using the proposed system for scalable collection of biosignals and the formation of data sets suitable for further application of machine learning methods.

Keywords: data collection automation; Internet of Things; biosignal registration, data streaming processing, functional state determination, dataset generation

DOI: 10.64915/RADAP.2026.103.78-84

Introduction

Advances in machine learning and neural networks are opening up new possibilities for medical diagnosis, particularly in the areas of disease screening and monitoring [1–3]. The effectiveness of such methods depends to a large extent on the availability of representative data samples that include both healthy patients and those with medical conditions [4–6]. At the same time, the quality of statistical conclusions is determined by the extent to which the sample under study corresponds to the population; in practice, this requires the collection of large volumes of diverse data and their systematic organisation. At the same time, there is a trend towards the creation of personalised diagnostics, which necessitates the creation of proprietary datasets of biosignals to enable the further training of artificial intelligence models that were originally developed based on a generalised set of biosignals.

In recent years, there has been a significant increase in the number of studies focusing on the use of wearable sensors and biosensor devices to monitor physiological parameters in everyday settings [3, 5–9]. Such devices are typically based on the Internet of Things concept [10–12]. This enables the recording of biosignals, supplemented with metadata regarding the user's condition, environmental factors or subjective assessments of their well-being. Machine learning models are built using this data to analyse a person's physiological state, in particular to assess stress levels or predict changes in well-being [13].

1 Statement of the Problem

At the same time, most existing approaches involve collecting data and then analysing it offline. Biosignals and corresponding labels (e.g. survey results) often

have different temporal resolutions, which complicates the construction of models with high temporal resolution. Furthermore, publicly available datasets are not always relevant to specific research tasks, which limits their potential for assessing an personal's functional state [14].

Traditional medical equipment does not usually support the automated creation of structured databases: biosignals are stored locally on diagnostic devices and require additional export, processing and conversion steps before they can be used in research. This significantly complicates the process of creating datasets and limits the scalability of experiments. Of course, this also hinders the creation of personalised datasets for the active retraining of neural network models [13].

In this context, there is a pressing need to develop information systems capable of ensuring the automated collection, storage and structuring of biomedical signals, the parameters extracted from them, and the associated metadata. Of particular importance is the ability to generate datasets in real time, which allows for the rapid accumulation of representative datasets for the subsequent application of machine learning transfer methods.

The aim of this article is to establish an information infrastructure that enables the direct transmission of signals from wireless sensor devices to a server, followed by their storage and processing, as well as the recording of metadata and the conditions under which the biosignals were recorded.

2 System Architecture and Real-Time Data Acquisition Method

Achieving this goal requires developing a system that forms a closed loop with three main components: data collection, annotation collection, analysis, parameter extraction, their analysis and training of neural network models. Also important are the presence of a dashboard and an administrator API that facilitate the processes of organizing and coordinating data collection according to the task of forming datasets.

The overall structure of the system consists of several layers (Fig. 1): a sensor layer for recording biosignals; a layer for sensor data transmission and user interaction; and a cloud layer for data processing, detection, and decision-making [11]. In our implementation, these layers correspond to sensor devices, a Smartphone/Wi-Fi router, and a server.

This architecture focuses on three goals: the ability to collect input data and analyze it quickly; the ability to scale in the data processing process; real-time operation.

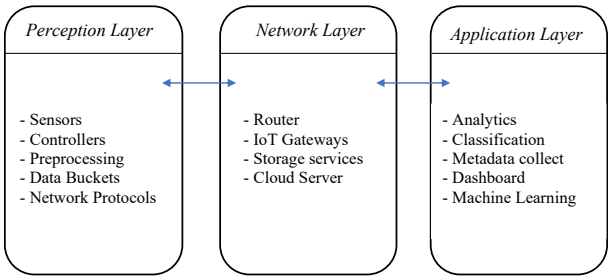

Fig. 1. Architecture of the information infrastructure for wireless biosignal registration

To create a semi-automatic sensor device for collecting user data and determining their functional state, it is advisable to develop a system that includes a server part for saving data from different devices to a single database (Fig. 2). In addition, the server should have services for training neural networks. Currently, it is most convenient to work with neural networks in the Python programming language. The system should also have an interface for adjusting and entering metadata that characterizes the biosygnal.

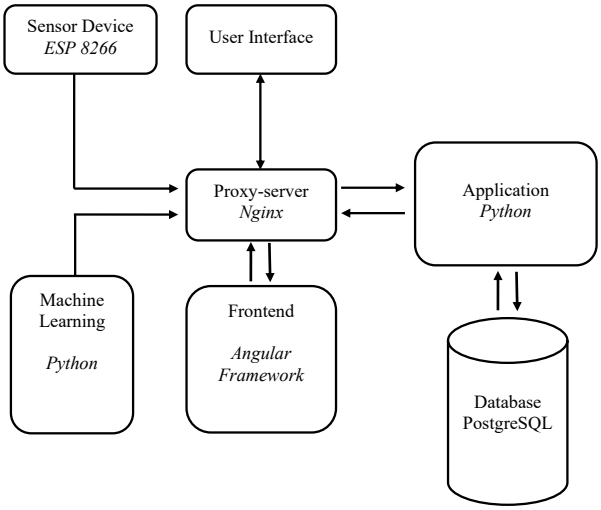

Fig. 2. Structure of the biosignal collection and analysis system

Data processing pipeline

Open-source software was used to implement the system's server infrastructure. This approach provides deployment flexibility, the absence of licensing restrictions, and the support of an active developer community. The system operates in a virtualized environment, which allows it to be quickly deployed on different hardware or migrated to cloud platforms.

The server operating system used is Ubuntu Server LTS, which is widely used for web infrastructure and server applications. Virtualization provides isolation of system components, simplifies scaling, and allows for efficient management of computing resources. The Nginx web server is used, which performs the functions

of a reverse proxy server, load balancing, and provides secure access to services via HTTPS.

Infrastructure deployment automation is implemented using Vagrant and Ansible tools. This approach enables the automatic generation of type-safe servers with the same software configuration, as well as manage network parameters, node IDs, and other configuration elements. The use of automation scripts ensures infrastructure reproducibility and simplifies its scaling.

The PostgreSQL database management system is used for data storage, which provides reliable storage and processing of streaming data from sensors. The server is configured taking into account performance parameters, including the placement of data files, transaction logs, and limits on the number of simultaneous connections. Server access rules are also configured.

During the testing of the system, a digital photoplethysmographic sensor MAX30102 with an I2C interface [1] was used to register biosignals. The sensor supports a sampling rate of up to 400 kHz and has an 18-bit analog-to-digital converter with the ability to adjust the intensity of the LED radiation. In standby mode, the device consumes only 0.7 µA, which makes it suitable for use in wearable systems with autonomous power supply.

To an automated data collection system, a sensor device based on the ESP8266 Wi-Fi module [12] was developed. The module contains a built-in wireless adapter supporting IEEE 802.11 b/g/n standards. The ESP8266 module supports basic network protocols, including IPv4, TCP, UDP, and HTTP, which allows it to be used to transmit data to the server infrastructure.

During the development of the device, various methods of data transfer to the server were investigated, including HTTP POST requests over TCP and packet transmission over the UDP protocol. Using UDP provides the ability to transfer up to 200 measurements per second, but requires a stable network connection.

The device transmits values from an 18-bit ADC as six-digit decimal numbers for two measurement channels. One UDP packet transmits 64 sampling samples, which is approximately 1 kB of data. This size was chosen experimentally taking into account the Ethernet MTU limitation (1500 bytes). Increasing the packet size can lead to data loss due to fragmentation, while decreasing the packet size reduces the useful information content and transmission efficiency.

To obtain data from a wireless sensor device, a service was created in the OS based on a program that collects fragments of data from sensors as they arrive and stores them in a database with automatic assignment of a unique identifier to each record. The data collected during the collection period was recorded in a signal table that contains the readings of each sensor with an indication of the time of their registration.

Fig. 3 shows an example of implementing an interface with the ability to automatically collect a database of a target group of patients with a description of their functional state.

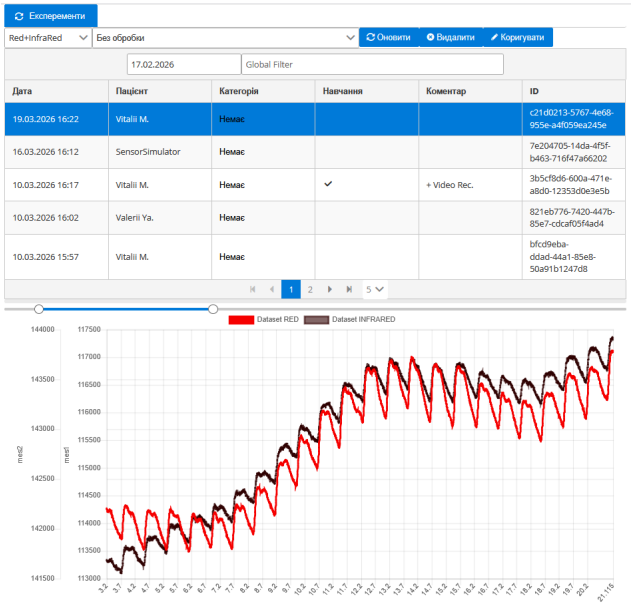

Fig. 3. Interface of biosignal registration mode

Algorithms for signal processing, data packet filtering, and neural network training are based on the Python 3 programming language. Its use is driven by the ability to rapidly develop prototypes and the presence of a developed ecosystem of scientific libraries for numerical computing and machine learning, including NumPy, SciPy, TensorFlow, PyTorch, Keras, and Matplotlib.

To process streaming data from sensor devices, a web service has been developed that implements a pipeline architecture for query processing. Each query contains a biosignal measurement session and a unique user identifier. At the initial stage, the authentication token is checked by comparing it with the user database data; queries from unauthorized users are rejected.

In the photoplethysmographic testing mode, the database stored the primary PPG signals in the red and infrared ranges, as well as the parameters obtained as a result of their analysis [15,16]. The main attention was paid to heart rate variability (HRV) indicators, which are used to assess the functional state of the user [4,6]. These parameters include: heart rate (BPM); mean interval between consecutive heartbeats (IBI); statistical characteristics of the variability of NN intervals (SDNN, SDSD, RMSSD); the proportion of NN intervals exceeding the corresponding threshold values (pNN20 and pNN50); mean absolute deviation of NN intervals (MAD); respiratory rate (BR).

The obtained parameters were used for further analysis of the physiological state and decision-making on the need for additional measurements or metadata collection.

Web-API system architecture

User interaction with the biomedical signal collection and analysis system is implemented using Web-API, which provides data exchange between the client web interface and server processing modules. The client part is built on the Angular framework using the PrimeNG component library. After compilation, the client code (HTML, JavaScript and static resources) is placed on the NGINX web server, which is responsible for servicing HTTP requests and delivering the user interface.

The server part of the system is implemented as a separate HTTP server, integrated with the web resource via API using REST-like requests. The client part interacts with the server using JSON format requests, receiving data for further visualization in a web browser in the form of tables, graphs, and indicators.

The Web-API architecture includes several functional groups of services:

- *Measurement access service.* Provides a list of registered biosignals, their filtering by time parameters or search attributes, as well as access to sensor data for visualization of biosignals.
- *Sensor data processing service.* Provides the ability to receive signal samples in a given time range with support for various display modes of the channels of the registered biosignals. At this stage, signal preprocessing algorithms can be applied, in particular, the extraction of the constant component of the signal, window filtering and data normalization.
- *Measurement monitoring service.* Designed to display the current status of the signal registration process and the formation of progress indicators during the measurement.
- *Metadata management service.* Provides work with user (patient) cards, categories for marking target groups and other auxiliary parameters necessary for the formation of datasets and training samples.
- *Data management service.* Implements the storage of measurement results and related metadata in the database, as well as the operations of editing and deleting records.

The proposed Web-API architecture provides system modularity, separation of client and server logic, as well as the ability to scale and integrate with biosignal processing modules and machine learning algorithms

A key performance indicator of the developed system is the delay between the registration of a biosignal by a sensor device, its transmission to the server, and receipt of the processing result. This delay determines the total time of the data processing pipeline and depends on the number of active users, which affects the loading of data processing queues.

The system performance was evaluated using the Tsung load testing tool, which allows emulating the simultaneous operation of a significant number of users and analyzing the behavior of web services and IoT applications under different load levels. During testing, the system receives requests that simulate typical interaction of sensor devices with the server. A gradual increase in the number of virtual users makes it possible to determine the maximum load at which acceptable performance indicators are maintained.

To simulate a typical device operation cycle with a data collection server, a test session was created that includes two main stages. The first stage involves registering the device with the assignment of a unique identifier, and the second stage involves transferring measured data to the appropriate processing channel.

The experimental study was conducted on a virtualized infrastructure using a virtual machine with 2 CPU cores and 2 GB of RAM with a web server and database. This architecture allows for isolation of system components and avoidance of resource access conflicts. The results of the testing stages are shown in Fig. 4.

The test results showed that the distributed information system deployed on two leading cloud providers, Digital Ocean (fig. 4a) and AWS (fig. 4b), is capable of handling over 100 simultaneous requests. The results indicate the possibility of operating up to 30 sensor devices in real time without the need to take measures to scale servers in the cloud environment.

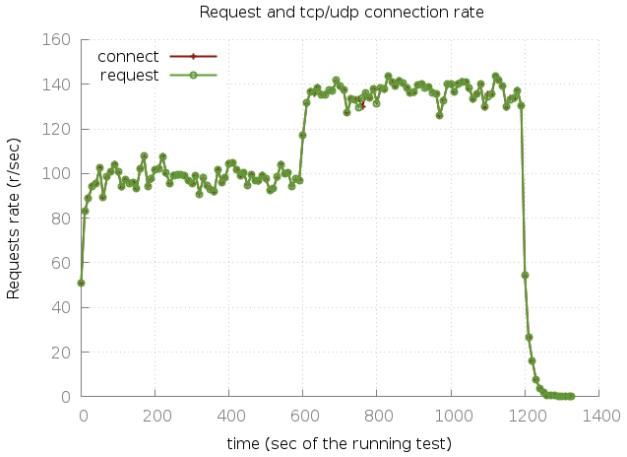

Fig. 4. Maximum number of requests per second during load testing

3 Discussion and Further Research

The obtained results confirm the feasibility of using the proposed information infrastructure for the automated generation of biosignal datasets in real time. Unlike most existing approaches, which focus on data collection followed by offline analysis, the proposed system implements an integrated ‘collection-processing-

annotation' pipeline, ensuring temporal consistency between biosignals and metadata.

Two approaches predominate in contemporary research into biosignals. The first involves the use of open datasets, such as the Pulse Transit Time PPG Dataset [17] and the BIDMC PPG and Respiration Dataset [18], which provide standardised samples, but are limited in terms of the specifics of a particular task, data collection conditions and sensor characteristics. The second type consists of experimental data collection systems using wearable devices, where data is collected over a specific period for subsequent analysis. Such systems generally lack a mechanism for real-time intervention in the data collection process [6].

The proposed approach is distinctive in that it allows not only for data collection but also for real-time management of the data acquisition process. By integrating the calculation of biosignal parameters with decision-making mechanisms, the system can initiate additional measurements or refine metadata precisely at the moments when this is most informative. In the current implementation, this is achieved through threshold rules for key physiological indicators; however, the architecture allows this mechanism to be extended to more complex models. The work most relevant to the approach proposed in this article is works [13,14], in which active reinforcement learning is used to diagnose an individual's functional stress state.

A key advantage of the proposed approach and infrastructure is the ability to rapidly assemble target datasets tailored to a specific task. Unlike static datasets, this enables the rapid planning of experiments and the collection of the necessary biosignals and corresponding metadata for a specific user, which is critical for the personalisation of machine learning models. This opens up the possibility of tailoring neural network models to a patient's individual characteristics, including physiological parameters, behavioural patterns and responses to external factors.

Furthermore, the proposed approach allows for the variability of sensor devices to be taken into account. It is well known that different types of sensors can generate signals with varying noise characteristics, sampling rates and sensitivity. The ability to quickly collect additional data for a specific device allows models to be adapted to its specific characteristics, which improves the accuracy of the analysis and reduces the impact of hardware differences.

Performance analysis has shown that the proposed architecture ensures low query processing latencies, even under heavy load. It was also established that the main factor affecting the total processing time is not the web server's response time, but rather delays in the data processing queues and database writes. This confirms the validity of using an asynchronous pipeline architecture and indicates the potential for further scaling of the system's computational components in particular.

However, the study has a number of limitations. In particular, the experimental evaluation was conducted in a controlled environment and does not cover the full range of real-world usage scenarios. Furthermore, the decision-making mechanism implemented for data augmentation is currently heuristic and requires further development using adaptive or learning strategies. Further research is also needed to quantitatively analyse the impact of the proposed approach on the quality of machine learning models compared with traditional static datasets.

Further research should focus on integrating active learning methods with reinforcement learning to manage the data collection process, as well as on conducting comparative experiments to assess the effectiveness of personalised models. A separate area of research involves investigating the impact of sensor characteristics on signal quality and developing methods for the automatic calibration of models.

Thus, the proposed information infrastructure facilitates the transition from the static accumulation of biomedical data to its controlled and adaptive generation, thereby laying the groundwork for the development of personalised systems for analysing physiological status based on machine learning methods.

Conclusion

A multilevel information infrastructure of the system for collecting and forming biosignal datasets in real time has been developed, which includes a sensor level of signal registration, a communication level of data transmission, and a server level of information processing and storage. Such a multi-level architecture provides centralized data accumulation, their synchronization with metadata, and the possibility of further use in personal physiological state analysis systems.

The conducted load testing confirmed the effectiveness of the proposed infrastructure for working with streaming biosignal dat. The experimental results showed that the system is capable of processing over 100 simultaneous requests and the ability to work in real time with up to 30 sensor devices simultaneously, which indicates the possibility of its application in scalable cloud environments.

The results confirmed the relevance of creating information systems focused on automated collection and structuring of biomedical signals together with metadata for further use in machine learning tasks. The proposed approach allows overcoming the limitations of traditional medical systems, in which biosignals are stored locally and require additional processing before forming datasets. The system can be used as a technological basis for forming representative datasets of biomedical signals and further application of machine learning methods for analyzing the functional state of a person.

References

[1] Mohan, D.C. and R, Y.R. (2026) Cloud-Based IoT and AWS Architecture for Real-Time Cardiovascular Patient Monitoring. *SSRG International Journal of Electronics and Communication Engineering*, vol. 13, no. 2, pp. 281-290, DOI: 10.14445/23488549/IJECE-V13I2P121

[2] K. Nguyen, H., & V. Pham, M. (2026) An edge AIoT system for non-invasive biological indicators estimation and continuous health monitoring using PPG and ECG signals. *International Journal of Reconfigurable and Embedded Systems (IJRES)*. Vol. 15, No. 1, pp. 97-108, DOI: 10.11591/ijres.v15.i1.pp97-108

[3] Neri, L., Oberdier, M. T., van Abeelen, K. C. J., Menghini, L., Tumarkin, E., Tripathi, H., Jaipalli, S., Orro, A., Paolocci, N., Gallelli, I., Dall'Olio, M., Beker, A., Carrick, R. T., Borghi, C. and Halperin, H. R. (2023) Electrocardiogram Monitoring Wearable Devices and Artificial-Intelligence-Enabled Diagnostic Capabilities: A Review. *Sensors*, 23(10), 4805. DOI: 10.3390/s23104805

[4] Jingye Xu and Yuntong Zhang and Wei Wang and Mimi Xie and Dakai Zhu (2025) A Comprehensive PPG-based Dataset for HR/HRV Studies. *arXiv*, eprint 2505.18165

[5] Baigutanova, A., Park, S., Constantinides, M. et al. (2025) A continuous real-world dataset comprising wearable-based heart rate variability alongside sleep diaries. *Sci Data* 12, 1474. DOI: 10.1038/s41597-025-05801-3

[6] Tasmurzayev, N., Amangeldy, B., Imankulov, T., Imanbek, B., Postolache, O. A. and Konysbekova, A. (2025) A Wearable IoT-Based Measurement System for Real-Time Cardiovascular Risk Prediction Using Heart Rate Variability. *Eng*, 6(10), 259. DOI: 10.3390/eng6100259

[7] Kim, K. B. and Baek, H. J. (2023) Photoplethysmography in wearable devices: a comprehensive review of technological advances, current challenges, and future directions. *Electronics* 12, 2923, DOI: 10.3390/electronics12132923

[8] Baigutanova, A., Park, S., Lee, S. W. & Cha, M. (2023) End-to-end system based on wearable devices for measuring HRV and its potential as an insomnia indicator. *J. Korea Inst. Inf. Sci. Pract.*, 29, 403-409. DOI: 10.5626/KTCP.2023.29.9.403

[9] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain and K. -S. Kwak (2015) The Internet of Things for Health Care: A Comprehensive Survey, *IEEE Access*, vol. 3, pp. 678-708, DOI: 10.1109/ACCESS.2015.2437951

[10] Li, Z., Xiao, H., Xia, Z., Zhou, F., Huang, X. and Liu, T. (2025) Three-Branch Network for Multi-Scale Spatiotemporal Feature Fusion in Remote Physiological Measurement. *IEEE Transactions on Consumer Electronics*, 71, 10270-10282. DOI: 10.2139/ssrn.5113756

[11] Ali, I. (2021) Data Collection in Studies on Internet of Things (IoT), Wireless Sensor Networks (WSNs), and Sensor Cloud (SC): Similarities and Differences. *IEEE Access*, 10, 33909-33931. DOI: 10.1109/ACCESS.2022.3161929

[12] Hercog, D., Lerher, T., Truntič, M. and Težak, O. (2023) Design and Implementation of ESP32-Based IoT Devices. *Sensors*, 23(15), 6739. DOI: 10.3390/s23156739

[13] A. Tazarv, S. Labbaf, A. Rahmani, N. Dutt and M. Levorato (2023) Active Reinforcement Learning for Personalized Stress Monitoring in Everyday Settings, *2023 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 44-55, DOI: 10.1145/3580252.3586979.

[14] Aqajari, S.A., Wang, Z., Tazarv, A., Labbaf, S., Jafarlou, S., Nguyen, B., Dutt, N.D., Levorato, M., and Rahmani, A.M. (2024) Enhancing Performance and User Engagement in Everyday Stress Monitoring: A Context-Aware Active Reinforcement Learning Approach. *ArXiv*, eprint 2407.08215. DOI: 10.48550/arXiv.2407.08215

[15] Yali Zheng and Chen Wu and Peizheng Cai and Zhiqiang Zhong and Hongda Huang and Yuqi Jiang (2023) Tiny-PPG: A Lightweight Deep Neural Network for Real-Time Detection of Motion Artifacts in Photoplethysmogram Signals on Edge Devices, *arXiv*, eprint 2305.03308. DOI: 10.48550/arXiv.2305.03308

[16] Moraes, J. L., Rocha, M. X., Vasconcelos, G. G., Vasconcelos Filho, J. E., De Albuquerque, V. H. C. and Alexandria, A. R. (2018) Advances in Photopletysmography Signal Analysis for Biomedical Applications. *Sensors*, 18(6), 1894. DOI: 10.3390/s18061894

[17] Mehrgardt, P., Khushi, M., Poon, S. and Withana, A. (2022). Pulse Transit Time PPG Dataset (version 1.1.0). *PhysioNet*. RRID:SCR_007345. DOI: 10.13026/jpan-6n92

[18] Pimentel, M.A.F. et al. (2016) Towards a Robust Estimation of Respiratory Rate from Pulse Oximeters. *IEEE Transactions on Biomedical Engineering*, 64(8), pp.1914-1923. DOI: 10.1109/TBME.2016.2613124

Інфраструктура для збору датисетів біосигналів у реальному часі

Мосійчук В.С., Шарпан О.Б., Ялосоветський В.І.

Ефективність сучасних методів аналізу біосигналів значною мірою залежить від наявності структурованих наборів даних, що містять як первинні сигнали, так і пов'язані з ними метадані. Водночас більшість існуючих систем реєстрації біосигналів не забезпечують автоматизованого накопичення таких даних у централізованих базах даних, що ускладнює формування навчальних вибірок та подальше застосування алгоритмів машинного навчання.

У статті розглядається автоматизована система формування репрезентативних наборів біомедичних даних для підвищення ефективності завдань машинного навчання при аналізі фізіологічного стану людини. Розроблено інфраструктуру інформаційної системи для автоматизованого збору біосигналів і метаданих та формування наборів даних у режимі реального часу. Система має багаторівневу архітектуру, що включає сенсорний рівень реєстрації біосигналів, комунікаційний рівень передачі даних та серверний рівень для їх обробки та зберігання. Реєстрація сигналів під час тестування системи виконано з застосуванням фотоплетизмографічного датчика, інтегрованого з бездротовим модулем на базі мікроконтролера з підтримкою Wi-Fi. Серверна частина реалізована у віртуалізованому середовищі з використанням програмного забезпечення з відкритим кодом, веб-сервера, системи керування базами даних та програмних модулів обробки сигналів.

Для візуалізації біосигналів та взаємодії з користувачами розроблено Web-API та веб-інтерфейс, які забезпечують доступ до вимірювань, управління метаданими та візуалізацію сигналів. Система реалізує конвеєр потокової обробки даних, що включає перевірку запитів, зберігання сигналів та розрахунок

біомедичних параметрів для оцінки функціонального стану користувача.

Експериментальне дослідження продуктивності системи було проведено в режимі навантажувального тестування, яке імітує одночасну роботу значної кількості сенсорних пристроїв. Результати показали, що розроблена інфраструктура здатна обробляти понад сто одночасних підключень із середнім часом обробки запитів менше 100 мс. Отримані результати підтверджують можливість використання запропонованої системи для масштабованого збору біосигналів та формування наборів даних, придатних для подальшого застосування методів машинного навчання.

Ключові слова: автоматизація збору даних; Інтернет речей; реєстрація біосигналів, потокова обробка даних; визначення функціонального стану; генерація наборів даних