Internet of Things 35 (2026) 101847

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot

Research article

Human-centered and context-aware smart ML-based IoT framework for online fatigue detection: A real-world study of football training

Abdelkarim Mamen [a], Elisabetta De Giovanni [b], Teodoro Montanaro [a], Ilaria Sergi [a], Luigi Patrono [a,*]

[a] Department of Engineering for Innovation, Università del Salento, Via per Monteroni, 73100, Lecce, Italy
[b] Basque Center for Applied Mathematics (BCAM), Alameda de Mazarredo 14, 48009, Bilbao, Bizkaia (Basque Country), Spain

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Deep learning
Internet of things
Framework
Athlete monitoring
Fatigue detection
Futsal
Sports
Dataset

ABSTRACT

Fatigue is one of the factors that most influences competitive athletes' performance, leading to injuries and overtraining. To effectively monitor and predict fatigue levels during real-world training, it is necessary to integrate Internet of Things (IoT) technology with machine learning (ML). In this context, the paper presents three main contributions : a) a smart IoT framework that integrates edge and cloud-based modules to collect physiological parameters, monitor fatigue during real-world sessions, and assist coaches in optimizing exercise strategies ; b) a dataset collected through the proposed framework in a real pilot study with eight futsal players over five training sessions, each lasting between 35 and 50 m depending on performed exercises, using ECG and PPG-based sensors ; c) an online ML-based fatigue detection module and on-cloud analysis of various ML models, traditional and deep learning, including CNN + GRU, XGBoost, and Transformer architectures, and context-aware feature sets. We evaluated the accuracy of our fatigue detection method using standard metrics, achieving an F1-score of up to 95 % with pilot study data. Our framework incorporates a context-aware design, where contextual information (exercise type) and sensing modality (ECG- or PPG-based) are explicitly integrated with physiological features (HRV and HR) in the fatigue prediction model to adapt it to different settings, improving robustness and interpretability. Finally, we evaluated the framework's efficacy and the value of user and expert input, highlighting the benefits of integrating IoT and ML within a human-centered, context-aware approach to balance sensor accuracy, comfort, and efficiency in competitive sports training.

1. Introduction

In competitive sports, elite athletes manage a fine equilibrium, striving for peak performance while carefully minimizing the risks of injuries and overtraining. In 2021, a study found that 62.9 % of professional football players prematurely retired due to injuries [1]. Therefore, it is crucial to prevent their occurrence while analyzing the factors that cause them [2]. As demonstrated in the literature, even though injuries are usually interpreted as unavoidable incidents, they are often connected to athletes' fatigue levels. This is

* Corresponding author.
E-mail addresses: abdelkarim.mamen@unisalento.it (A. Mamen), edegiovanni@bcamath.org (E. De Giovanni), teodoro.montanaro@unisalento.it (T. Montanaro), ilaria.sergi@unisalento.it (I. Sergi), luigi.patrono@unisalento.it (L. Patrono).

https://doi.org/10.1016/j.iot.2025.101847
Received 1 July 2025; Received in revised form 26 November 2025; Accepted 3 December 2025
Available online 8 December 2025

a state of physical exhaustion, often arising during competitions or as a result of prolonged and intense training sessions, which has been recognized as one of the most important factors impairing an athlete's performance [3]. Moreover, fatigue makes athletes more susceptible to injuries and overtraining [4], leading to a growing need to focus on fatigue detection, especially in the research field [5,6]. To this aim, Imran et al. [7] highlight the importance of monitoring fatigue in a real-world environment instead of a controlled one to capture its true dynamics. However, this remains a complex challenge in athletes' fatigue detection due to motion artifacts and users' comfort [8]. For this reason, there remains a significant gap in online detection of athletes' fatigue levels due to a lack of data during real-world training sessions [7,9]. Another relevant challenge is the ground-truth assessment of fatigue in real-world scenarios. The current measurements used in the literature are based on subjective perception of effort, such as the NASA-TLX, Karolinska Sleepiness Scale (KSS), and the Borg's Rating of Perceived Exertion (RPE) scale, which are prone to bias [7,10] or infrequent testing [11], and do not consider quantitative measurements or the coach's knowledge. To overcome these challenges, intelligent Internet of Things (IoT) systems are emerging as powerful tools for athletes and their coaches [12]. In fact, these systems combine wearable sensors and smart data analytics, through machine learning (ML) [13], to track key physiological parameters in real-world scenarios and support coaches in designing optimal training strategies to prevent injuries and overtraining [14,15]. By analysing prior research, we identified important limitations that restrict progress in this field (see Section 2). These include the limited availability of open and diverse datasets, the absence of standardized fatigue assessment metrics, and the need to validate existing models in real-world training scenarios.

For the reasons mentioned above, this paper presents three main contributions:

- a smart IoT framework to collect and process data from football players and to monitor their fatigue level during real-world training sessions. The framework seamlessly integrates advanced non-invasive wearable sensors capable of capturing a wide range of physiological parameters. The system combines edge and cloud-based modules, respectively, the former for data collection and inference, and the latter for advanced analysis of fatigue detection. The system design followed two main ensemble approaches: human-centered, that is, integrating the needs of the athletes and knowledge of the coach into the system; context-aware, that is, adding information about the exercise-related scenarios to the system design [16]. In our framework, context awareness is operationalized by explicitly modeling the relationship between physiological responses and situational variables. The contextual state includes the type of exercise, and the sensing modality (i.e., the type of wearable sensor and physiological signal, ECG or PPG). These factors influence both the physiological response and the reliability of the measured signals. For instance, the ECG-based (HRV) features generally yield higher model accuracy, but during high-intensity exercises, the chest-band sensor may lose contact due to displacement, whereas the arm-band PPG sensor remains stable and provides more consistent heart rate readings. This dynamic understanding of context allows the system to adapt its predictions, balance sensor accuracy and comfort, and ultimately support the coach in tailoring subsequent training sessions.
- a dataset collected through the proposed IoT framework in a real experimental pilot involving a local futsal team (a variant of football) during their regular training sessions. We expose the collected data for future interested stakeholders in a public dataset, which includes demographic data, vital parameters acquired from two types of commercial wearable devices based on electrocardiogram (ECG) and photoplethysmogram (PPG), and fatigue labels reported by the players through our system via the Borg's RPE. However, to fix bias issues, we allow the coach to correct the scores using their expertise and the collected physiological parameters, integrating their knowledge into the system. Specifically, the coach visually inspects graphs plotting heart rate trends to identify key inflection points that reflect players' actual perceived exertion and adjusts the reported scores accordingly.
- the integration in the IoT framework of an ML-based online fatigue detection module, which exploits its human-centered and context-aware approach with its analysis of the trade-off between performance of the algorithms, aided by the human knowledge, and feasibility of deployment (in terms of complexity and user needs) in different contexts of physical exercise. For this analysis, the system integrates a cloud-based module to train a wide range of ML models, traditional and deep learning (DL), and select features and models suitable to the context. Moreover, it includes an online prediction of fatigue on newly collected data on edge.

To demonstrate the efficacy of our proposed framework, we conduct an analysis through all the experiment stages, from data collection to user's feedback. The framework effectively enables the trade-off between accuracy, computational efficiency, and user comfort in predicting the athletes' fatigue level, achieving an F1-score of up to 95.0 %, in a preliminary study during real-world training sessions. Additionally, we show how the context directly affects the ML-based model accuracy, using feature selection. Specifically, if the exercise type feature is removed, the fatigue detection accuracy highly decreases, from an F1-score of approximately 95 % to 82 % on cross-validation results. Finally, by integrating data from wearable sensors, the framework provides online fatigue detection and feedback to the coaches.

The paper is structured as follows: Section 2 reviews the related works, and Section 3 introduces the proposed smart IoT system for fatigue detection via wearable sensors in real-world scenarios. Section 4 presents the data acquisition process used to validate the system and collect a new dataset. Section 5 describes the proposed fatigue detection methodology and optimization analysis targeted to real-world training. Section 6 presents the results obtained through the analysis of a wide range of ML techniques and feature sets on the newly collected dataset. Finally, Section 7 concludes the paper with key findings and future directions.

2. Related works

Intelligent IoT systems for online monitoring athletes have been already discussed and proposed in many existing works, however, only a few of them are focused on investigating fatigue as an indicator of the performances and the status of the players during real-world training sessions. Therefore, the present section analyzes two main categories of existing works. In the first category, we present

the works focused on frameworks for monitoring athletes and providing inferences based on ML or DL. In the second category, we present the most significant works focused on fatigue detection using a posterior analysis (not integrated into an IoT system that can predict fatigue online) via ML techniques to highlight the few results already obtained in the field.

Among the first category, we have identified two works whose scope and objectives are similar to ours, specifically regarding athlete monitoring and IoT-based solutions. However, these works remain conceptual without extending to practical development or testing. One of these works is presented by Balachandar et al. [5] which highlights the potential of the digital twin (DT) to monitor athletes in real time. A DT is a digital model of an intended or actual real-world physical product, system, or process, that exploits IoT and intelligent techniques, AI-based or more common mathematical models, to provide services, like our proposed framework. The proposed system leverages connected devices such as wearables, RFID tags, and sensors to gather players' data and offer coaches a tool for informed decision-making. Another similar approach is introduced by Ejiofor et al. [17] that proposes an IoT-based model to monitor athletes during physical exercise aiming to reduce health risks resulting from fatigue. It exploits a system that integrates heart rate data collection, heart rate data analysis, and a user interface to allow the results visualization. Lukač et al. [18] introduce a DT system that provides online feedback and guidance to athletes during training sessions. Despite implementing predictive algorithms to transmit training plans and monitor performance, the authors do not incorporate artificial intelligence (AI) techniques in their work, focusing instead on traditional computational and mathematical models for prediction. Beyond sports, Ding et al.[19] develop a smart framework to detect pilot mental fatigue using ECG data collected from wearable devices. Although the study is not focused on sport, it is mentioned due to its usage of commercial devices to monitor fatigue through Heart Rate Variability (HRV) during a one-hour flight simulation. In addition, it is worth mentioning that it exploits the Naïve Bayes model for fatigue classification.

The second category of works includes articles on advanced data analysis and the deployment of ML/DL for fatigue detection. The discussion will highlight the interest of the researchers in the challenges faced in identifying the onset and intensity of fatigue, as well as the lack of universally accepted methodologies and measurement standards for unobtrusive markers. Goodwin et al. [20] provide an analysis of blood lactate concentration ([La-] b) as an indicator of fatigue during exercise and emphasize the benefit of [La-]b as a marker of metabolic stress and exercise intensity. Unfortunately, the invasive nature of blood sampling required for [La-]b measurement limits its applicability in real-world scenarios. Hao et al. [21] conduct an experiment where volunteers performed weight-bearing walking under controlled conditions, carrying a load equivalent to 30 % of their body weight at a constant speed of 7 km/h. Their analysis reveals a variation in HRV metrics as fatigue increased. Buchheit et al. [22] explore the relationship between heart rate, exercise intensity, and fatigue in athletes, highlighting the utility of heart rate variability (HRV) as a non-invasive monitoring tool. They found a moderate to strong correlation between improvements in aerobic fitness and HRV indices suggesting that HRV can reflect both physical fitness and fatigue levels during training. Husom et al. [23] evaluate six ML models for fatigue detection using physiological data from a commercial device. Among the tested models, the Fully Connected Neural Network (FCNN) achieved the best performance. Liu et al. [24] proposed a medical intelligence framework for fatigue detection using photoplethysmography (PPG) signals combined with hybrid deep learning. Their model integrates ResNetCNN, Xception, and Bidirectional LSTM (BiLSTM) architectures to analyze physiological signals and classify fatigue and non-fatigue states in student athletes achieving a classification accuracy of 91.8 %. Gan et al. [25] proposed a heart rate variability (HRV)-based nonlinear analysis method for exercise-induced fatigue detection. Using a support vector machine (SVM) classifier, the method distinguished three fatigue levels, achieving an accuracy of 82.9 %. Guan et al. [26] employ a Bidirectional Long Short-Term Memory (Bi-LSTM) network to classify three levels of fatigue during running exercises. The classification is based on ECG signals, acceleration, and angular velocity. The Bi-LSTM model showcased its potential for online monitoring in sports applications. Caroppo et al. [27] present an approach for fatigue detection by combining physiological and activity data using a wearable device. Heart rate data and activity information are integrated using a rule-based expert system to compute a fatigue level score. Albert et al. [9] aim to predict perceived exertion during resistance training. They apply Gradient Boosting Regression Trees (GBRT) to the combined IMU and HRV features. However, this work is conducted in a controlled environment and focused solely on a single exercise, limiting its generalizability to more diverse training scenarios. Table 1 reports a comparative summary of the mentioned studies on fatigue detection with the aim of providing a graphical overview of the differences.

By analyzing the mentioned current state-of-the-art research, it is evident that the integration of data analytics and ML offers significant potential in detecting and monitoring fatigue. However, several research gaps were identified. At first, the discussion demonstrated the need for testing a wide range of advanced ML/DL models to evaluate their use in different physical exercise contexts and adapt their optimization to the user needs. In addition, a lack of existing datasets hinders further exploration in this domain. Furthermore, there are no established "gold" standard metrics to accurately detect fatigue, emphasizing the need for further research to establish reliable evaluation criteria. Lastly, most existing studies have been conducted in controlled environments, raising concerns about the generalization of their findings to real-world applications.

3. Smart IoT framework for fatigue detection in real-world scenarios

The innovative study presented in this paper is focused on detecting futsal athletes' fatigue during real-world training sessions. To this end, we designed a smart IoT framework that detects the athletes' fatigue and provides coaches and players with information about the latter's physiological status during the session. We designed the framework to:

- incorporate advanced sensors that can seamlessly capture a broad spectrum of physiological data without disrupting athletes;
- locally gather, process, integrate and visualize information coming from different types of sensors;

Table 1
Comparison of fatigue detection studies.

Study	Fatigue Metric	Model(s) Used	Sensor(s) Type	Activity Type	Limitations
[23]	FAS-scores	Random Forest, Gradient Boosting, Linear Regression	Fitbit wearable	Daily / occupational activity	No sport-specific validation (general population focus).
[9]	Borg RPE 6–20 scale	Support Vector Regression, Random Forest, Gradient Boosted Trees	IMU, ECG (HRV)	Squat exercise	Controlled environment; focused on one exercise.
[24]	Fatigued / Non-fatigued	ResNetCNN, Xception, BiLSTM, ResNet-BiLSTM, Xception-BiLSTM	PPG	Student sports / physical activity	Limited to lab data; focus on students, not trained athletes.
[26]	Low, medium, high	Bidirectional LSTM (Bi-LSTM)	ECG (HR, HRV), Inertial sensors	Running on treadmill	High computational complexity; lab treadmill setup.
[25]	Rested, slightly tired, tired	Support Vector Machine (SVM)	Polar H7, Polar V800	Smith squat and resistance training with blood flow restriction	Controlled experimental conditions; fixed exercise type.
[19]	Fatigued / Non-fatigued	Naïve Bayes (best), SVM, Random Forest, AdaBoost, KNN	Polar H10, Polar Ignite	Simulated flight (1 hour continuous mission)	Simulation study; not real-world.

- exploit cloud-based ML methods to train and optimize the models based on the context and user needs, later deployed for edge inference; ML locally provide the online fatigue detection inference during training sessions;
- expose the collected data to relevant stakeholders through a public dataset.

Figs. 1–3 present the designed architecture and the workflow of the proposed system, separating the ML training stage and the inference. The figures describe the different blocks, their role, and interactions in the context of online optimal detection of athletes' fatigue. As shown in Fig. 1, the proposed system encompasses three main blocks:

- Smart Device;
- Edge;
- Cloud.

The "Smart Device" block represents the wearable sensors responsible for capturing information from the athletes. The architecture has been designed to adapt to the integration of similarly interfacing commercial devices, which openly allow the acquisition of physiological signals and vital parameters in real-world athletes' training. In our experiments, the devices were divided into two main categories: electrical-based chest straps and optical-based arm bands. The data acquired from these sensors is then transmitted to the "Edge" block via wireless communication, through common protocols like the Bluetooth Low Energy (BLE). Our system can adapt to the integration of new devices in the future, thanks to its modularity. For our study, we selected the Polar H10 and Polar Verity Sense devices because they have demonstrated accuracy comparable to medical-grade equipment for measuring RR intervals and heart rate [28,29]. Additionally, they are specifically designed for sports, making them a reliable choice for athlete monitoring.

The "Edge" block has two distinct functions. First, it is responsible for collecting and preprocessing data, extracting heart rate variability and heart rate features, and transferring them to the cloud to create a training dataset. Later, once a trained model is available, the Edge performs inference to detect fatigue and support decision-making. At the end of the training session, the system transfers the results and newly collected data to the cloud for further model optimization and retraining.

In our implementation, the wearable sensors provided RR intervals and heart rate data at a sampling rate of one record per second. Data transmission between the sensors and the edge device was established through Bluetooth Low Energy (BLE), ensuring efficient and low-power communication. The edge-to-cloud connection was achieved using a Wi-Fi link through standard HTTP requests. To optimize bandwidth usage, the collected physiological and contextual data were synchronized and uploaded to the cloud at the end of each training session, enabling model retraining and performance benchmarking. This configuration ensures reliable data transfer, energy efficiency.

The "Cloud" block handles the ML model training, leveraging its substantial computational resources to generate predictive models that can be adapted to different contexts and user needs. For this reason, the cloud supports data and model analysis to identify the optimal models and groups of features for different scenarios. Once trained, these models can be deployed to the "Edge" block, where they run locally on edge devices. This transition from cloud to edge allows inference without relying on continuous cloud connectivity. In addition, the cloud provides storage for both raw data and computed features, enabling periodic evaluation and update of the ML models.

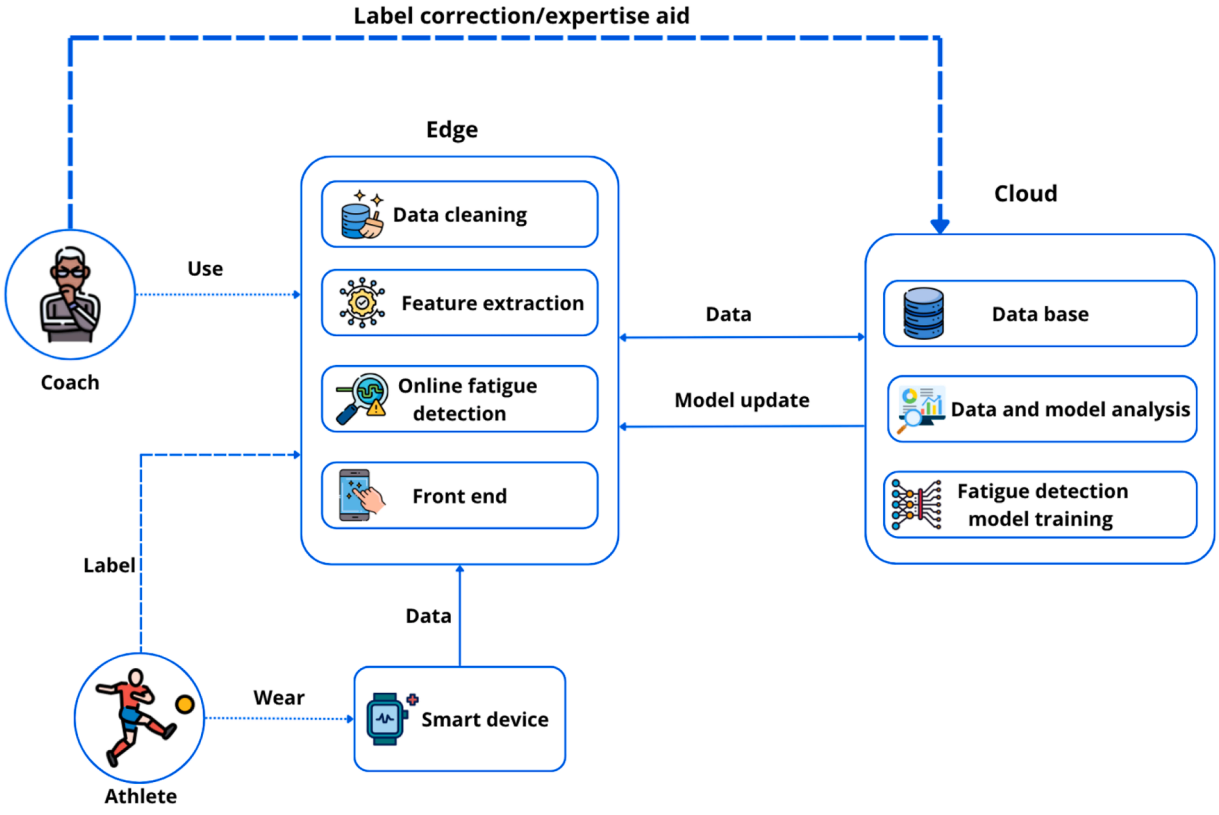

Fig. 1. Architecture of our proposed smart IoT framework for fatigue detection during real-world training sessions.

To clarify the role of each block and the way they interact with each other, Fig. 2 illustrates the system's workflow. The system handles three phases: data collection, edge processing, and cloud-based operations. The process begins with data collection, where the athlete's physiological parameters are continuously recorded using a smart device. This data, primarily in the form of time series, is then sent to the edge device, which processes the information. During this phase, the athletes provide subjective feedback regarding their perceived fatigue levels recorded on the edge device. At the end of the training session, the processed data is transmitted to the cloud, where it is securely stored in a database. The coach can also intervene at this stage by reviewing and correcting fatigue labels combining their expertise and the information from the collected time series. The cloud layer is responsible for model benchmarking, where various models are trained and optimized using the extracted features on edge. This phase includes evaluating different models, fine-tuning their parameters, The final model is selected manually by domain experts based on these benchmarking results and the specific operational constraints and based on the trade-off of complexity, accuracy, and user needs. After benchmarking, the optimal AI model is stored and deployed back to the edge device.

Fig. 3 illustrates the online athlete monitoring stage of the system workflow. This consists of three main phases: continuous online monitoring, cloud data storage, and periodic model updates. The first phase is the same for both workflows: the athlete engages in physical activity while a smart device continuously collects physiological data streamed to an edge device. The edge performs data preprocessing, which includes cleaning the time series and extracting relevant features. Then, the system classifies fatigue levels using the selected optimal model at training time. Moreover, the edge device provides online visualization of time series and fatigue predictions, allowing the coach to monitor the athlete's physical condition during the session, and provide feedback to adjust the training strategies. Once the session ends, the edge device transmits the newly collected data and fatigue predictions to the cloud for periodic data analysis and model update. The model updating ensures that the online fatigue classification adapts to real-world physiological changes between sessions.

4. Data acquisition and dataset creation

This section outlines the methodology and tools used to collect data during the study, focusing on the participants, the study setup, and the employed protocols. Finally, we describe the dataset that is made public for future investigations. We aimed to capture physiological data to study fatigue detection during sports activities. The data acquisition process was conducted in a real environment, integrating seamlessly with the routine training sessions of a local futsal team.

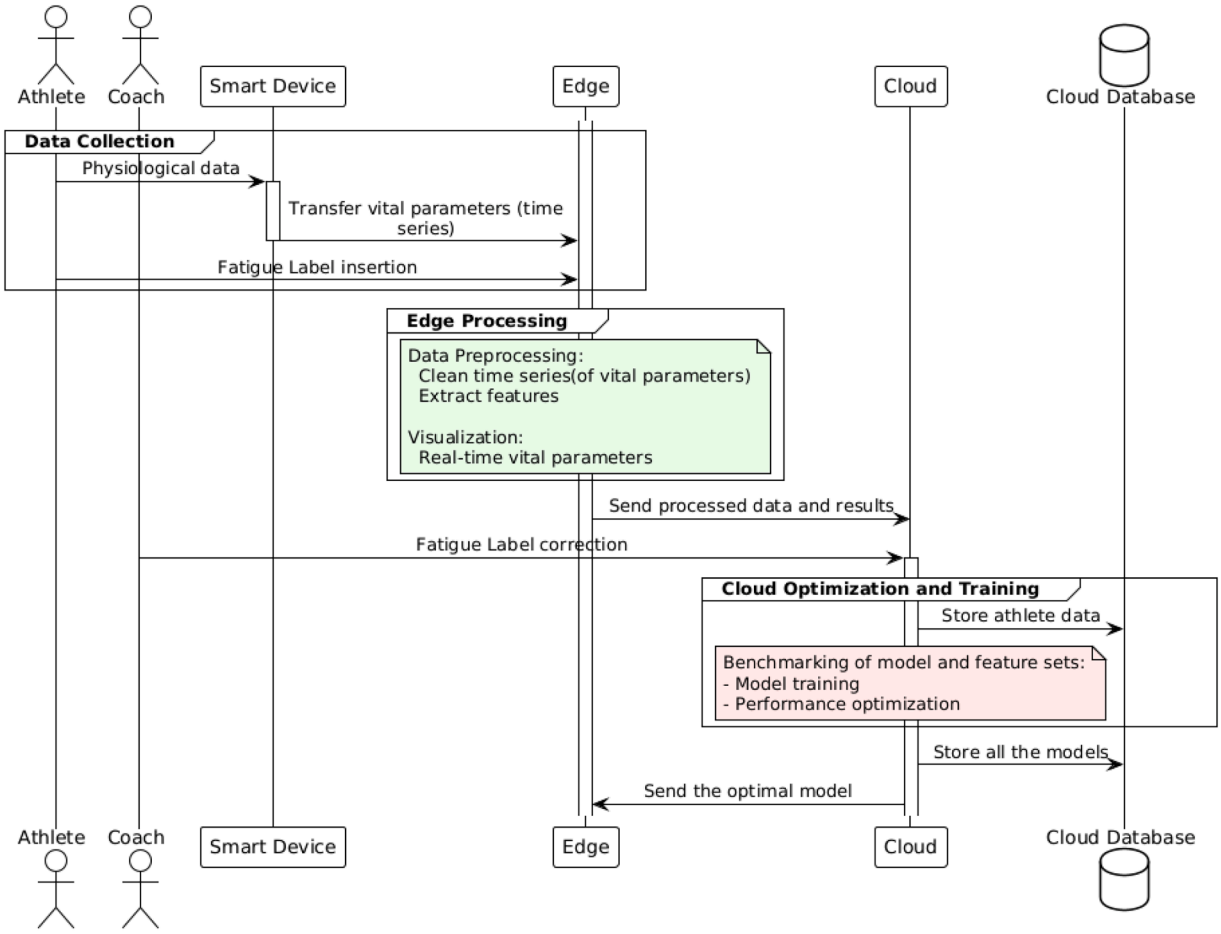

Fig. 2. Athlete monitoring system workflow including 1) data collection, processing, feature extraction and labeling on edge, 2) cloud-based label correction, model optimization, training and deployment, and storage of public dataset.

4.1. Participants

For our study, we focused on futsal, a variant of football played indoors on a smaller court with a smaller ball. In particular, after receiving the authorization from the Italian Ethical Committee through the protocol n. 1752/CEL - Studio IoTSport2024, we asked the Lecce futsal team to join our experiments. The experiments have been conducted in a real football field. Each participant signed a written informed consent describing the purpose, methodology, and minimum risks of the experiments, and they were explicitly informed about the implications of monitoring during training sessions. All the data has been anonymized and collected after the authorization of each involved stakeholder. The study involved eight healthy male adult athletes aged 30 to 47, showcasing a diverse range of athletic experience, physical fitness levels, and roles within the team (i.e., side, central, pivot).

Table 2 presents the collected demographic data for each participant, i.e., age, height, weight, smoking status, gender, and distinct role in the team.

4.2. Study setup and protocol

The data collection took place at the futsal stadium (“Stadio, Tensostatico Antistadio - A. Montinaro - Via del Mare, 73100 - Lecce”) where athletes train and compete regularly. In our study we exploited two types of wearable devices, arm band and chest strap, to compare their suitability in monitoring the players’ physiological responses. For this comparison, we took into account the accuracy of the output signals and vital parameters, the ease of use and comfortability, and the performance of our fatigue prediction methodology via different sets of extracted features related to each specific sensor. To this aim, we selected two devices produced by Polar, a well-established company in the sports domain, due to their characteristics: the Polar H10 [30] is a chest strap whose ECG signal and RR intervals have been validated against standard medical devices [29]; the Polar Verity Sense [31] is an arm band, whose main output parameter, the heart rate (HR), has been validated in several sports contexts [28]. Specifically, Polar H10 collects ECG

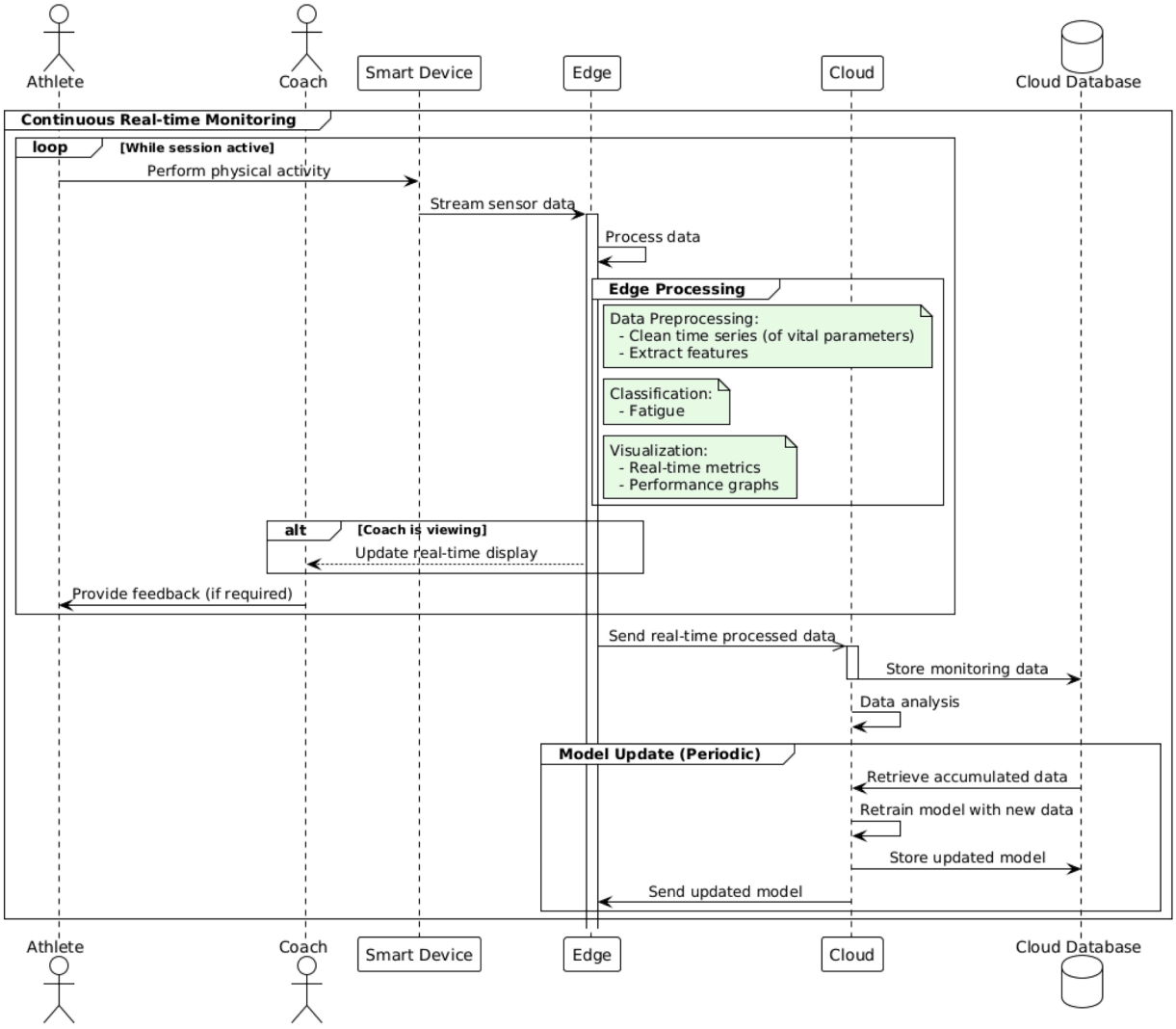

Fig. 3. Online athlete monitoring workflow. Physiological data are streamed from the wearable to the edge device, where preprocessing, feature extraction, and fatigue inference are performed. The edge also visualizes the signals for the coach. After each session, data and predictions are uploaded to the cloud for storage and periodic model retraining for continuous performance optimization.

Table 2
Player characteristics.

Player	Age	Height(cm)	Weight(kg)	Position	smoker
1	47	173	68	Side-central	yes
2	30	171	60	Side	no
3	47	172	68	Pivot	no
4	40	176	68	Central	yes
5	35	175	74	Side-pivot	yes
6	31	181	90	Central	no
7	41	174	71	Side	no
8	38	175	83	Pivot	yes

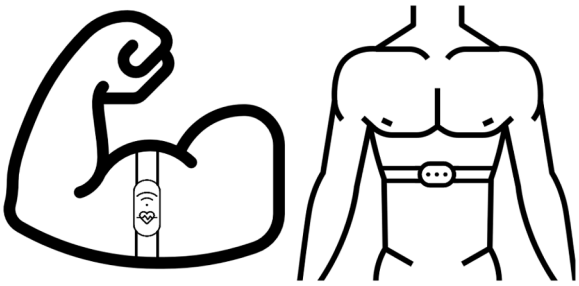

Fig. 4. Sensor Placement for ECG (chest band on the right) and PPG (arm band on the left) Measurements.

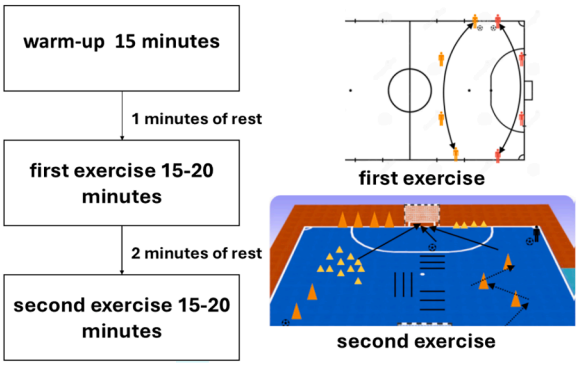

Fig. 5. Graphic representation of one training session from the study. The session begins with a general warm-up phase (15 m). After a 1-minute rest, the first exercise focused on metabolic work (15-20 m). After a 2 m rest, the second exercise starts, consisting of a dynamic shooting circuit (15-20 m).

data and computes the main R peaks and the consequent RR intervals every second. The Polar Verity sense collects PPG data and computes the pulse rate (i.e., HR) every second.

We placed the two sensors on the chest (ECG) and the arm (PPG), as shown in Fig. 4.

The training protocol was designed with the help of the team's coach and is reported in Fig. 5. The figure shows a graphic representation of one training session, specifically two main match-related exercises each preceded by a set of general warm-up. The protocol followed the coach's guidelines of exercise prescriptions based on the season and the following match to play. The 15 m warm-up included strength and cardiovascular exercises and 1 m of rest. Then, the first main exercise focused on metabolic work, which lasted 15 to 20 m according to the performance of each subject. After a 2 m rest, the second main exercise starts and is normally a dynamic shooting circuit of 15 to 20 m. To minimize disruption, at the end of each bout of exercise (warm-up, metabolic, circuit), we asked the athletes to estimate their perceived exertion levels using the Borg Scale (i.e., RPE), representative of their subjective feelings of fatigue (cf. Section 4.3). Later, we asked the coach to review these levels analyzing the physiological data acquired within each bout and using their expertise. The data collection was performed during five of the team's regular training sessions.

4.3. Collected data

In this study, we collected five main types of data from the athletes:

- Demographic information, such as age, weight, height, smoking status, gender, and the distinct role in the team;
- ECG-related data provided by Polar H10, specifically the HR and RR time series that we use for the heart rate variability (HRV) analysis;
- PPG-related data provided by Polar Verity Sense, specifically HR from the pulse;
- RPE as a subjective fatigue score declared by the athletes.
- Exercise identifier, uniquely associated with each exercise in the training session, used to link physiological and subjective data to specific exercise events. Each exercise identifier is mapped to a type of exercise that describes the activities performed by the athletes.

In addition, the coach's expert evaluations were used to validate and refine the athletes' self-reported RPE values, thereby improving consistency and objectivity in the dataset. Drawing on their background knowledge and experience with the team, the coach assessed whether the reported levels of perceived exertion were consistent with the expected physiological responses to each exercise intensity. This evaluation involved visually examining the heart rate trends recorded during the sessions and correlating them with the RPE

Table 3
HRV and HR parameters.

Category	Parameters
HRV Time Domain	SDNN, RMSSD, MinNN MaxNN, SDNNd, SDNNa
HRV Non-linear Domain	SD1, SD2, SD1SD2, ApEn DFA_alpha1, MFDFA_alpha1_Width MFDFA_alpha1_Peak, MFDFA_alpha1_Mean MFDFA_alpha1_Max
HR Parameters	% HRR, HR Index, HR Net, VO_2

scores reported by the athletes for each segment of the session. Through this process, the coach was able to confirm or adjust the scores to better reflect the athletes' effort.

In our work, the physiological data collected from the wearable sensors represents the key information to define the fatigue detection framework. Therefore, it is necessary to describe this data in detail. The Polar H10 chest strap acquires the ECG signal as a measure of the electrical activity of the heart. Moreover, it computes the RR intervals, defined as the time intervals between successive R peaks, and the HR. The R-wave represents the highest point of the QRS complex of an ECG, which describes the ventricular depolarization during the heart's electrical cycle [32]. From these RR intervals, we can derive the HRV, which is a physiological marker regulated by the autonomic nervous system and, specifically, is highly related to fatigue [33–35]. The Polar Verity Sense is an arm band with an optical sensor that detects changes in blood volume by measuring the light reflected by the tissues (i.e., PPG signal). From the PPG, the Polar Verity sense automatically extracts the HR from the pulse, which can similarly give information about the fatigue level of an athlete [36]. The device is generally more comfortable than a chest band, although the PPG signal is more prone to motion artifacts than the ECG[37]. The raw physiological signals are first acquired and preprocessed on the wearable devices, which extract key features such as RR intervals and heart rate, as implemented by the manufacturer. These data, are then transmitted to the Edge block, which performs preprocessing and send the data to the cloud, where more computationally demanding tasks, such as long-term trend analysis, and model retraining are performed.

In addition, with the aim of having ground-truth label of our experiments, we employed the Borg Scale (RPE), a subjective scale ranging from 6 ("no exertion at all") to 20 ("maximum exertion") used by the players to self-assess their perceived effort during physical activity [38]. The Borg Scale provides a practical method for evaluating exercise intensity by focusing on overall effort, complementing objective measures like HR. Its simplicity and standardization make it a widely accepted measure in sports for tracking perceived exertion and fatigue during training.

4.4. Public dataset description

Our dataset is available at [39]. It includes the demographic information, physiological data, and fatigue collected from the eight participants of our study while performing their regular training sessions. The data has been anonymized according to the European General Data Protection Regulation (GDPR) [40].

The physiological data includes:

- Heart Rate (bpm): Recorded in beats per minute, it captures the participants' cardiac response during training sessions. The dataset includes two different HR values: one automatically extracted by the Polar H10 device from the ECG signal, and one automatically extracted by the Polar Verity Sense device from the PPG signal.
- RR Intervals (ms): Measured in milliseconds, they represent the time between consecutive heartbeats. The values reported in the dataset are automatically extracted by the Polar H10 device from the ECG signal.
- Heart Rate Variability (HRV) features: they are established biomarkers to evaluate the response of the autonomic nervous system to training in both individual and team sports [35]. They can be grouped into two different categories: Time and Non-linear domains. Time domain features quantify the amount of variability in measurements of the interbeat interval (IBI), which is the time period between successive heartbeats, while nonlinear methods assess both short-term and long-term variability, providing insights into autonomic balance [41]. In HRV analysis, there exist another category of features, the frequency domain one. However, for our specific scenario we chose not to extract them as we aimed to analyze windows shorter than one minute, which the tool used required as minimum length for the frequency domain features [42]. Nonetheless, as we provide the RR time series, future researcher can extract different types of features. Table 3 summarizes the HRV features included in our dataset.
- Heart Rate features: offer valuable insight into how the body responds to physical activities and stress, thereby acting as reliable indicators of fatigue. By examining variations in HR at rest, during exercise, and at maximum effort, several derived values can be analyzed. The Heart Rate Reserve (%HRR) highlights the percentage of the available HR range used during activity [43], The HR Index which is the ratio of the current HR to resting HR [44], Net HR the difference between current and resting HR [44], Lastly, estimating VO_2 from HR data leverages equations that map HR Index to oxygen consumption, providing an indirect yet effective way to track aerobic demand without cumbersome gas-analysis equipment [45]. Those features are summarized in Table 3.

The demographic data encompasses:

- Age, Weight, Height, and Gender: they describe the physical characteristics of participants;
- Smoking Status: indicates whether participants are smokers or non-smokers at the beginning of the study;
- Position: it represents the distinct role of the player in the team.

The exercise data includes:

- Exercise ID: uniquely identifies each exercise performed by an athlete, allowing understanding of the exercise intensity and context associated with the recorded physiological data. Each exercise ID is mapped to a type of exercise performed during each training session.

As the ground-truth labels for the dataset (i.e., fatigue annotation), we collected the fatigue data in the form of the RPE declared by the athletes, based on the Borg scale during physical activity [46]. However, after a comprehensive analysis of the collected data, we found out that these subjective ratings were often influenced by personal biases, leading to underestimation or overestimation of fatigue levels. Therefore, to resolve those biases, the coach was asked to confirm or adjust the athletes' subjective ratings by analyzing heart rate plots from each session based on their expertise. The final fatigue scores reported in the dataset are the values declared by the athletes and adjusted by the coach. Finally, to enhance analysis and address data imbalance, we mapped the Borg scale ratings into four levels of fatigue. The original Borg scale uses a score range from 6 to 20, and we mapped it to a 4-level categorical scale to improve interpretability, following the correspondence between the 6–20 scale and the 4-level physical fatigue scale reported in the literature [47]. The resulting categories are low, moderate, heavy, and severe. These levels are used as categorical labels for the supervised fatigue classification process, mapped as low = class 0, moderate = class 1, heavy = class 2, and severe = class 3.

5. Fatigue detection in real-world training

This section provides an overview of the preprocessing steps, feature extraction methods, and ML models used to detect fatigue in real-world training scenarios.

5.1. Data preprocessing

Data preprocessing is an important task to ensure the reliability of results. The key preprocessing steps include detecting and correcting data errors in time series to minimize error propagation during analysis, as detailed by Kandel et al. [48]. These steps involve handling missing values, standardizing formats, and detecting outliers. Specifically, Radivsic et al. [49] highlighted that if noisy data is not properly handled, results may be biased, which undermines the credibility and applicability of the findings. One relevant issue in our dataset was the presence of outliers, especially in RR intervals and HR measures. These outliers may be caused by algorithm misdetection (integrated into the specific sensor), particularly with R peaks and HR calculations. Additionally, they could have been related to sporadic physiological phenomena, like ectopic beats (i.e., disturbances in the cardiac rhythm of a non-pathological nature). Finally, in the context of real-world scenarios, the outliers could be linked to device connection issues, or sensor misplacement, causing episodic inaccurate measurements.

All preprocessing operations were performed at the edge, allowing for on-device correction and ensuring that only clean and reliable data were transmitted to the cloud for further analysis and model training. To correct these outliers, RR interval signals were first processed using the Interquartile Range (IQR) method to detect and remove abnormal values while preserving the integrity of the dataset [50]. This method identifies outliers by calculating the spread between the 25th and 75th percentiles (Q1 and Q3) and flagging observations outside the interval $Q1 - 1.5 \times IQR$ or $Q3 + 1.5 \times IQR$. Next, we applied linear interpolation to maintain the main trend of the RR or HR series. This correction was followed by applying two standard filtering methods: a median filter [51] (kernel size = 9) and a second-order low-pass Butterworth filter [52]. The median filter smooths the signal by sliding a window (kernel) across the data and replacing each point with the median value within that window. This approach reduces isolated outliers while preserving the underlying structure of the RR-interval time series [51]. Finally, we applied a second-order low-pass Butterworth (cutoff = 0.5 Hz) filter [52] to remove high-frequency noise and preserve meaningful variations.

5.2. Feature extraction

To assess fatigue, it is essential to extract relevant features that serve as quantitative indicators of physiological states. In this study, we focused on testing features derived from the RR series, that is HRV, and the HR series, following the existing literature. These features were extracted at the edge, reducing the need for continuous data transmission to the cloud. To extract the features, we use a window of analysis of 30s, with a sliding window of 10s, following the guidelines of short-term HRV analysis [41]

5.2.1. HRV features

We extracted the HRV features using the NeuroKit2 toolbox, an open-source Python package intended for neurophysiological signal processing [42]. Specifically, NeuroKit2 provides methods to filter ECG signals, extract their R peaks and RR intervals, and perform HRV analysis. we chose to extract time-domain and non-linear HRV parameters for short-window analysis. This choice was made because it enables the detection of small changes [53,54]. Table 3 in Section 4.4 shows all the HRV features extracted using the NeuroKit2 tool.

Table 4
Tested AI models categorized into machine learning and deep learning approaches.

Approach	Tested AI models
Machine learning	SVM, XGBOOST
Deep learning	LSTM, Transformer, CNN + LSTM, CNN + GRU

5.2.2. HR features

For deriving features related to the HR, we have adapted some recent works to apply established equations and verified techniques. One of the most important methodologies used in this work is the formula developed by Martti J. Karvonen [43], which computes the percentage of heart rate reserve (% HRR), described in Eq. (1).

$$\%\ \text{HRR} = \left(\frac{\text{Actual HR} - \text{Resting HR}}{\text{Maximal HR} - \text{Resting HR}} \right) \times 100 \tag{1}$$

This equation is useful in determining exercise intensity levels personalized to the specific fitness conditions of individuals, represented by the HR at rest and the HR at maximal effort.

Furthermore, we used methods proposed by Wicks et al. [44], who introduced the Eq. (2), to calculate the HR Index, that is the ratio of absolute HR to resting HR. This ratio is a simple method to estimate energy expenditure via HR measurements. He also defined the net HR, shown in Eq. (3), that is the difference between absolute (i.e. at the lowest activity) and resting HR. The performance of HR Net to estimate energy expenditure was lower in Wicks's findings compared to HR Index, although we chose to test it on our data regardless to avoid bias.

$$\text{HR Index} = \frac{\text{HR Absolute}}{\text{HR Rest}} \tag{2}$$

$$\text{HR Net} = \text{HR Absolute} - \text{HR Rest} \tag{3}$$

To integrate the information extracted from the two sensing modalities, we applied a multimodal sensor fusion step before model training. Specifically, we adopted an early-fusion strategy, in which the ECG-derived HRV features and the PPG-based HR features were concatenated into a single unified feature vector. This fused representation was then used as the input to the machine-learning models during training and inference.

After analyzing our feature set, we found a class imbalance, such as we had more samples in low-level fatigue than in extreme fatigue. To address this imbalance, we employed an undersampling technique to have a balanced sub-dataset, despite reporting all the data in our public dataset [55]. This approach involved randomly selecting a representative subset of samples from the majority class to match the size of the minority class, preserving the integrity of our dataset. Given that our project involves fatigue detection using physiological signals, the patterns and relationships within the data are highly dependent on real-world variability. Synthetic generation methods may struggle to accurately replicate these complex patterns, potentially leading to distortions that compromise the model's reliability. Undersampling allowed us to achieve a balanced dataset while maintaining the authenticity and validity of the data. After balancing the dataset, we proceeded with data normalization to ensure that all features contributed equally to the model.

5.3. Model training and inference

The present subsection describes the methodology performed to a) integrate ML algorithms in the system to predict fatigue, b) train the models with different feature sets and evaluate their impact in terms of prediction accuracy, and c) evaluate the benefits that each combination of ML model and feature set can provide to the selected scenario. These operations are carried out in the cloud environment, which provides the necessary computational resources for data aggregation, model training, validation, and performance analysis before deployment to the Edge block for inference. The trained models were transferred from the cloud to the edge via a Wi-Fi connection using standard HTTP requests, enabling seamless synchronization between the two environments.

The framework has been exploited to test and compare different emerging DL models alongside traditional ML models. We evaluated the validity of these models considering complexity and accuracy for fatigue level classification. Moreover, we analyzed this trade-off considering the two sets of HRV and HR features, which directly correlated with the selected wearable devices. As fatigue classification is inherently a time series problem exhibiting both short-term and long-term dependencies, we analyze the DL models that handle time series.

Table 4 shows all the tested AI models. We selected four DL models: Transformer, LSTM, CNN + LSTM, and CNN + GRU. Additionally, we incorporated two traditional ML models, support vector machine (SVM) and XGBoost, following the existing literature and providing a comparative analysis between DL and ML. During the validation process, we optimized the models' hyperparameters to maximize performance.

The deep learning models were trained for 150 epochs with a batch size of 64 using the Adam optimizer, set with a learning rate of 0.00015. These Training hyperparameters were selected based on preliminary experiments that balanced training stability

Table 5
Summary of model hyperparameter settings for all trained models.

Model	Model Hyperparameters
XGBoost	max_depth = 6 n_estimators = 12 eval_metric = 'logloss' use_label_encoder = False
SVM (RBF Kernel)	kernel = 'rbf' γ = 'scale' C = 1 probability = True
LSTM	units = 128 input_shape = (30, 9) Dropout = 0.5
CNN + LSTM	Conv1D(filters = 64, kernel_size = 3, activation = 'relu', input_shape = (30, 9)) Dropout = 0.3 LSTM(128) Dropout = 0.5
CNN + GRU	Conv1D(filters = 64, kernel_size = 3, activation = 'relu', input_shape = (30, 9)) Dropout = 0.3 GRU(128) Dropout = 0.5
Transformer	head_size = 32 num_heads = 2 ff_dim = 16 num_transformer_blocks = 1 mlp_units = [128] mlp_dropout = 0.2 dropout = 0.2

and convergence speed. A learning rate of 0.00015 provided smooth optimization without overshooting local minima [56], while a batch size of 64 offered an effective trade-off between computational efficiency and gradient stability [57]. The number of epochs (150) ensured sufficient learning without overfitting [58]. The categorical cross-entropy loss function was employed to optimize model performance on multi-class classification tasks. Evaluation was conducted using multiple metrics, including accuracy, AUC, precision, recall, and F1-score, to ensure a comprehensive assessment of the model's predictive capability and generalization performance. The machine learning models were configured and trained using both ensemble and kernel-based approaches. The XGBoost classifier (`XGBClassifier`) was initialized with a maximum tree depth of 6, 12 estimators, and `logloss` as the evaluation metric. Additionally, a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel was implemented in a one-vs-rest scheme to handle multi-class classification. The SVM was configured with a regularization parameter C = 1, kernel coefficient γ = scale, and probability estimates enabled for calibrated output. Table 5 summarizes the Model hyperparameters used for each machine learning and deep learning model during training.

Beyond model evaluation, we explored various feature groups, to identify the optimal feature selection for fatigue detection. Additionally, as these groups are directly related to the selected devices, we conducted a comparative analysis of the two separate sets of features (HRV from Polar H10 and HR from Polar Verity Sense) to assess the impact on model accuracy and determine in which context they can be used independently. This approach is particularly relevant to select the most suitable device for practical applications, as an arm band, such as the Polar Verity Sense, is widely accessible and affordable, and for specific settings more comfortable than a chest band.

6. Results

In this paper, we studied the athletes' fatigue during training sessions by designing and developing an IoT framework to support trainers and players. This section analyzes the fatigue prediction performance in different aspects.

To evaluate the proposed models, we used standard classification metrics, including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). These metrics provide a comprehensive assessment of predictive capability across all classes. The metrics are computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \tag{8}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively, and TPR and FPR represent the true positive rate and false positive rate.

First, we present the impact of applying the label correction of the expert (i.e., coach) on the model performance. Next, we examine the impact of feature selection on model performance and computational complexity, considering the specific features and feature groups presented in Section 5.2. Then, we analyze the different traditional ML and DL models presented in Section 5.3. To highlight the trade-off between accuracy and complexity of the prediction in the test set, we use F1-score, as the optimal metric of performance for unbalanced datasets, floating point operations (FLOPs) for the ML models, and big O notation for the features, as metrics of computational complexity. Nonetheless, we report additional standard performance metrics for each combination of model and feature group, including precision, recall, accuracy, and loss [59]. Moreover, to analyze the discriminative power of the models for each class, we plot the Receiver operating characteristic (ROC) curves and compute the Area Under the Curve (AUC) using two

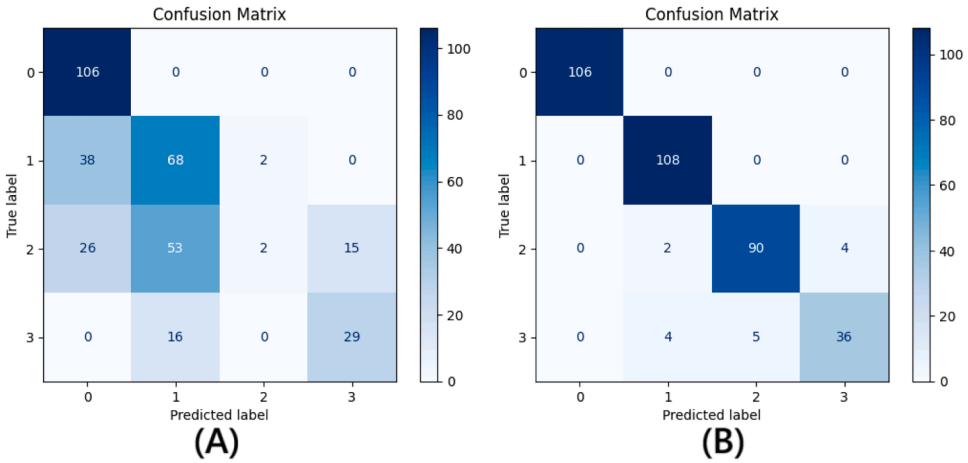

Fig. 6. (A) Confusion matrix of the XGBoost model trained with labels assigned only by athletes, using HRV time-domain features. (B) Confusion matrix of the XGBoost model trained with labels corrected by the coach, using HRV time-domain features.

Table 6
Detailed class-wise and macro-averaged performance of XGBoost models with different labeling sources.

Label Source	Class	Precision (%)	Recall (%)	F1-score (%)
Coach	Class 0 (low)	100	100	100
	Class 1 (Moderate)	95	100	97
	Class 2 (Heavy)	95	94	94
	Class 3 (Severe)	90	80	85
	Macro avg	**95**	**93**	**94**
Athletes	Class 0 (low)	62	100	77
	Class 1 (Moderate)	50	63	56
	Class 2 (Heavy)	50	2	4
	Class 3 (Severe)	66	64	65
	Macro avg	**57**	**57**	**50**

approaches: one-vs-all and one-vs-one. In addition, for the model complexity analysis, we measure execution time and memory usage, on a system equipped with a 12th Gen Intel® Core™ i7-1255U (12 cores), 16.0 GiB of RAM, and OS Ubuntu 24.04.2, the same used for computing the FLOPs. Next, we evaluate the usability and effectiveness of the selected wearable devices. Finally, we assess the overall efficacy of the framework, emphasizing practical deployment considerations.

6.1. Analysis of expert label correction

To evaluate the impact of label quality on model performance, we conducted a comparative analysis using two XGBoost classifiers trained on fatigue level data annotated by different sources. The first model was trained using self-reported fatigue levels from athletes, while the second used annotations validated by expert coaches. Both models were subsequently evaluated on a test set labeled by coaches, which served as the gold standard due to the higher reliability and consistency of expert assessments. This setup allowed us to quantify how discrepancies in data labeling affect the generalization performance of the predictive model.

The model trained on athlete-provided labels exhibited considerable misclassification, particularly between fatigue class 0 = Low and class 1 = moderate, as well as between class 0 = low and class 2 = heavy, as shown in the confusion matrix in Fig. 6 (A). Most notably, class 3 = sever, representing high fatigue, was frequently misclassified, indicating a strong presence of label noise in self-reported data. This noise is likely attributable to subjective variability in fatigue perception, which compromises model generalization and degrades inter-class discriminability. Conversely, the model trained using coach-validated labels demonstrated significantly improved classification accuracy across all classes, as shown in the confusion matrix in Fig. 6 (B). Predictions closely matched the ground truth, especially in classes 0 through 2, with minimal errors, and class 3 improved significantly compared to the model trained with athlete-provided labels. The model trained on athlete-labeled data achieved an overall F1 score of 50 %, whereas the model trained on coach-labeled data achieved 94 %, as shown in Table 6, with expected acceptable lower performance in class 3. This underscores the model's ability to learn more robust decision boundaries when trained with labeled data corrected by experts.

Future research could explore automated label correction techniques grounded in physiological signal patterns and domain knowledge, potentially reducing reliance on manual expert annotation while maintaining high model performance.

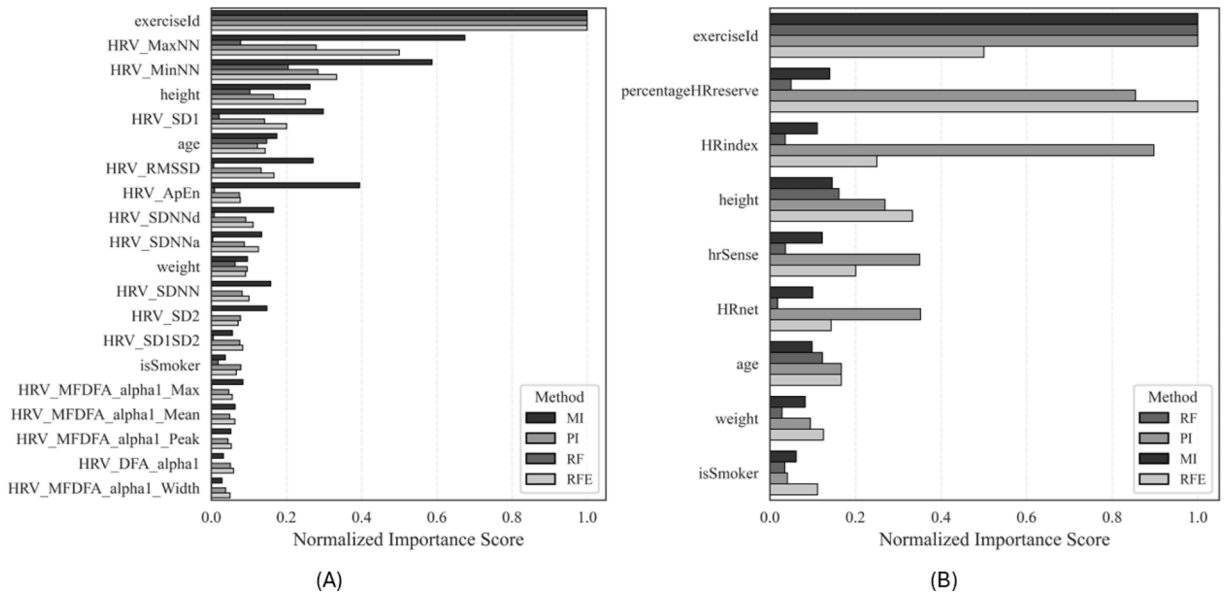

Fig. 7. Comparison of normalized feature importance scores for fatigue prediction across four methods: Mutual Information (MI), Permutation Importance (PI), Random Forest (RF), and Recursive Feature Elimination (RFE). (A) HRV features from ECG signal; (B) HR features from PPG signal. Both include the contextual feature exercise type.

6.2. Analysis and comparison of feature sets

The choice of feature sets significantly influenced the models' performance, demonstrating the role of feature engineering. Our study utilized data from two types of devices: ECG-based and PPG-based devices, testing various combinations of HRV features and HR features to identify the most effective and efficient feature sets, for specific contexts.

To evaluate the contribution of each feature in predicting the fatigue level, four feature importance techniques were applied: Mutual Information (MI), Random Forest (RF) importance, Permutation Importance (PI), and Recursive Feature Elimination (RFE). Among the four feature selection methods evaluated in Fig. 7, the Permutation Importance (PI) method achieved the most reliable results for fatigue prediction. PI effectively identified the most physiologically meaningful variables, such as %HRR, HR Index, and HRVMaxNN, reflecting their strong association with fatigue levels. By directly assessing the impact of each feature on model performance, PI provided a more interpretable and unbiased ranking compared to the other methods. Hence, it was considered the most appropriate technique for selecting relevant features in the proposed fatigue detection framework.

As shown in Fig. 7 (A) and (B), the feature representing the type of exercise (exerciseId) is continuously ranked as the most influential element. This dominance highlights a contextual dependency in fatigue modelling and emphasizes how the type of exercise session, and consequently its intensity, affects the fatigue levels. In Fig. 7 (A), "HRV Time domain" features, extracted from the ECG-based RR interval series, such as "HRV MaxNN" and "HRV MinNN" ranked highly, particularly under MI and RF, demonstrating their close association with fatigue. "HRV Non-linear domain" such as "HRV SD1" and "HRV ApEn" demonstrated moderate importance, indicating their sensitivity to short-term variability and signal entropy under physiological stress. In contrast, multifractal and Detrended Fluctuation Analysis (DFA)-based features were consistently ranked low across all methods, suggesting limited discriminative power when used independently in this mixed-feature setting. In Fig. 7 (B), HR features extracted from the PPG-based HR series, such as "percentageHRreserve" and "HRindex", demonstrated strong importance under PI and RFE, indicating that these relative heart rate measures effectively reflect the cardiovascular load experienced by the subject. Additionally, "hrSense" and "HRnet" displayed moderate relevance, pointing to their utility in representing dynamic heart rate patterns. Across all feature categories, demographic attributes such as age, weight, and "isSmoker" exhibited consistently low importance scores. This pattern underscores a critical insight: online physiological signals and contextual parameters are substantially more predictive of fatigue than static individual characteristics, affirming the value of human-centered and context-aware monitoring systems.

To further evaluate the practical impact of feature importance on the model's decision logic, we additionally trained the XGBoost model, as it is the best performing (see Section 6.3), using only the most influential features identified by permutation importance. We assessed each reduced feature set using a 5-fold cross-validation procedure. When using only the two most relevant HR-based features (percentageHRreserve and HRindex), the model maintained robust performance (mean accuracy: 90.4 % ±0.8), confirming that PPG-derived HR features alone can support reliable online inference when ECG is unavailable. In contrast, using the most important HRV features (HRV MaxNN, HRV MinNN, HRV SD1, HRV RMSSD, HRV ApEn) together with the exercise type yielded substantially higher performance (mean accuracy: 95.6 % ±0.8), highlighting that short-term HRV dynamics and contextual information jointly drive accurate fatigue detection. Removing the exercise type from this feature set resulted in a significant drop in performance (mean

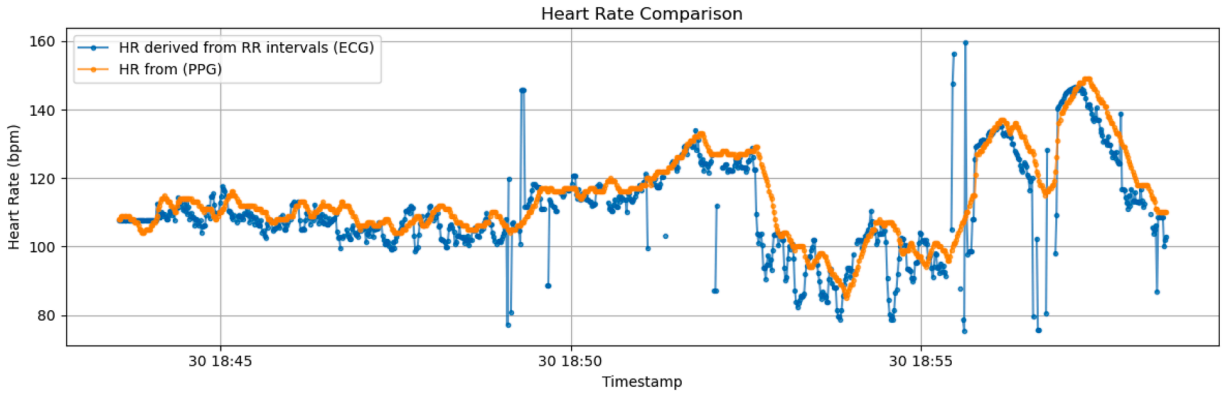

Fig. 8. Heart-rate from ECG (blue) and PPG (orange) sampled every 1 s from athlete (Id: 4) during a ground-based core-stability exercise. The ECG-based HR trace shows several dropouts due to chest-strap contact loss, while the PPG-based HR trace is more stable. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Features and their extraction computational complexities.

Feature Domain	Feature	Extraction Complexity
HRV Time	SDNN, MinNN, RMSSD, MaxNN, SDNNd, SDNNa	$(O(n))$
HRV Non-linear	SD1, SD2, SD1SD2	$(O(n))$
	ApEn	$(O(n^2))$
	DFA_alpha1	$(O(n \log n)$ - $O(n^2))$
	MFDFA_Width, MFDFA_Peak	$(O(n \log n)$ - $O(n^2))$
	MFDFA_Mean, MFDFA_Max	$(O(n \log n)$ - $O(n^2))$
HR Parameters	% HRR, HR Index, HR Net VO_2	$(O(n))$

accuracy: 81.9 % ±1.9), demonstrating that contextual information is important for interpreting physiological changes. These results provide quantitative evidence that the ranked features directly influence prediction behaviour, thereby improving the interpretability of the model. However, during the exercises where athletes performed drills lying prone with their chests against the floor, the chest-band ECG sensor frequently lost contact with the skin, resulting in data interruptions. These disconnections occurred in approximately 20 % of all recorded sessions for that exercise type, leading to missing RR intervals sequences. In contrast, the arm-band PPG sensor maintained stable optical readings and provided continuous heart rate data. Fig. 8 shows an example of how these ECG disconnections cause breaks and sudden changes in the recorded heart rate signal. This observation illustrates a practical trade-off between accuracy and robustness and highlights the value of our context-aware framework, which dynamically adapts sensor choice according to the exercise type. A concrete example of this context dependency emerged during one of the recorded training sessions. The session began with a set of ground-based core-stability exercises, during which athletes frequently leaned or lay on their torsos. In this context, the Polar H10 experienced repeated disconnections as shown in Fig. 8, leading to unreliable or missing RR intervals values. As a result, HRV features could not be used reliably, and the analysis relied on heart-rate features extracted from the Polar Verity Sense, whose signal remained stable throughout these movements. After this phase, the session progressed to upright running activities, where the Polar H10 maintained proper contact and provided consistent RR intervals data. In this part of the session, HRV time-domain features were therefore utilised to achieve higher fatigue detection accuracy. This example shows how different movement conditions within the same training session naturally determine whether HR-based or HRV-based features are more appropriate for fatigue detection. This context-aware analysis further supports coaches in tailoring training sessions by interpreting both physiological responses and sensor reliability. Similar context-aware approaches, such as those by Duque et al. [60] in environmental modelling and Alegre et al. [16] in adaptive system design, demonstrate how integrating contextual information enhances model robustness and interpretability across domains.

Regarding the features complexity, time-domain HRV features, such as the standard deviation of NN intervals (SDNN) and the root mean square of successive differences (RMSSD), involve straightforward statistical calculations, as “MaxNN” and “MinNN”. As shown in Table 7, these operations have a linear computational complexity, denoted as O(n). This linear complexity ensures that time-domain features are computationally efficient, making them suitable for online applications. In our models, these features consistently demonstrated robust performance, indicating that they can capture meaningful patterns in the data. Non-linear HRV features, including Approximate Entropy (ApEn) and DFA, are designed to assess the complexity and dynamic properties of the RR interval series but require more computationally demanding calculations, as shown in Table 7. ApEn measures the regularity

Table 8
Performance metrics at inference for different models and feature combinations.

Model	Metric (%)	Feature Sets				
		HRV Non-linear	HRV Time domain	HRV-Non-linear + HRV-Time domain	Heart Rate	All Features
LSTM	Precision	87.0	93.2	91.2	92.4	93.2
	Recall	87.0	92.9	90.7	89.8	92.9
	Accuracy	87.0	93.2	90.9	91.5	92.9
	F1-score	85.2	**91.0**	89.2	**89.5**	**91.0**
	Loss	55.0	18.8	45.4	20.9	28.9
Transformer	Precision	85.5	89.6	91.8	90.6	92.6
	Recall	83.1	87.8	89.8	85.0	92.1
	Accuracy	84.2	89.0	90.4	89.3	92.1
	F1-score	82.0	86.4	**88.1**	**87.3**	**89.7**
	Loss	37.7	25.5	28.6	26.4	18.9
CNN + LSTM	Precision	86.0	93.2	91.2	91.1	94.3
	Recall	85.3	93.2	91.2	90.1	94.0
	Accuracy	85.6	93.2	91.2	90.9	94.0
	F1-score	82.7	**91.0**	88.5	**88.5**	**92.6**
	Loss	53.7	16.4	31.7	22.8	22.0
CNN + GRU	Precision	88.6	94.3	92.9	92.4	94.3
	Recall	87.6	93.5	92.6	89.3	94.3
	Accuracy	88.1	93.8	92.6	91.5	94.3
	F1-score	86.2	**91.8**	90.2	**89.2**	**93.1**
	Loss	34.3	16.0	23.6	21.1	20.3
SVM RBF	Precision	88.0	89.0	92.0	86.0	93.0
	Recall	86.0	89.0	92.0	84.0	93.0
	Accuracy	87.8	91.2	93.2	86.7	94.0
	F1-score	87.0	89.0	**92.0**	84.0	**93.0**
XGBoost	Precision	88.0	95.0	95.0	89.0	95.0
	Recall	88.0	94.0	95.0	88.0	95.0
	Accuracy	89.5	96.0	96.3	90.7	96.0
	F1-score	88.0	**95.0**	95.0	88.0	**95.0**

and unpredictability of fluctuations in a time series, involving pairwise comparisons of data points, which results in a quadratic complexity of $O(n^2)$. Similarly, DFA evaluates long-range correlations within a time series by analyzing the relationship between fluctuation magnitudes and time scales, with complexity ranging from $O(n \log n)$ to $O(n^2)$ depending on the implementation. Due to their higher computational demands, non-linear features are less suited for online analysis. Moreover, when used independently, these features often exhibited weaker predictive performance in our study (c.f. Table 8), indicating that they may not fully capture the subtle shifts in autonomic regulation associated with fatigue, especially in shorter data segments.

The combination of "HRV Non-linear Domain" and "HRV Time Domain" features did not lead to a significant improvement in performance compared to other feature sets, suggesting that the time domain features have more impact on our fatigue prediction problem. Given the increased computational load associated with non-linear features and their limited added benefit, the combination may not be justified in contexts where computational resources are constrained. HR features are computationally less demanding, with a linear complexity of $O(n)$, as shown in Table 7. Despite their simplicity, these features demonstrated strong predictive capabilities in our models, especially the DL ones (c.f. Table 8), offering an efficient and lightweight option for various applications. The feature set containing all the types of features result in the best performance across most of the models. However, the increased complexity of extracting these additional features, particularly those from the non-linear domain, paired with a negligible gain in accuracy compared to the set of time domain features, makes it a less suitable choice. This finding emphasizes that an optimal feature selection strategy should balance predictive accuracy with computational burden, prioritizing features that offer the most meaningful physiological insights with minimal computational overhead.

6.3. Analysis and comparison of the ML-based fatigue detection models

This section presents several analysis performed to compare the evaluated models. First, we focus on the results obtained on the cross-validation set. Then, we compare the performance obtained on the test set highlighting strengths and weakness of each model and their discriminative power over all the fatigue levels. Finally, we discuss the trade-off between accuracy and computational cost in the selected platform.

Table 9
Cross-validation performance (mean ± standard deviation) for different models and feature combinations.

Model	Metric	Feature Sets				
				HRV-		
	(%)	**HRV Non-linear**	**HRV Time domain**	**Non-linear + Time domain**	**Heart Rate**	**All Features**
LSTM	Accuracy	88.3 ± 1.1	92.2 ± 0.5	92.9 ± 0.7	89.3 ± 0.9	93.8 ± 1.4
	Precision	88.5 ± 0.9	92.3 ± 0.5	93.0 ± 0.7	89.3 ± 0.9	93.9 ± 1.4
	Recall	88.3 ± 1.1	92.2 ± 0.5	92.9 ± 0.7	89.3 ± 0.9	93.8 ± 1.4
	F1-score	88.3 ± 1.1	92.2 ± 0.5	92.8 ± 0.8	89.1 ± 0.9	93.8 ± 1.4
Transformer	Accuracy	87.2 ± 1.2	88.4 ± 1.8	90.6 ± 1.5	87.1 ± 1.8	93.1 ± 1.6
	Precision	87.4 ± 1.3	88.5 ± 1.9	90.6 ± 1.5	87.1 ± 1.6	93.1 ± 1.6
	Recall	87.2 ± 1.2	88.4 ± 1.8	90.6 ± 1.5	87.1 ± 1.8	93.1 ± 1.6
	F1-score	87.1 ± 1.3	88.3 ± 1.9	90.5 ± 1.5	86.9 ± 1.8	93.0 ± 1.7
CNN + LSTM	Accuracy	88.8 ± 2.0	93.4 ± 0.5	93.9 ± 0.9	89.8 ± 1.1	95.0 ± 1.4
	Precision	89.2 ± 1.9	93.5 ± 0.6	94.0 ± 0.9	89.8 ± 1.0	95.1 ± 1.4
	Recall	88.8 ± 2.0	93.4 ± 0.5	93.9 ± 0.9	89.8 ± 1.1	95.0 ± 1.4
	F1-score	88.8 ± 2.0	93.4 ± 0.6	93.9 ± 0.9	89.6 ± 1.1	95.0 ± 1.4
CNN + GRU	Accuracy	89.1 ± 1.3	92.8 ± 0.9	94.4 ± 0.9	89.5 ± 1.3	95.4 ± 1.5
	Precision	89.3 ± 1.2	92.8 ± 0.9	94.4 ± 0.9	89.6 ± 1.4	95.5 ± 1.4
	Recall	89.1 ± 1.3	92.8 ± 0.9	94.4 ± 0.9	89.5 ± 1.3	95.4 ± 1.5
	F1-score	89.1 ± 1.2	92.8 ± 0.9	94.4 ± 1.0	89.1 ± 1.4	95.4 ± 1.5
SVM RBF	Accuracy	88.8 ± 0.7	92.4 ± 2.3	93.5 ± 1.0	88.3 ± 1.5	94.6 ± 1.0
	Precision	89.1 ± 0.8	92.4 ± 2.3	93.5 ± 1.0	88.3 ± 1.6	94.7 ± 1.0
	Recall	88.8 ± 0.7	92.4 ± 2.3	93.5 ± 1.0	88.3 ± 1.5	94.6 ± 1.0
	F1-score	88.7 ± 0.7	92.4 ± 2.3	93.4 ± 1.0	88.0 ± 1.5	94.6 ± 1.0
XGBoost	Accuracy	89.3 ± 0.9	95.4 ± 0.5	95.8 ± 0.4	89.9 ± 1.2	95.7 ± 0.9
	Precision	89.5 ± 0.9	95.4 ± 0.5	95.8 ± 0.4	89.9 ± 1.3	95.7 ± 1.0
	Recall	89.3 ± 0.9	95.4 ± 0.5	95.8 ± 0.4	89.9 ± 1.2	95.7 ± 0.9
	F1-score	89.3 ± 0.9	95.3 ± 0.5	95.8 ± 0.4	89.7 ± 1.2	95.6 ± 0.9

6.3.1. Model evaluation using K-fold cross-validation

To evaluate the generalization performance of the proposed models, a 5-fold cross-validation procedure was conducted across all feature sets. The results are presented in Table 9. Overall, the results demonstrate consistent and robust performance across folds, with relatively low standard deviations indicating stable model behavior. Among the tested models, XGBoost achieved the highest overall performance across most feature combinations, particularly when using the HRV time-domain and mixed HRV feature sets, obtaining mean accuracies of 95.43 ± 0.58 and 95.88 ± 0.42, respectively. Deep learning models such as CNN + GRU and CNN + LSTM also achieved competitive performance, especially when trained on the combined feature sets, with accuracies above 94 %. In contrast, the Transformer model generally exhibited lower performance compared to the other architectures, although it remained consistent across folds. When all features were combined, the models achieved their best results, with XGBoost (95.71 ± 0.95), SVM (94.69 ± 1.07), and CNN + GRU (95.48 ± 1.52) showing particularly strong generalization capabilities.

6.3.2. Accuracy of the fatigue level classification

As described in Section 6.2 and shown in Table 8, each model showed strengths and weaknesses, based on the feature set used.

For the analysis of the models, we focus mostly on the HRV “Time domain” feature set, as from the results it is the best performing or comparable to other sets or a combination of them across most of the models. Among the DL models analyzed, convolutional neural network (CNN) + gated recurrent units (GRU) is the one that performs better, achieving an F1-score of 91.8 %, precision of 94.3 % and recall of 93.5 %. However, long short-term memory (LSTM) and CNN + LSTM show comparable results. Only the transformer has a lower performance with F1-score of 86.4 % considering the same feature set. This could be caused by the need for a larger amount of data, as, in general, DL models require large quantities of data for proper training. Our proposed framework (c.f., Section 3) enables future research to address this problem by collecting new additional data and continuously retraining the models. Traditional ML models such as XGBoost and SVM show comparable results. XGBoost reports the highest performance, achieving an F1-score of 95.0 %, while SVM has a lower F1-score of 89.0 %.

To further evaluate the performance of the compared models in our collected dataset, we perform a ROC curve-based analysis, with a one-vs-rest approach, where each class is compared against the rest. We also compute the AUC score for each model over each class and report it in Table 10. The results show consistently high AUC values across classes 0 through 2, while class 3 shows lower results, although acceptable. Specifically, the XGBoost model demonstrated the highest overall performance, with AUC scores of 1.00, 0.99, 0.96, and 0.92 for classes 0 to 3, respectively.

Table 10
AUC scores per class across models and feature sets with a one-vs-rest approach.

Model	AUC	Feature Sets				
				HRV-		
		HRV Non-linear	**HRV Time domain**	**Non-linear + HRV-Time domain**	**Heart Rate**	**All Features**
LSTM	Class 0	0.99	1.00	1.00	1.00	0.99
	Class 1	0.90	0.98	0.94	0.96	0.97
	Class 2	0.86	0.91	0.91	0.89	0.93
	Class 3	0.87	0.87	0.89	0.87	0.92
Transformer	Class 0	0.99	1.00	0.99	0.99	0.99
	Class 1	0.88	0.96	0.95	0.95	0.97
	Class 2	0.82	0.84	0.90	0.86	0.92
	Class 3	0.83	0.87	0.86	0.87	0.85
CNN + LSTM	Class 0	1.00	1.00	0.99	1.00	1.00
	Class 1	0.87	0.98	0.95	0.96	0.97
	Class 2	0.88	0.93	0.93	0.89	0.90
	Class 3	0.84	0.89	0.85	0.85	0.86
CNN + GRU	Class 0	0.99	1.00	1.00	1.00	1.00
	Class 1	0.91	0.98	0.96	0.96	0.97
	Class 2	0.89	0.92	0.94	0.90	0.93
	Class 3	0.85	0.89	0.83	0.85	0.92
SVM RBF	Class 0	0.99	1.00	1.00	1.00	1.00
	Class 1	0.92	0.97	0.96	0.94	0.96
	Class 2	0.85	0.88	0.92	0.83	0.94
	Class 3	0.89	0.87	0.90	0.83	0.91
XGBoost	Class 0	1.00	1.00	1.00	1.00	1.00
	Class 1	0.93	0.99	0.99	0.96	0.99
	Class 2	0.88	0.96	0.96	0.88	0.95
	Class 3	0.87	0.92	0.94	0.86	0.93

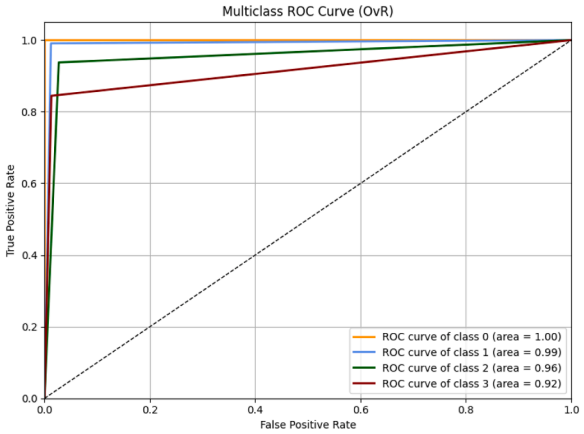

Fig. 9. One-vs-rest ROC curves of the XGBoost classifier trained on HRV Time-Domain features, showing the discrimination of the four fatigue classes (0-3). Each plots the true-positive rate (y-axis) against the false-positive rate (x-axis). Class 3, representing the highest fatigue level, shows the lowest performance, possibly due to the different subject responses at high fatigue levels.

Considering that XGBoost shows the best overall performance, we show its ROC curve in Fig. 9 to further discuss the results. The figure shows that classes 0 and 1 (representing low fatigue levels) exhibit near-perfect true positive rates, while classes 2 and 3 (high fatigue) have a lower performance. Since, in the XGBoost ROC curve, classes 2 and 3 show lower true positive rates compared to classes 0 and 1, we performed the one-vs-one approach to evaluate the discriminative power between classes. Fig. 10 illustrates the ROC curves of the XGBoost model for four class comparisons: Low = class 0 vs Moderate = class 1, Heavy = class 2 vs Severe = class 3, 1 vs Heavy = class 2, and Moderate = class 1 vs Severe = class 3. Specifically, the pairwise analysis reveals that the model struggles more when discriminating between adjacent fatigue classes. For instance, the AUC score for class 0 vs class 1 is 0.72, and for class 2 vs class 3 is 0.79, highlighting limited separability between neighboring fatigue levels. In contrast, higher separability is observed between more distant classes, with class 1 vs class 2 reaching an AUC score of 0.82, and for class 1 vs class 3 reaching the

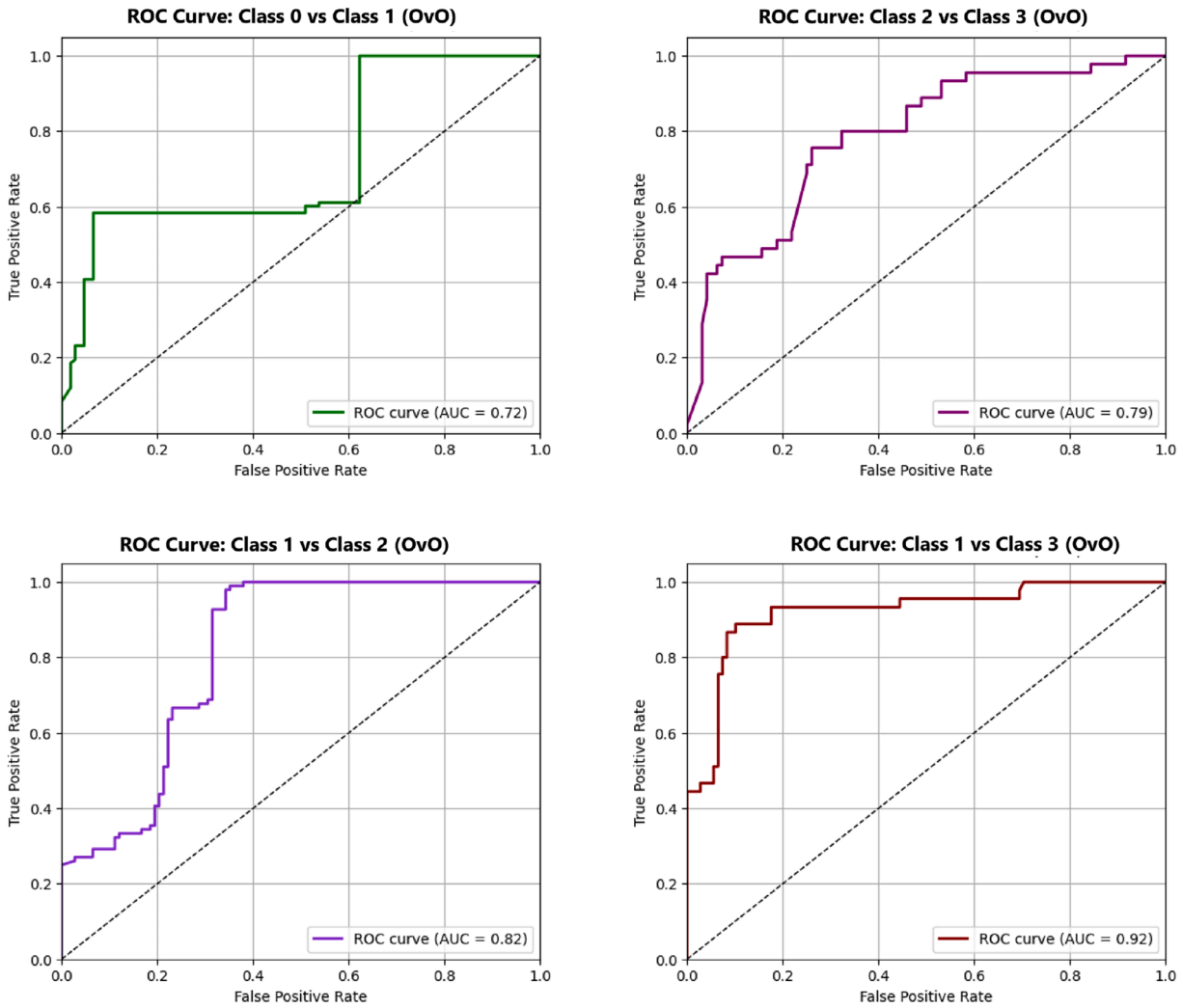

Fig. 10. One-vs-one ROC curves of the XGBoost model trained on HRV Time-Domain features. Each curve compares a pair of fatigue classes (0 vs 1, 0 vs 2, 0 vs 3, 1 vs 3), with true-positive rate (y-axis) plotted against false-positive rate (x-axis). The curves show a limited separability between neighboring fatigue levels, while higher separability is observed between more distant classes, suggesting a more discriminative power between very low and very high fatigue levels.

highest AUC score of 0.92. These results suggest that while the model struggles with fine-grained differentiation between close fatigue levels, it is more effective at separating low and high fatigue levels. This insight could inform future work in two ways: (1) improving model sensitivity for nuanced fatigue distinctions through targeted feature engineering or data augmentation, and (2) considering a binary classification approach for practical use cases, where distinguishing between “low” and “high” fatigue may suffice in the decision support to coaches or sports professionals.

6.3.3. Trade-off between model accuracy and computational cost

As the F1-score metrics of all the models fall in an acceptable and statistically comparable range, we compared them in terms of computational cost, through FLOPs, memory usage, and execution time, as shown in Table 11. LSTM, CNN + LSTM and CNN + GRU have a relatively similar computational cost during inference, while the transformer is one order of magnitude less expensive. Nonetheless, the least computationally intensive model is among the traditional ML ones, the XGBoost, approximately 64x, 139x, and 106x less than CNN + GRU, CNN + LSTM and LSTM, respectively, and approximately 12x less than the transformer. The SVM high computational cost is due to a high number of support vectors trained by the model, suggesting that the SVM model with the RBF kernel is overfitting. The model performs better considering the combination of non-linear and time domain feature sets, though it incurs additional computational costs compared to the time domain set. Despite the literature reporting the RBF SVM as one of the utilized models, in our case, another kernel might be more suitable, or the model needs further optimization of its hyperparameters, which goes beyond the immediate scope of this work. Execution time shows consistent results with the FLOPs, with SVM as the slowest,

Table 11
FLOPs, memory usage and execution time for different models.

Model	FLOPs	Memory Usage (KB)	Execution Time (s)
LSTM	83,183,478	565.2	0.137
Transformer	15,791,757	531.4	0.079
CNN + LSTM	181,444,629	588.0	0.139
CNN + GRU	137,654,826	662.2	0.132
SVM	10,675,716,488	13.4	5.643
XGBoost	1,301,772	15.0	0.051

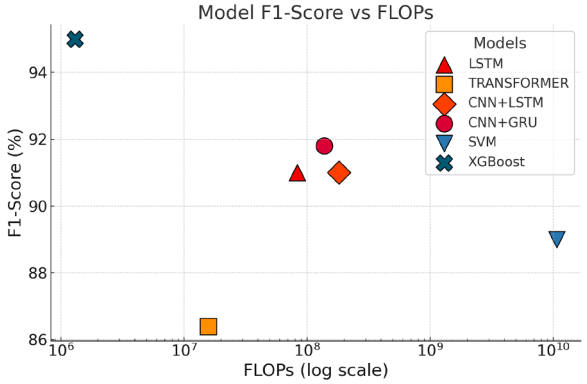

Fig. 11. Computational cost (FLOPs, x-axis) versus classification performance (F1-score, y-axis) for all evaluated models trained on the HRV Time-Domain feature set. Each point represents one model, illustrating the trade-off between F1-score and computational cost. The XGBoost is the model with the optimal trade-off, while CNN + LSTM, CNN + GRU, and LSTM are comparable with an acceptable trade-off. In contrast, transformer and SVM show, respectively, low accuracy vs fairly low complexity, and fairly high in accuracy vs high complexity.

followed by CNN + LSTM LSTM, and CNN + GRU, with similar values. XGBoost is the fastest, approximately 2.7x compared to the DL models, except the transformer with 1.5x. In terms of memory usage, the DL models have similar values, while the traditional ML models are lighter, as expected.

To evaluate the adaptability of the system to different resource-related contexts and user needs, we analyze the trade-off between computational cost and fatigue prediction accuracy. Fig. 11 compares all the analyzed models, in terms of FLOPs, represented in the logarithmic scales, and F1-score, considering the "HRV Time Domain" feature set. As discussed before, SVM is the model with the highest computational cost and with the lowest F1-score. Therefore, it is discarded from this analysis and set aside for future study. The XGBoost shows the best trade-off as it is accurate and computationally less expensive than the other models, suggesting that it can be suitable for deployment in resource-constrained devices. LSTM, CNN + LSTM, and CNN + GRU have a similar trade-off, being comparable in accuracy to the XGBoost, although more computationally expensive. In our study, the limited dataset size does not allow these models to perform effectively. However, our system is designed to continuously collect data and retrain the fatigue prediction model with the goal of building a larger dataset and improving performance. In this case, these DL models can be deployed in a cloud-based module or in an edge device with enough computational resources. Among the DL models, the transformer does not show a good trade-off, despite being less computationally expensive than the others, because of the lower F1-score. Further study is needed to address the suitability of these types of models for fatigue prediction in real-world settings.

As a final evaluation, we breakdown the computational cost of the full fatigue detection process in different steps by simulating online continuous data processing to assess responsiveness and computational efficiency. Table 12 summarizes the execution times for each step of the fatigue detection process: feature extraction, data scaling, and inference using two different models, XGBoost and CNN + GRU. The execution time of each step was computed as average over multiple windows of analysis of the full test set. In the case of our fatigue detection process, the window of analysis is 30s with a sliding window of 10s. Results show that HRV-based feature extraction (ECG-derived RR intervals) requires approximately 475.5 ms, while HR-based feature extraction (PPG-derived heart rate) requires approximately 69.3 ms. The data scaling stage adds less than 2 ms across all feature groups, and the inference stage is extremely lightweight, requiring approximately 0.1 ms for XGBoost and 0.4 ms for CNN + GRU. Consequently, even the most computationally intensive pipeline (HRV features + scaling + XGBoost inference) remains below 480 ms, well within the time restrictions of the window of analysis. Moreover, the latency of our fatigue detection method will be 30s (time to collect the data according to the window of analysis) + up to approximately 0.5s by design, which is acceptable considering the physiological aspect of fatigue detection and short-term HRV analysis [41]. These results confirm that the system operates in near real-time with minimal latency on the selected platform, reinforcing its suitability for real-world deployment and providing a solid foundation for future research.

Table 12
Process execution time summary.

Process	Details	Execution Time (ms)
Feature Extraction	HRV Features	475.5
	HR Features	69.3
Data Scaling	HR Features	1.3
	Time Domain Features	1.6
	Non-Linear Domain Features	1.6
	Mixed Features	1.7
Inference (Time Domain)	XGBoost	0.1
	CNN + GRU	0.4
Full Process Example	HRV Feature Extraction + Time Domain Data Scaling + XGBoost Inference	**477.2**

In conclusion, by selecting models that offer an optimal trade-off between performance and resource consumption, our system can adapt to various contexts and user needs. This strategy enables the development of a technically robust and practically viable solution, supporting our broader objective of enabling effective and sustainable detection of fatigue levels in diverse, real-life contexts.

6.4. Evaluation of the selected wearable devices

One of the objectives of the performed experiments regarded the evaluation of the benefits that each of the two types of wearable devices can provide to the fatigue prediction problem in real-world settings. This evaluation completes the trade-off analysis for the athletes' fatigue prediction during training sessions, adding user comfort to accuracy and computational burden. We conducted this evaluation considering the feedback received from dedicated interviews with the athletes. In this occasion, all athletes declared that the more comfortable device between the two throughout the whole training was the arm band, Polar Verity Sense. For specific exercises, especially the warm-up session, they declared that the chest band was uncomfortable. In addition, the results discussed in Sections 6.2 and 6.3 indicate that the "HRV Time Domain" features set, derived by the Polar H10 chest band, captures the fatigue prediction better than the other set. Although players complain about its wearability, the Polar H10 seems the most effective device due to its capability of acquiring ECG data and the derived parameters, as shown in the literature [61]. Nonetheless, considering the acceptable results obtained in the experiments with only HR features, the arm band could be a suitable replacement in detecting the athletes' fatigue in specific exercise conditions. Additionally, the results have demonstrated that the improvements obtained by including broader feature sets are marginal, suggesting the use of only one device. Therefore, in the absence of a chest band, opting for a lighter model with acceptable accuracy trained with fewer features, such as the ones derived from PPG data, i.e., the arm band, can be a suitable solution for resource-constrained edge devices.

6.5. Framework efficacy in real-world training sessions

The proposed system integrates the analyzed fatigue prediction models within a complete framework that collects and processes data to provide insights to the coaches. The efficacy of the framework has been tested within the performed experiments at different stages: data collection, ML analysis, online predictions, and feedback. The framework effectively helped in analyzing different models with various feature sets, i.e., types of sensors, to demonstrate the trade-off between accuracy, computational efficiency, and user comfort in predicting the athletes' fatigue level. Additionally, by incorporating data from wearable sensors, the framework offered online detection and feedback to the coaches.

Fig. 12 presents representative results from unseen test sessions. We trained the XGBoost model using "HRV Time Domain" features over a set of consecutive training sessions and tested it in the following one, that is, a new training session. We also compared the fatigue prediction with the RR interval series. Overall, the model effectively captures fatigue trends; however, it tends to misclassify fatigue levels during periods of rapid physiological fluctuations, particularly when the RR intervals exhibit abrupt drops or spikes within the analysis window. These transient changes directly affect the computed HRV features, which are highly sensitive to short-term irregularities in the RR series. This is most likely due to the choice of the 30s window of analysis, which might not capture rapid changes in the RR series. As a result, the feature set no longer reflects a stable physiological state, causing the model to mispredict the fatigue level. This effect is visible in Fig. 12, where brief misclassification episodes coincide with fast RR dynamics. It would be interesting to explore additional features from the ECG signal that capture beat-to-beat changes in a window of a few beats [53]. Despite these fluctuations, the overall patterns remain consistent, indicating that the model generalizes well across sessions while leaving room for improvement through personalized calibration and temporal smoothing.

The experimental results underscore two principles for model selection and system design:

- **Context-aware deployment:** A context-aware approach to model deployment and feature selection can effectively guide the choice of both sensor and model for inference. Lightweight models are ideal for low-power edge devices, while moderately complex models can be deployed on edge devices with more robust resources. Meanwhile, high-complexity models can be offloaded to the

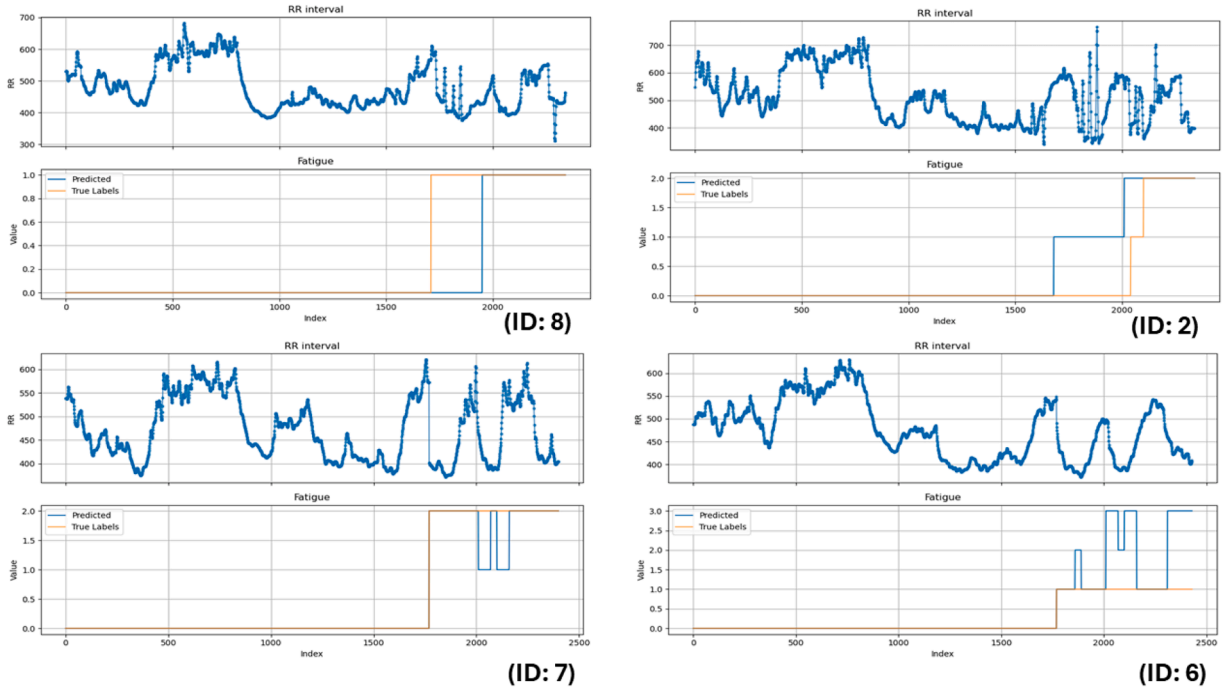

Fig. 12. Predicted and true fatigue levels aligned with RR-interval series for four athletes (Id: 2,6,7,8) during one full training session. Each sequence (2430 s) shows how XGBoost ("using HRV Time Domain features") predictions follow major RR-interval trends, correctly capturing the main fatigue transitions while producing occasional misclassification. This is due to rapid short-term fluctuations in the RR series that might not be captured by the 30s window of analysis.

cloud, which can support more advanced algorithms. In parallel, features must be matched to available resources: for instance, time-domain metrics have linear extraction complexity of ($O(n)$), making them well-suited for resource-limited platforms, while non-linear domain features can scale quadratically $O(n^2)$, demanding greater computational capacity. Such a context-aware design ensures a balance between performance and feasibility across different deployment environments.

- **Human-centered design:** Together with the context, a human-centered design of the system is necessary when tackling a fatigue prediction problem in real-world settings. Specifically, the needs and knowledge of athletes and coaches should be taken into account:
  (a) The athletes' comfort is a key factor in sensor selection, ensuring a balance between technical specifications and user-friendliness.
  (b) The coaches' expertise plays a critical role in ground-truth labeling, while athletes benefit from receiving predictive feedback. This iterative process facilitates future model retraining with coach-refined labels to enhance personalization.

Based on Tables 8 and 10, it can be concluded that the XGBoost model consistently outperforms other models across most metrics and feature combinations, achieving the highest F1-score (up to 95 %) and AUC values (ranging from 0.92 to 1.00 across classes). Among deep learning models, CNN + GRU and CNN + LSTM also show strong and comparable performance, particularly with time-domain HRV and combined feature sets. However, while all models perform well for lower fatigue levels (classes 0-1), their ability to distinguish between higher or adjacent fatigue levels (classes 2-3) is relatively lower, indicating that fine-grained fatigue classification remains challenging. Overall, the results confirm the robustness of the proposed IoT framework and suggest that XGBoost offers the best balance between accuracy and computational efficiency for fatigue detection.

Table 13 summarizes and compares the results of our proposed model with the most relevant state-of-the-art studies on fatigue detection. These works were selected because they employ similar input modalities (ECG or PPG signals), focus on physical or sports-related fatigue, and represent different modeling paradigms – traditional machine learning (e.g., Naïve Bayes, SVM, XGBoost) and deep learning (e.g., Bi-LSTM, ResNet-BiLSTM). As illustrated in Table 13, the reported accuracies and F1-scores across studies vary between 80 % and 92 %, depending on the signal modality, experimental setup, and model complexity. Our XGBoost-based model achieved an accuracy of 96.0 % and an F1-score of 95.0 %. These findings indicate that the proposed method performs competitively with other approaches while maintaining lower computational complexity, making it suitable for online inference in IoT environments. The observed performance differences can be mainly attributed to three factors: (i) the inclusion of contextual information such as exercise type and sensing modality, (ii) the integration of coach-validated labels reducing subjectivity in the training data, and (iii) the combination of efficient feature sets capturing key physiological responses. In contrast, most previous works were conducted

Table 13
Comparison of results for related state-of-the-art fatigue detection studies.

Study	Result	Model
[24]	Accuracy: 91.8 % F1-score: 91.8 % Precision: 91.6 %	ResNet-BiLSTM
[26]	Accuracy: 80.55 % F1-score: 76.88 %	Bi-LSTM
[25]	Accuracy: 82.9 % F1-score: 79.3 % Precision: 80.0 %	SVM
[19]	F1-score: 92.50 %	Naïve Bayes
Our Work	Accuracy: 96.0 % F1-score: 95.0 % Precision: 95.0 %	XGBoost

in controlled laboratory conditions or relied solely on self-reported fatigue scores, limiting their applicability to real-world sports contexts.

The framework shows significant promise for athletic training and performance optimization. Coaches can employ the provided insights to customize training loads and prevent overtraining. Moreover, the system can integrate multiple sensor types to ensure adaptability to various sports. Finally, the use of edge devices, such as smartphones, for initial data storage and preprocessing can enhance portability and ease of use, while also enabling seamless integration with cloud-based systems for advanced analysis and long-term storage.

7. Conclusions

Achieving peak performance while minimizing injury risks in competitive sports remains a fundamental challenge for athletes and coaches. One of the factors that mostly influences athletes' physical and mental capacities and increases the probability of injuries is fatigue. For this reason, this paper presented a smart IoT framework to detect athlete fatigue in real time during training sessions. Additionally, we designed a real-world study to collect data from eight futsal players over five training sessions to validate the IoT framework and provide a public dataset for future research. Throughout the sessions, the system was used to gather and analyze physiological data from the athletes using two types of commercially available devices capturing two key types of data: ECG and PPG related parameters.

The proposed framework has been designed with several key objectives: a) incorporate advanced sensors that can unobtrusively and efficiently capture a broad spectrum of physiological data without interfering with athletes' performance, b) enable local collection, management, and preprocessing of data from both wearable sensors, c) utilize cloud-based ML approach to explore and optimize the fatigue detection problem for different contexts in terms of computation and user needs, d) provide online access to collected data and fatigue predictions to both athletes and coaches during training sessions, e) make the gathered data available to relevant stakeholders through a publicly accessible dataset, supporting further research and analysis. The collected dataset was used to conduct a preliminary exploratory analysis to compare the performance of various ML algorithms to predict fatigue using different combinations of HRV and HR features, derived from ECG and PPG. The study demonstrated the valuable role of integrating IoT and ML in improving athlete performance and ensuring their safety. We also emphasize the importance of leveraging these technologies to optimize the balance between model accuracy, user comfort, and computational efficiency, enabling future personalized and effective training strategies in competitive sports. The results demonstrate that contextual information is essential for reliable and accurate fatigue prediction, especially exercise type and sensing modality. By jointly analyzing ECG-derived HRV and PPG-derived HR features, the proposed framework adapts to each exercise context, achieving a practical balance between accuracy, comfort, and data reliability.

Our study paves the way for various future works. At first, further experiments needs to be performed by expanding the study to a wider group of players and by incorporating additional physiological parameters. In addition, due to the difficulties in obtaining precise and unbiased fatigue scores during training, future works should evaluate quantitative and reliable scores or a more advanced label correction. Other wearable devices should be evaluated and tested within the provided framework to further validate the adaptability and extensibility of the solution. Another promising direction is fatigue forecasting within the same training session and between sessions, providing coaches with additional insights to personalize training programs, prevent overtraining, and optimize athletic performance.

In conclusion, by bridging the gap between cutting-edge IoT systems and practical sports applications, this research contributed to the field of intelligent athlete monitoring by paving the way for more data-driven methodologies and systems. Beyond its application to fatigue detection, the proposed IoT system represents an initial step toward intelligent sports performance management. By offering online feedback to both athletes and coaches, our proposed system can facilitate data-driven decision-making regarding training loads, recovery times, and injury prevention strategies. Furthermore, the published dataset provides a basis for exploring fatigue predictions

using physiological and demographic data collected from a real-world experimental study, offering valuable insights for the research community.

CRediT authorship contribution statement

Abdelkarim Mamen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Data curation, Conceptualization; **Elisabetta De Giovanni:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization; **Teodoro Montanaro:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Data curation, Conceptualization; **Ilaria Sergi:** Writing – review & editing, Visualization, Supervision, Conceptualization; **Luigi Patrono:** Writing – review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Data availability

We have shared the link to the generated data in the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Abdelkarim Mamen currently exploits a research grant partially funded by Inmatica S.p.A. and the Italian Ministerial Decree No. 352/2022 - PNRR, Mission 4, Component 2, "From Research to Enterprise" - Investment 3.3. "Introduction of innovative doctoral programs that respond to the needs of enterprises and promote the recruitment of researchers from enterprises" - Innovative doctoral scholarship with industrial connotation. In addition, the work has been partially funded by "MUR, Ministero dell'Universit‘a e della Ricerca" under the project title "IDAS - Innovazione Digitale in Ambito Salute" through the Italian Ministerial Decree "Decreto Ministeriale n. 737 - 25-06-2021 - Criteri di riparto e utilizzazione del Fondo per la promozione e lo sviluppo delle politiche del Programma Nazionale per la Ricerca (PNR)" - CUP: F84D22000270001. Furthermore, it has been partially funded by the grant RYC2021-032853-I from MCIN/AEI/ 10.13039/501100011033, the European Union NextGenerationEU/PRTR, and the European Union's Horizon Europe research and innovation programme under the Marie Sklodowska-Curie actions (Grant Agreement 101202439 - SWEATHEART).

References

[1] M. Koch, M. Klügl, B. Frankewycz, S. Lang, M. Worlicek, D. Popp, V. Alt, W. Krutsch, Football-related injuries are the major reason for the career end of professional male football players, Knee Surg. Sports Traumatol. Arthroscopy 29 (11) (2021) 3560–3568. https://doi.org/10.1007/s00167-021-06684-8
[2] M. Stojchevska, B. Steenwinckel, J. Van Der Donckt, M. De Brouwer, A. Goris, F. De Turck, S. Van Hoecke, F. Ongenae, Assessing the added value of context during stress detection from wearable data, BMC Med. Inform. Decis. Mak. 22 (1) (2022) 268. https://doi.org/10.1186/s12911-022-02010-5
[3] F. Dambroz, F.M. Clemente, I. Teoldo, The effect of physical fatigue on the performance of soccer players: a systematic review, PloS one 17 (7) (2022) e0270099. https://doi.org/10.1371/journal.pone.0270099
[4] A. Zajac, M. Chalimoniuk, A. Maszczyk, A. Gołaś, J. Lngfort, Central and peripheral fatigue during resistance exercise–a critical review, J. Hum. Kinet. 49 (2015) 159. https://doi.org/10.1515/hukin-2015-0118
[5] S. Balachandar, R. Chinnaiyan, Reliable digital twin for connected footballer, in: International Conference on Computer Networks and Communication Technologies: ICCNCT 2018, Springer, 2019, pp. 185–191. https://doi.org/10.1007/978-981-10-8681-6_18
[6] A.-C. Phan, T.-N. Trieu, T.-C. Phan, Driver drowsiness detection and smart alerting using deep learning and IoT, Internet Things 22 (2023) 100705. https://doi.org/10.1016/j.iot.2023.100705
[7] M.A. Al Imran, F. Nasirzadeh, C. Karmakar, Designing a practical fatigue detection system: a review on recent developments and challenges, J. Saf. Res. 90 (2024) 100–114. https://doi.org/10.1016/j.jsr.2024.05.015
[8] N. Hernandez, L. Castro, J. Medina-Quero, J. Favela, L. Michán, W.B. Mortenson, Scoping review of healthcare literature on mobile, wearable, and textile sensing technology for continuous monitoring, J. Healthcare Inf. Res. (2021) 1–30. https://doi.org/10.1007/s41666-020-00087-z
[9] J.A. Albert, A. Herdick, C.M. Brahms, U. Granacher, B. Arnrich, PERSIST: a multimodal dataset for the prediction of perceived exertion during resistance training, Data 8 (1) (2022) 9. https://doi.org/10.3390/data8010009
[10] R. Eston, Use of ratings of perceived exertion in sports, Int. J. Sports Physiol. Perform. 7 (2) (2012) 175–182. https://doi.org/10.1123/ijspp.7.2.175
[11] C. Foster, A.C. Snyder, N.N. Thompson, K. Kuettel, Normalization of the blood lactate profile in athletes, Int. J. Sports Med. 9 (03) (1988) 198–200. https://doi.org/10.1055/s-2007-1025005
[12] F. Oliveira, D.G. Costa, F. Assis, I. Silva, Internet of intelligent things: a convergence of embedded systems, edge computing and machine learning, Internet Things (2024) 101153. https://doi.org/10.1016/j.iot.2024.101153
[13] K.S. Hahm, B.W. Anthony, Machine learning-based gait health monitoring for multi-occupant smart homes, Internet Things 26 (2024) 101154. https://doi.org/10.1016/j.iot.2024.101154
[14] F.-Y. Leu, P.-J. Chiang, H. Susanto, R.-T. Hung, H.-L. Huang, Mobile physiological sensor cloud system for long-term care, Internet Things 11 (2020) 100209. https://doi.org/10.1016/j.iot.2020.100209
[15] A. Mamen, S. Kovaçi, T. Montanaro, I. Sergi, L. Patrono, A digital twin architecture for minimizing injuries risks with personalized regimens via IoT and machine learning, in: 2024 9th International Conference on Smart and Sustainable Technologies (SpliTech), IEEE, 2024, pp. 1–5. https://doi.org/10.23919/SpliTech61897.2024.10612411
[16] U. Alegre, J.C. Augusto, T. Clark, Engineering context-aware systems and applications: a survey, J. Syst. Softw. 117 (2016) 55–83. https://doi.org/10.1016/j.jss.2016.02.010

[17] C. Ejiofor, I.J. Mgbeafuluike, Model monitoring physical exercise heart rate using internet of things (MMPEH-IOT), Int. J. Artif. Intell. Tools 8 (2018) 21–26. https://api.semanticscholar.org/CorpusID:26073593. https://doi.org/10.5121/ijait.2018.8202

[18] L. Lukač, I. Fister, Jr, I. Fister, Digital twin in sport: from an idea to realization, Appl. Sci. 12 (24) (2022) 12741. https://doi.org/10.3390/app122412741

[19] S. Ding, X. Pan, R. Han, X. Zeng, Y. Li, X. Zheng, A machine learning approach to reduce mental fatigue risk of pilots based on HRV data (2022). https://doi.org/10.1049/icp.2022.2923

[20] M.L. Goodwin, J.E. Harris, A. Hernández, L.B. Gladden, Blood lactate measurements and analysis during exercise: a guide for clinicians, J. Diabetes Sci. Technol. 1 (4) (2007) 558–569. https://doi.org/10.1177/193229680700100414

[21] X. Hao, Q. Zhou, Z. Liu, Experiment study of weight-bearing walking fatigue of human body based on ECG signal characteristics, in: Man–Machine–Environment System Engineering: Proceedings of the 17th International Conference on MMESE 17, Springer, 2018, pp. 269–277. https://doi.org/10.1007/978-981-10-6232-2_32

[22] D.J. Plews, P.B. Laursen, J. Stanley, A.E. Kilding, M. Buchheit, Training adaptation and heart rate variability in elite endurance athletes: opening the door to effective monitoring, Sports Med. 43 (2013) 773–781. https://doi.org/10.1007/s40279-013-0071-8

[23] E.J. Husom, R. Dautov, A. Nedisan Videsjorden, F. Gonidis, S. Papatzelos, N. Malamas, Machine learning for fatigue detection using fitbit fitness trackers, in: Proceedings of the 10th International Conference on Sport Sciences Research and Technology Support (icSPORTS 2022), SciTePress, 2022. https://doi.org/10.5220/0011527500003321

[24] P. Liu, Y. Song, X. Yang, D. Li, M. Khosravi, Medical intelligence using PPG signals and hybrid learning at the edge to detect fatigue in physical activities, Sci. Rep. 14 (1) (2024). https://doi.org/10.1038/s41598-024-66839-8

[25] L. Gan, Z. Yang, Y. Shen, R. Cao, Y. Xia, Y. Shi, B. Cao, Heart rate variability analysis method for exercise-induced fatigue monitoring, Biomed. Signal Process. Control 92 (2024). https://doi.org/10.1016/j.bspc.2024.105966

[26] X. Guan, Y. Lin, Q. Wang, Z. Liu, C. Liu, Sports fatigue detection based on deep learning, in: 2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2021, pp. 1–6. https://doi.org/10.1109/CISP-BMEI53629.2021.9624395

[27] A. Caroppo, A.M. Carluccio, G. Rescio, A. Manni, A. Leone, Fatigue estimation through multimodal data retrieved from a commercial wearable device, in: AIxAS@ AI* IA, 2023, pp. 40–52.

[28] J.J. Merrigan, J.H. Stovall, J.D. Stone, M. Stephenson, V.S. Finomore, J.A. Hagen, Validation of Garmin and Polar devices for continuous heart rate monitoring during common training movements in tactical populations, Meas. Phys. Educ. Exerc. Sci. 27 (3) (2023) 234–247. https://doi.org/10.1080/1091367X.2022.2161820

[29] K. Hinde, G. White, N. Armstrong, Wearable devices suitable for monitoring twenty four hour heart rate variability in military populations, Sensors 21 (4) (2021) 1061. https://doi.org/10.3390/s21041061

[30] Polar H10 heart rate sensor - official website, 2024, (https://www.polar.com/it/sensors/h10-heart-rate-sensor).

[31] Polar verity sense - official website, 2024, (https://www.polar.com/it/products/accessories/polar-verity-sense).

[32] A.M. Amiri, Abhinav, K. Mankodiya, m-QRS: an efficient QRS detection algorithm for mobile health applications, in: 2015 1th International Conference on E-health Networking, Application & Services (HealthCom), 2015, pp. 673–676. https://doi.org/10.1109/HealthCom.2015.7454590

[33] L. Schmitt, J. Regnard, G.P. Millet, Monitoring fatigue status with HRV measures in elite athletes: an avenue beyond RMSSD?, Front. Physiol. 6 (2015) 343. https://doi.org/10.3389/fphys.2015.00343

[34] J.-G. Dong, The role of heart rate variability in sports physiology, Exp. Ther. Med. 11 (2016) 1531–1536. https://doi.org/10.3892/etm.2016.3104

[35] D. Tao, Y. He, A. Cole, R. Awan-Scully, R. Supriya, Y. Gao, J.S. Baker, The benefits of heart rate variability (HRV) in the assessment of health and exercise performance, Imaging J. Clin. Med. Sci. 9 (1) (2022) 011–014. https://doi.org/10.17352/2455-8702.000136

[36] Y.-X. Chen, C.-K. Tseng, J.-T. Kuo, C.-J. Wang, S.-H. Chao, L.-J. Kau, Y.-S. Hwang, C.-L. Lin, Fatigue estimation using peak features from PPG signals, Mathematics 11 (16) (2023) 3580. https://doi.org/10.3390/math11163580

[37] F. Scardulla, G. Cosoli, S. Spinsante, A. Poli, G. Iadarola, R. Pernice, A. Busacca, S. Pasta, L. Scalise, L. D'Acquisto, Photoplethysmograhic sensors, potential and limitations: is it time for regulation? a comprehensive review, Measurement 218 (2023) 113150. https://doi.org/10.1016/j.measurement.2023.113150

[38] M. Kent, Rating of perceived exertion, 2007. https://www.oxfordreference.com/view/10.1093/acref/9780198568506.001.0001/acref-9780198568506-e-5798. https://doi.org/10.1093/acref/9780198568506.013.5798

[39] A. Mamen, E. De Giovanni, T. Montanaro, I. Sergi, E. Russo, L. Patrono, Fatigue monitoring in futsal athletes using physiological wearable sensors, (2025). https://doi.org/10.5281/zenodo.15076182

[40] The general data protection regulation, (2025). Accessed: 2025-03-24, https://www.consilium.europa.eu/en/policies/data-protection-regulation/.

[41] F. Shaffer, J.P. Ginsberg, An overview of heart rate variability metrics and norms, Front. Public Health 5 (2017) 258. https://doi.org/10.3389/fpubh.2017.00258

[42] D. Makowski, T. Pham, Z.J. Lau, J.C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S.H.A. Chen, NeuroKit2: a python toolbox for neurophysiological signal processing, Behav. Res. Methods (2021) 1–8. https://doi.org/10.3758/s13428-020-01516-y

[43] K. Mj, The effects of training on heart rate: a longitudinal study, Ann. Med. Exp. Biol. Fenn. 35 (1957) 307–315.

[44] J.R. Wicks, N.B. Oldridge, L.K. Nielsen, C.E. Vickers, HR index–a simple method for the prediction of oxygen uptake, Med. Sci. Sports Exerc. 43 (10) (2011) 2005–2012. https://doi.org/10.1249/MSS.0b013e318217276e

[45] A.L. Colosio, M. Lievens, S. Pogliaghi, J.G. Bourgois, J. Boone, Heart rate-index estimates aerobic metabolism in professional soccer players, J. Sci. Med. sport 23 (12) (2020) 1208–1214. https://doi.org/10.1016/j.jsams.2020.04.015

[46] S. Morishita, A. Tsubaki, T. Takabayashi, J.B. Fu, Relationship between the rating of perceived exertion scale and the load intensity of resistance training, Strength Cond. J. 40 (2) (2018) 94–109. https://doi.org/10.1519/SSC.0000000000000373

[47] D. Bustos, F. Cardoso, M. Rios, M. Vaz, J. Guedes, J. Torres Costa, J. Santos Baptista, R.J. Fernandes, Machine learning approach to model physical fatigue during incremental exercise among firefighters, Sensors 23 (1) (2022) 194. https://doi.org/10.3390/s23010194

[48] S. Kandel, A. Paepcke, J. Hellerstein, J. Heer, Wrangler: interactive visual specification of data transformation scripts, in: Proceedings of the Sigchi Conference on Human Factors in Computing Systems, 2011, pp. 3363–3372. https://doi.org/10.1145/1978942.1979444

[49] B. Radišić, S. Seljan, I. Dundjer, Impact of missing values on the performance of machine learning algorithms, in: 5th International Conference on Recent Trends and Applications in Computer Science and Information Technology (RTA-CSIT), University of Tirana, Faculty of Natural Sciences, Department of Informatics, 2023, pp. 54–62.

[50] D. Dallah, H. Sulieman, A.A. Zaatreh, F. Kamalov, Empirical evaluation of the relative range for detecting outliers, Entropy 27 (7) (2025) 731. https://doi.org/10.3390/e27070731

[51] R. Verma, R. Mehrotra, V. Bhateja, An integration of improved median and morphological filtering techniques for electrocardiogram signal processing, in: 2013 3rd IEEE International Advance Computing Conference (IACC), 2013, pp. 1223–1228. https://doi.org/10.1109/IAdCC.2013.6514402

[52] scipy.signal.butter, (https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.butter.html#scipy.signal.butter).

[53] E. De Giovanni, A.A. Valdes, M. Peon-Quiros, A. Aminifar, D. Atienza, Real-time personalized atrial fibrillation prediction on multi-core wearable sensors, IEEE Trans. Emerg. Top. Comput. 9 (4) (2020) 1654–1666. https://doi.org/10.1109/TETC.2020.3014847

[54] E. De Giovanni, T. Teijeiro, G.P. Millet, D. Atienza, Adaptive r-peak detection on wearable ecg sensors for high-intensity exercise, IEEE Trans. Biomed. Eng. 70 (3) (2022) 941–953. https://doi.org/10.1109/TBME.2022.3205304

[55] What Is Undersampling?, (2022), (https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/?utm_source=chatgpt.com).

[56] A.M. Albarrak, R. Alharbi, I.A. Ibrahim, Detection and classification of unhealthy heartbeats using deep learning techniques, Sensors 25 (19) (2025) 5976. https://doi.org/10.3390/s25195976

[57] R. Muralidhar, M.L. Demory, M.M. Kesselman, M.D. Beckler, Exploring the impact of batch size on deep learning artificial intelligence models for malaria detection, Cureus 16 (5) (2024). https://doi.org/10.7759/cureus.60224

[58] S. Afaq, S. Rao, Significance of epochs on training a neural network, Int. J. Sci. Technol. Res. 9 (06) (2020) 485–488.

[59] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, Int. J. Data Min. Knowl. Manage. Process 5 (2) (2015) 1. https://doi.org/10.5121/ijdkp.2015.5201
[60] H. Duque, K. Diao, R. Villa, J.P. Leitao, S. Djordjević, M. Abdel-Aal, Context-aware data driven sensor data analysis: with application to H2S concentration prediction in urban drainage networks, Water Res. X (2025) 100346. https://doi.org/10.1016/j.wroa.2025.100346
[61] J.M. Mühlen, J. Stang, E.L. Skovgaard, P.B. Judice, P. Molina-Garcia, W. Johnston, L.B. Sardinha, F.B. Ortega, B. Caulfield, W. Bloch, S. Cheng, U. Ekelund, J.C. Brønd, A. Grøntved, M. Schumann, Recommendations for determining the validity of consumer wearable heart rate devices: expert statement and checklist of the INTERLIVE network, Br. J. Sports Med. 0 (2021) 1–13. https://doi.org/10.1136/bjsports-2020-103148