Li *et al. Journal of Cloud Computing* (2024) 13:73
https://doi.org/10.1186/s13677-024-00636-6

Journal of Cloud Computing:
Advances, Systems and Applications

RESEARCH

Open Access

Time-aware outlier detection in health physique monitoring in edge-aided sport education decision-makings

Yanjie Li[1], Liqin Kang[2], Zhaojin Li[3], Fugao Jiang[3], Nan Bi[4], Tao Du[5] and Maryam Abiri[6*]

Abstract

The increasing popularity of various intelligent sensor and mobile communication technologies has enabled quick health physique sensing, monitoring, collection and analyses of students, which significantly promoted the development of sport education. Through collecting the students' physiological signals and transmitted them to edge servers, we can precisely analyze and judge whether a student is in poor health (e.g., an outlier). However, with time elapsing, the accumulated physiological signals of students become massive, which places a heavy burden on the quick storage and in-time processing of physiological data of students. In this situation, it is becoming a necessity to develop a time-aware outlier detection technique for health physique evaluation of students in a time-efficient way. Considering this challenge, we propose a novel time-aware outlier detection method named TOD based on Locality-Sensitive Hashing. TOD condenses extensive physiological student data into a concise set of health indices. Leveraging these indices, we can efficiently identify potential student outliers from a large pool of candidates with precision and speed. Finally, we have designed a group of simulated experiments based on WS-DREAM dataset. Experiment results prove the feasibility and superiority of the TOD method compared with other existing methods.

Keywords Health physique monitoring, Edge computing, Outlier detection, Time-efficiency, Locality-Sensitive Hashing

Introduction

The dawn of the 21st century has seen a paradigm shift in the domain of sport education, largely fueled by the advent and integration of intelligent sensor technologies and advanced mobile communication systems [1–3]. These technological advancements have opened new vistas in the realm of health physique monitoring, particularly in the educational context, where the physical well-being of students is of paramount importance. In recent years, the proliferation of smart, sensor-based technologies has revolutionized the way we perceive and interact with our immediate environment [4, 5]. In the field of sport education, these technologies have been instrumental in facilitating rapid and accurate sensing, monitoring, and analysis of students' health parameters. This enhanced capability has not only contributed to a more nuanced understanding of student health dynamics but has also paved the way for more personalized and effective physical education strategies.

However, this influx of data presents a significant challenge. As time progresses, the accumulated physiological data from students burgeons to an enormous

*Correspondence:
Maryam Abiri
m_abiri@sbu.ac.ir
[1] Physical Education Department, Shandong University (Weihai), Weihai, China
[2] Marine College, Shandong University (Weihai), Weihai, China
[3] Department of Exercise Science and Sports Studies, Qufu Normal University, Rizhao, China
[4] School of Competitive Sports, Shandong Sport University, Jinan, China
[5] School of Continuing Education and Training, Shandong Sport University, Jinan, China
[6] Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran

volume. This exponential growth in data volume places a substantial strain on storage capacities and the timely processing of this information [6–8]. In such a scenario, the need for a sophisticated, time-aware outlier detection mechanism becomes evident. This requirement is not just for the sake of data management efficiency but also for ensuring that the health monitoring process remains robust and effective, particularly in identifying students whose health parameters deviate from the norm, termed as 'outliers'.

Addressing this need, our research introduces a novel technique named Time-aware Outlier Detection (TOD). TOD represents a significant leap in the realm of health physique monitoring in sport education. By employing locality-sensitive hashing, TOD transforms the voluminous physiological data of students into a concise set of health indices. These indices, despite their compact nature, encapsulatively represent the health status of each student. This transformation is crucial in swiftly pinpointing potential outliers among a large pool of students, thereby ensuring that the detection process is not just accurate but also time-efficient.

The practical efficacy and superiority of TOD over existing methodologies are substantiated through a series of simulated experiments. These experiments, structured around the WS-DREAM dataset, offer a comprehensive evaluation of TOD's performance. The results from these experiments unequivocally demonstrate the feasibility of TOD and its superiority in comparison to other existing outlier detection methods. The significance of TOD lies in its ability to seamlessly integrate with edge computing environments. Edge computing, characterized by its proximity to data sources, enables quicker data processing and decision-making at the edge of the network, closer to where data is generated [9, 10]. This is particularly beneficial in the context of health physique monitoring in sport education, where real-time data analysis and immediate response are crucial. Through the utilization of edge computing, TOD maximizes the processing speed and efficiency of the abundant physiological data produced by students, resulting in prompt and precise health evaluations.

Furthermore, TOD's approach to outlier detection is not merely a technical advancement; it has profound implications for the holistic development of students in the realm of sport education. By enabling timely and accurate identification of students who may be at health risks, educators and health professionals can intervene proactively, ensuring that each student receives the attention and care they require. This not only fosters a safer and more supportive learning environment but also contributes to the overall well-being and physical development of students.

In summary, the contributions of the paper are briefly introduced as follows.

(1) We propose a novel time-aware efficient outlier detection method TOD based on Locality-Sensitive Hashing. TOD can generate a small number of student health indices based on the massive health data of students distributed in multiple edge servers. And with the few student health indices, we can precisely detect the possible student outliers among massive candidates in a time-efficient manner.
(2) In order to demonstrate the effectiveness of the TOD approach in outlier detection, we conducted a series of simulated experiments using the WS-DREAM dataset, which reflects real-world scenarios. By analyzing and contrasting the experimental outcomes with those of alternative methods, we substantiate the efficacy and superiority of TOD in identifying potential outliers in student health assessments.

The reminder of this article is organized as follows. Related literatures are investigated and introduced in "Related work" section. A motivating example is presented in "Motivation" section to emphasize the research significance of this article. The proposed TOD method is clarified in detail in "Our proposal: TOD" section. Evaluation is provided in "Experiments" section. At last, we conclude the paper and discuss the future directions in "Conclusion and future work" section.

Related work

Edge computing in health monitoring

Work [11] presents a novel approach to task scheduling in edge computing environments, specifically tailored for health monitoring systems in smart cities. The authors utilize Software-Defined Networking (SDN) to create a more secure and efficient task scheduling mechanism. The proposed solution aims to optimize the allocation of computing resources, reduce latency, and ensure secure data transmission, which are critical for effective health monitoring operations in urban environments. In [12], the authors address the need for advanced and secure health monitoring systems by proposing an Optimized Deep Recurrent Neural Network (O-DRNN) model integrated with edge computing. The model is designed to enhance the real-time collection and analysis of health data, ensuring faster and more personalized healthcare services. In [13], an innovative health monitoring system is introduced, which integrates IoT technologies with edge computing. At its heart lies a Convolutional Neural Network (CNN) model specifically tuned for cardiac classification. It operates within an IoT framework, where data from various health sensors are processed locally at edge servers, significantly reducing latency and energy consumption. In [14], the authors focus on developing a comprehensive patient health monitoring system utilizing the synergies of fog and edge computing. The system integrates a variety of sensors, including temperature, pulse, tilt, and flex sensors, all connected to a Raspberry Pi Pico W. This setup is particularly designed to assist in monitoring bedridden patients, enabling the tracking of vital signs and movements.

Intelligent sensor technologies in sport education

The paper [15] explores the application of big data analysis in personalized education management systems, using intelligent sensor networks. It provides a theoretical and empirical analysis of individualized teaching methods, emphasizing the role of big data in understanding students' habits and optimizing educational programs. The authors use data mining techniques, like the FP-Growth method, to analyze and optimize training schemes, demonstrating the dynamic and predictive capabilities of big data in education. In [16], the authors introduce the Low-cost Efficient Wireless Intelligent Sensor (LEWIS), a platform aimed at enhancing engineering, research, and education. It discusses the hardware and software architecture of LEWIS1, including a simplified version (LEWIS1 beta) that enhances user interfaces and simplifies both hardware and software. Literature [17] discusses the integration of artificial intelligence and sensor technologies in educational environments, particularly for students with developmental or intellectual disabilities. It highlights the revolution in computational power, brain mapping, wearable sensors, and AI, which enable real-time assessment of students' cognitive states and learning progress. The paper emphasizes the potential of these technologies to adapt teaching methods to individual needs, offering a multi-modal and multi-dimensional approach to education. In [18], the authors discuss the use of LEWIS in monitoring natural hazards like flooding and wildfires, emphasizing its cost-effectiveness and ease of deployment. It also covers the training and education components of LEWIS, showcasing its potential to engage communities, from students to industry professionals, in creating innovative monitoring solutions and combating climate change.

Time-aware outlier detection

In [19], the authors address the intersection of fairness and outlier detection (OD), a relatively unexplored area in fair machine learning. The paper develops FairOD, a fairness-aware outlier detector that aims to achieve treatment parity, flag equal proportions of samples from all groups (group fairness), and identify truly high-risk samples within each group. The study is significant in balancing fairness with detection performance, showing that FairOD can achieve fair outcomes while maintaining or even improving detection accuracy compared to fairness-agnostic detectors. Literature [20] introduces TADILOF, a novel algorithm for outlier detection in data streams, particularly addressing the challenges posed by the Internet of Things (IoT). TADILOF is a time-aware, density-based incremental local outlier detection method that adapts to changes in data over time, such as the emergence of new data clusters. It also features an "approximate LOF" for estimating the local outlier factor based on historical data. In [21], the authors propose a new framework for contextual outlier detection in big data streams. The framework integrates contextual attributes with the content of data streams for outlier detection, addressing both content and contextual anomalies. It includes both supervised and unsupervised detection methods, allowing the system to adapt to normal changes in stream behavior over time. The work [22] introduces AUTO, a novel approach for out-of-distribution (OOD) detection during test time using unlabeled online data. AUTO addresses challenges like catastrophic forgetting by incorporating an in-out-aware filter, an ID memory bank, and a semantically-consistent objective. It adaptively mines pseudo-ID and pseudo-OOD samples from test data to optimize networks in real-time during inference. The authors in [23] propose a method for temporal outlier detection and analysis in business process executions. The approach encompasses modeling temporal behaviors, taking into account various control-flow patterns within business processes, and generating an execution time matrix based on event logs. This matrix serves as the foundation for identifying temporal outliers and conducting correlation-based analyses. It holds particular significance for process-aware information systems, offering valuable insights into the temporal intricacies of business processes.

Outlier detection techniques in health and sport education

In [24], the authors focus on outlier detection in educational data, specifically in inquiry-based learning results of students. It uses the K-Means Clustering method with Minkowski-Chebyshev distances to detect outliers. The comparison results with related methods show the advantage of their proposal. The authors of [25] present a novel method for geocoding historical addresses, which is crucial for spatial data analysis in social science research, including studies on education, health, and emigration. The method uses an online geocoding service and employs outlier detection to improve the accuracy of locations assigned to addresses. Literature [26] uses Electromyography (EMG) signals to evaluate muscle activations in different genders during prolonged sitting tasks and investigates the influence of various types of exercise on muscle activation. The study aims to propose the best exercise to prevent low back discomfort, using EMG signals to record muscle activity and comparing Root Mean Square (RMS) values for muscle activation during prolonged sitting and after exercises. The study in [27] investigates the association between sedentary behavior, physical activity, and depression among sports university students. A cross-sectional survey was conducted in [27] to gather data on students' sedentary behavior, physical activity, and depression levels. The study found a significant association between recreational screen time, sedentary time spent on schoolwork, and participation in vigorous physical activity with depression. The findings highlight the importance of balancing screen time and physical activity to mitigate depression risks in sports university students.

Based on the literature analyses conducted above, it is evident that current literature lacks effective methods for timely outlier detection in monitoring physical health within sports education. Recognizing this gap, we introduce a pioneering approach called TOD, which utilizes Locality-Sensitive Hashing to achieve time-efficient outlier detection.

Motivation

Figure 1 presents an intuitive example to demonstrate the research motivation of this article. In the example, the health physique conditions of students are monitored and collected by smart devices such as smart watches, which form time-aware health signals. Later, health signal data are transmitted to nearby edge servers and then to a central cloud platform for uniform data fusion, analyses and processing. In general, the above data processing framework can accurately recognize the students' health conditions (e.g., outlier detection) through data mining; however, the accumulated health data of students are becoming more and more with time elapsing, which make it hard to detect potential outliers of students in a time-efficient manner. In this situation, we need to develop more lightweight outlier detection method with low time complexity to cope with the big data situation. Inspired by this observation, we propose a novel time-aware outlier detection method with low time complexity for better health physique evaluation of students in edge computing.

Our proposal: TOD

In "Locality-Sensitive Hashing" section, we'll provide a succinct overview of Locality-Sensitive Hashing, the primary technique employed in our paper. Subsequently, guided by Locality-Sensitive Hashing, we present TOD, a novel outlier detection approach adept at managing time-sensitive health monitoring and evaluation in edge environments with effectiveness and efficiency.

Locality-Sensitive Hashing

Locality-Sensitive Hashing is a technique designed for reducing the dimensions of large, high-dimensional datasets while maintaining the similarity relationships between data points. The fundamental concept of Locality-Sensitive Hashing involves hashing input items so that similar items are likely to be grouped into the same "buckets" due to the locality-sensitive nature of the hash functions. This approach facilitates approximate nearest neighbor searches in sub-linear time, offering a much faster alternative to linear search in extensive datasets.

The essence of Locality-Sensitive Hashing lies in employing a set of hash functions, collectively known as a family H, which ensures that similar items x and y have a high probability of ending up in the same hash bucket. This family H is termed "locality-sensitive" and adheres to two key principles for any pair of items x and y:

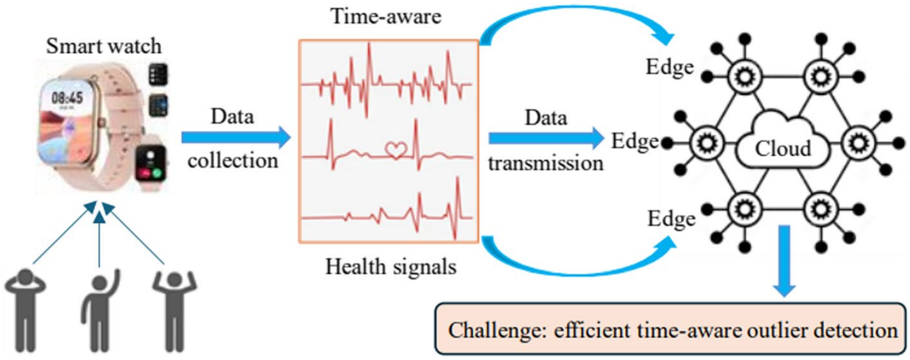


Fig. 1 Time-aware health physique monitoring and evaluation in edge computing: an example and challenge

(1) High Similarity Collision Probability: If x and y are close or similar, the probability$Pr[H(x) = H(y)]$ that they hash to the same value is high.
(2) Low Dissimilarity Collision Probability: Conversely, if x and y are distant or dissimilar, the probability $Pr[H(x) = H(y)]$that they hash to the same value is low.

To further define this concept, consider:

- $d(x, y)$ as the measure of distance between two items x and y.
- r as a specified threshold distance.
- p_1 as the probability that two points hash to the same value when their distance is less than r.
- p_2 as the probability that two points hash to the same value when their distance is greater than r.

A family H is called $(r, p1, p2)$-sensitive if for any x,y:

- If $d(x,y) \leq r$, then $Pr[H(x) = H(y)] \geq p1$
- If $d(x,y) > r$, then $Pr[H(x) = H(y)] \leq p2$

Through adjusting the values of parameters r, p_1 and p_2, we can approximately formulate the similarity-guarantee feature of Locality-Sensitive Hashing. More concretely, a data point can be converted to a hash index through Locality-Sensitive Hashing. With such hash indexes of data points, we can perform privacy-preserving approximate neighbor search: two neighboring data points are of the same or close index with high probability; two non-neighboring data points are of the different or distinguished index with high probability.

In essence, Locality-Sensitive Hashing provides an efficient way to group and search data based on similarity, which is particularly useful in handling large-scale, high-dimensional data in various applications including outlier detection. Locality-Sensitive Hashing can help to create an offline index table which can support time-efficient outlier detection. More concretely, the time complexity of our proposed TOD method based on Locality-Sensitive Hashing is approaching O(1), which is more suitable for the outlier detection applications with big data context.

TOD: time-aware outlier detection based on Locality-Sensitive Hashing

The major idea of TOD method is: according to Locality-Sensitive Hashing, we first classify the students into different clusters based on their monitored health conditions; afterwards, we judge whether a student is a health outlier by considering the number of students who belong to an identical cluster.

Next, we introduce the details of TOD method. TOD starts from analyzing the student-time health data which are often organized and stored within a matrix after the data are collected by various wearable sensors. Here, we use STM to denote the above matrix and the matrix can be represented in Eq. (1). In STM, each row is corresponding to a student's health physique data collected at n different time periods $t_1, ..., t_n$; each column is corresponding to the health physique data of m students $(s_1, ..., s_m)$ collected at a certain time period; the concrete value of the element $hp_{x,y}$ means the health physique data of s_x at time period t_y.

$$STM: \begin{array}{c|ccc} & t_1 & ... & t_n \\ \hline s_1 & hp_{1,1} & ... & hp_{1,n} \\ ... & ... & ... & ... \\ s_m & hp_{m,1} & ... & hp_{m,n} \end{array} \tag{1}$$

In STM, each student probably has a considerable amount of health physique monitoring records with time elapsing. In other words, if the health monitoring period is very long, the matrix STM will become extremely large. In this situation, it will be time-consuming to process and analyze the data in STM and as a consequence, detecting the outliers in STM will consume too much time and is not suitable for the big data context. To tackle this issue, we need to convert the large STM matrix into a much smaller one during which key valuable information contained in STM is still maintained and not lost. In our article, we use Locality-Sensitive Hashing introduced in "Locality-Sensitive Hashing" section to achieve above goal. Locality-Sensitive Hashing has been proven an effective tool to achieve high-quality data compression in big data context and is very useful and applicable to develop lightweight information retrieval and object clustering.

The basic concept of implementing data compression through Locality-Sensitive Hashing involves transforming the original student-time health matrix (STM) into a comparable student health index matrix. Subsequently, we will outline the methodology for achieving this conversion. We take student s_1 as an example to demonstrate the concrete conversion process. In STM matrix, s_1 is corresponding to an n-dimensional vector denoted by $v(s_1)$. Correspondingly, we need to create a new vector (denoted by v_{new}) which also contains n dimensions as $v(s_1)$ does. In concrete, we follow the rule in Eqs. (2) and (3) to create v_{new}. Here, function $f(x, y)$ is responsible to generate a random number which is larger than -1 and smaller than 1.

$$v_{new} = (\varphi_1, ..., \varphi_n) \tag{2}$$

$$\varphi_j = f(-1, 1) \tag{3}$$

With the created new vector v_{new}, we can map the n-dimensional vector $v(s_1)$ to be q_1 (0 or 1) by the conversion formulas in Eqs. (4) and (5). The physical meaning of the above mapping process is as follows: let us consider $v(s_1)$ to be a data point in an n-dimensional space, and the new vector v_{new} to be a plane which splits the n-dimensional space into two parts; then Eqs. (4) and (5) are responsible to judge whether the data point corresponding to $v(s_1)$ is above the plane represented by v_{new} or not: if the data point is above the plane, then $q_1 = 1$; otherwise, $q_1 = 0$.

$$v(s_1)^{\#} = v_{new} * v(s_1) = \sum (\varphi_i * hp_{1,j})(j = 1, 2, ..., n) \tag{4}$$

$$q_1 = \begin{cases} 1 \ \text{if} \ v(s_1)^{\#} > 0 \\ 0 \ \text{else} \end{cases} \tag{5}$$

Since the new vector v_{new} in Eqs. (2) and (3) is created based on a random function $f(x, y)$, a high randomness is inevitable during the judgment process of point-plane relative position formulated in Eqs. (4) and (5). To minimize such a randomness, we do not execute the above process (i.e., Eqs. (2) and (5)) only once for vector $v(s_1)$; instead, we repeat the above process $k(k \ll n)$ times and then obtain k values of q_1, denoted by $q_{1,1}$, . . ., $q_{1,k}$ for vector $v(s_1)$. We use Q_1 in Eq. (6) to represent the abovementioned k-dimensional 0-1 vector. Thus, the n-dimensional vector $v(s_1)$ is changed to be a k-dimensional vector Q_1. Since $k \ll n$, $v(s_1)$ is a real-value vector while Q_1 is a Boolean-value vector, Q_1 is regarded as a promising index of original $v(s_1)$ corresponding to the student s1. Likewise, for each student s_1, . . ., s_m, we can also obtain his or her index denoted by Q_1, . . ., Q_m, respectively.

$$Q_1 = (q_{1,1}, ..., q_{1,k}) \tag{6}$$

With the index of each student, i.e., Q_1, . . ., Q_m, we can classify all the students s_1, . . ., s_m into different clusters (here, we assume there are totally $d(d \leq m)$ clusters (i.e.,$Cluster_1$, . . .,$Cluster_d$) are obtained). The classification rationale or basis is the characteristics of Locality-Sensitive Hashing introduced in "Locality-Sensitive Hashing" section, i.e., the nature formulated by the following conclusion:

A family H is called (r, p_1, p_2)-sensitive if for any x, y:

- If $d(x, y) \leq r$, then $Pr[H(x) = H(y)] \geq p1$
- If $d(x, y) > r$, then $Pr[H(x) = H(y)] \leq p2$

More concretely, if two students own the same index, then their distance is small and could be classified into an identical community. This way, we can classify all the students into corresponding communities. The concrete classification process is demonstrated by the example in Fig. 2. In the example, four students belong to $Cluster_1$, five students belong to $Cluster_2$, . . ., one student belongs to $Cluster_d$. Therefore, in this situation, the last student s_m could be regarded as an outlier since his or her corresponding community contains the fewest students. This way, we use the lightweight student health indices, instead of the heavy-weight student-time health matrix STM in Eq. (1), to achieve the final goal of outlier detection in student health evaluation applications.

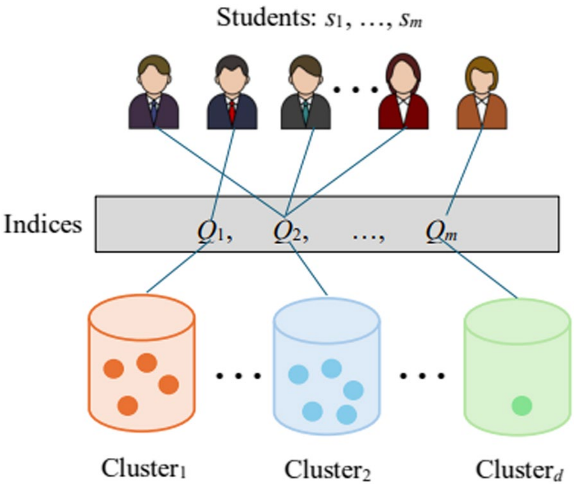


Fig. 2 Student-Index-Cluster correspondence relationship: an example

However, the above student clustering process also faces the challenge of misjudgment due to the randomness nature of Locality-Sensitive Hashing. Inspired by this limitation, we repeat the student health index creation process K times instead of only once and then we have obtained an improved version of Fig. 2, i.e., Fig. 3 which provides the Student-Index-Cluster correspondence relationships within K tables. The difference between Figs. 2 and 3 is not substantial as Fig. 3 is only updated from Fig. 2 by extending from one table to K tables. However, we argue that such an update is necessary and important for guaranteeing the accuracy of outlier detection in TOD method. This is because the Locality-Sensitive Hashing technique adopted in TOD method is essentially a probability-based approximate neighbor search technique and can lead to false-positive or false-negative outlier detection results inevitably. In this situation, we have to minimize the false-positive or false-negative probability through various optimization operations. The above update from Figs. 2 to 3 is for such an optimization since more hash tables can guarantee to reduce the false-positive or false-negative probability of Locality-Sensitive Hashing, which has already been proven by existing literature.

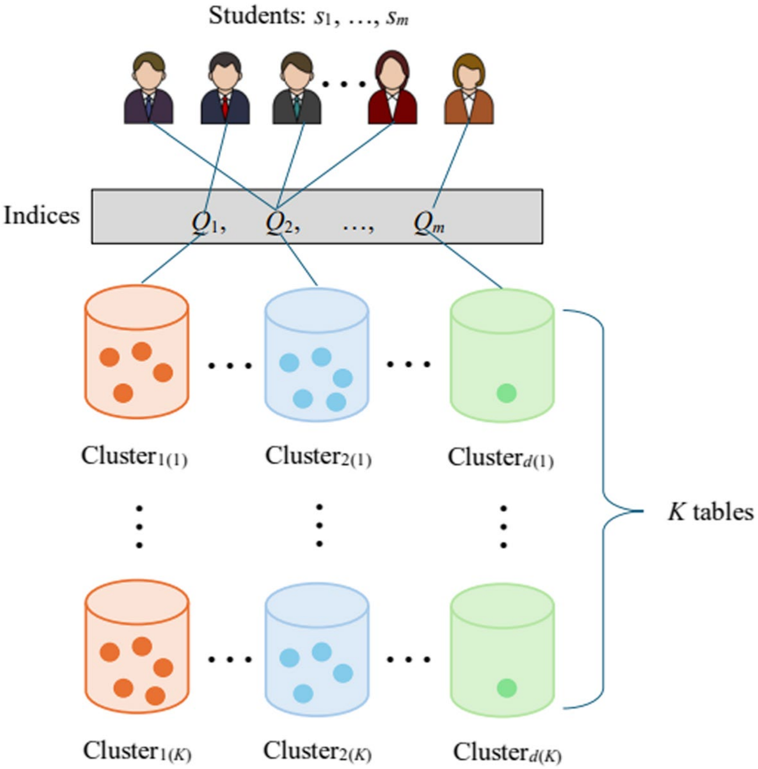


Fig. 3 Student-Index-Cluster correspondence relationships in K tables

With these K tables, we can minimize the randomness of our proposed TOD method in finding potential outliers existed in the health physique monitoring data of students. In concrete, in z-th table T_z $(1 \leq z \leq K)$, we store d_z communities that we classify all the m students into. Please note that in different repetitions, each student is probably classified into different clusters. Then for each cluster, we count the total students who belong to the cluster in any of the K tables. Concrete process is formalized in Eqs. (7) and (8). Here, D in Eq. (7) denotes the number of unrepeated clusters after K repetitions; Size-of($Cluster_\theta$) λ is used to count the total students who belong to θ-th cluster in λ-th table.

$$D = Sizeof\{\{Cluster_{1(1)}, ..., Cluster_{d(1)}\} \cup ... \cup \{Cluster_{1(K)}, ..., Cluster_{d(K)}\}\} \tag{7}$$

$$Sum_\theta = \sum Sizeof\,(Cluster_\theta)_\lambda (1 \leq \lambda \leq K) forany\theta (1 \leq \theta \leq D) \tag{8}$$

$$Sum_{\theta*} = Minimal\{Sum_1, ..., Sum_D\} \tag{9}$$

Next, the students belong to the θ-th cluster $Cluster_\theta$ are regarded as possible outliers since $Cluster_\theta$ own the fewest students compared to other clusters. This way, through the above TOD method (formalized with Eqs. (1) and (9)), we can successfully achieve our goal of outlier detection in finding abnormal students whose health physique conditions are different with other students. The major advantage of TOD method is that it can efficiently process the time-aware health physique monitoring data of students since it uses lightweight index mechanism for student clustering and outlier detection.

Regarding the time granularity adopted in this paper, we did not discuss more details in TOD since TOD can be applied to most time-aware scenarios as long as an identical time granularity is adopted in the scenario. In other words, no matter data are produced with a time granularity of second or minute or hour or day or week or month or year, TOD can always work well in detecting possible outliers with time stamps.

Experiments

Experiment setup

In this section, our proposed TOD method is compared with another four methods which are introduced briefly as follows, respectively. Experiments are based on WS-DREAM dataset (https://wsdream.github.io/) which is a time-aware service quality monitoring dataset. Evaluation metrics include MAE which indicates outlier detection accuracy and time cost which reflects outlier detection efficiency. Experiments are running under Win 11 operation system and python 3.10.0. Hardware settings include Intel(R) Core(TM) i5-1235U 2.50 GHz and 16.0 GB RAM. Each set of experiments are repeated 100 times to register their average performances.

(1) SI-CF [28]: Similar Items-based Collaborative Filtering.
(2) SU-CF [29]: Similar Users-based Collaborative Filtering.
(3) TL-CF [30]: Time-aware and Location-aware Collaborative Filtering.
(4) WSWalker [31]: Random walk with Location-aware Collaborative Filtering.

Performance comparison

Detection accuracy comparison

The accuracy of different methods is measured by MAE which belongs to "smaller is better" rule. The parameter values of TOD method in this profile are as follows: the size of student set, i.e., m is varied from 100 to 500; the size of time period set, i.e., n is varied from 1000 to 5000; size of function set in Eq. (6), i.e., $k = 10$; size of table set in Fig. 3, i.e., $K = 5$. MAE comparison results are reported in Fig. 4 which contains two figures: Fig. 4(a) shows the MAE values of different methods with respect to m (here, $n = 5000$); Fig. 4(b) shows the MAE values of different methods with respect to n (here, $m = 500$). As Fig. 4 reflects, the MAE values of five methods all vary with the increment of parameters m and n which means that these methods are all data-driven and data-related. Furthermore, compared with the four methods (i.e., SI-CF, SU-CF, TL-CF and WSWalker), our proposed TOD method performs the best in terms of accuracy. The advantage of TOD's accuracy is attributed to the fact that Locality-Sensitive Hashing used in TOD is an effective similar object searching technique and therefore, can guarantee a more accurate student clustering result as well as a more accurate outlier detection result.

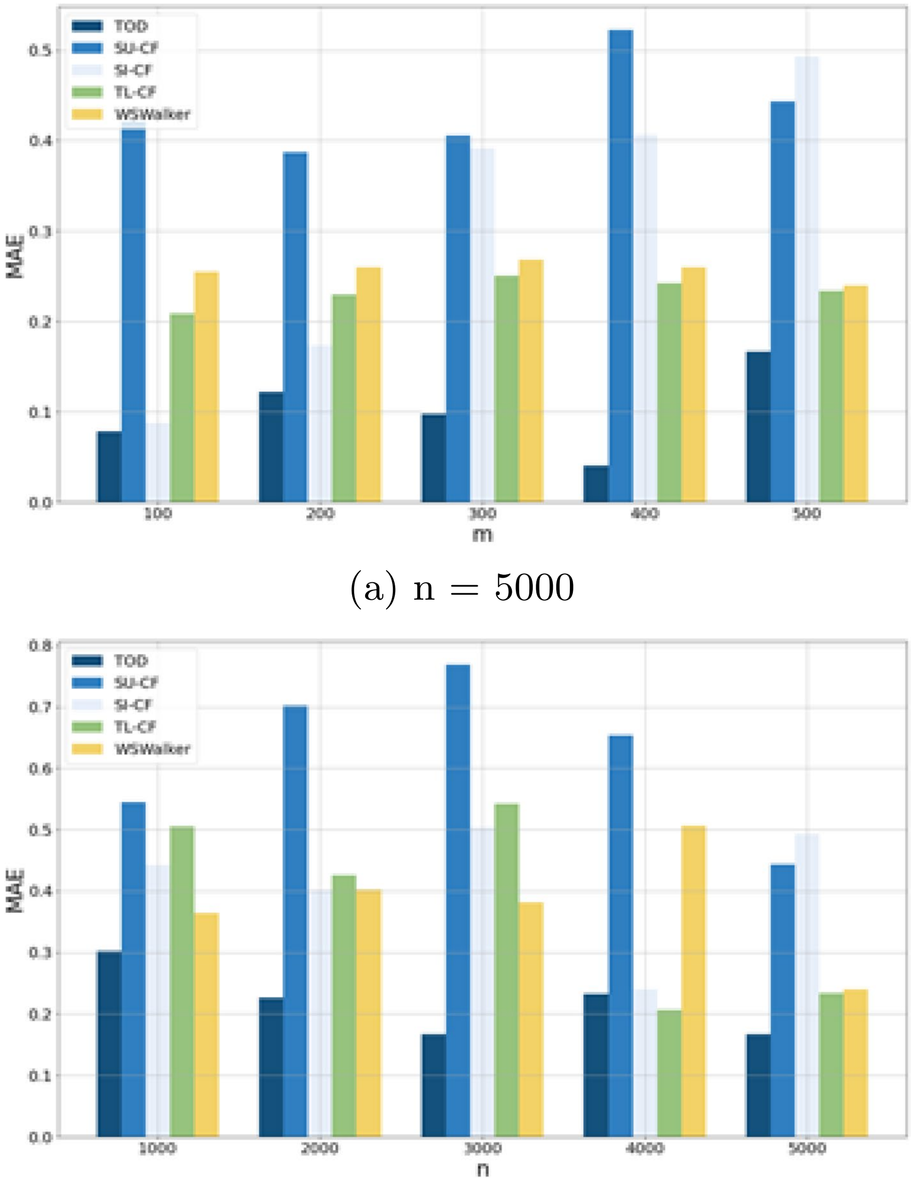
Detection efficiency comparison

As we introduced in the paper contribution part, one major contribution of our TOD method is its low time cost in tackling big data issues. To validate the efficiency advantage of TOD, we have devised a set of experiments and compare TOD's time cost with other baselines. The parameter setting is the same as that in the above profile, which is not repeated here. In concrete, the time cost comparison results of five methods are reported in Fig. 5 which also contains two figures: Fig. 5(a) shows the computational time of different methods with respect to m (here, $n = 5000$); Fig. 5(b) shows the computational time of different methods with respect to n (here, $m = 500$). As observed from Fig. 5, the time costs of five methods all grow with the rising of parameters m and n; this is because a larger m or n often means more student health data need to be processed, analyzed and mined in order to find potential outliers; correspondingly, more processing time is needed in every method. Moreover, our TOD method outperforms the other four baselines (i.e., SI-CF, SU-CF, TL-CF and WSWalker) significantly. This is due to the fact that the student indices based on Locality-Sensitive Hashing can be created and stored beforehand; while outlier detection operation only needs to read the stored indices and then cluster students based on student indices, whose time cost is rather small [32–34]. Hence, we can conclude that TOD method is suitable for time-aware outlier detection scenarios with big data.

Ablation study

Next, we observe the performances of TOD method with respect to two key parameters: size of function set in Eq. (6), i.e., k and size of table set in Fig. 3, i.e., K. Here, k is varied from 2 to 10 with a step-size of 2, K is varied from 5 to 25 with a step-size of 5. Other parameters are set as follows: $m = 500$, $n = 5000$. Concrete experiment results are shown in Fig. 6 which also includes two figures: Fig. 6(a) depicts the MAE variation trend of TOD with respect to k and K, while Fig. 6(b) reports the time cost variation trend of TOD with respect to k and K.

In Fig. 6(a), we can see an approximate decline of TOD's MAE with the increment of k; this observation may be caused by the relationship between the clustering condition and parameter k. More concretely, as Eq. (6) shows, each student index is corresponding to a k-dimensional vector. And theoretically, a larger k value often leads to a longer index vector as well as the resulted stricter evaluation conditions in student clustering and outlier detection. In this situation, only the really similar students with close health conditions will be put into an identical cluster; therefore, the accuracy of clustering and outlier detection is enhanced accordingly. Another observation from Fig. 6 is that TOD's MAE approximately increases with the growth of K; this observation may be caused by the relationship between the clustering condition and parameter K. More concretely, as Fig. 3 shows, each student index is corresponding to K tables containing student health indices. And theoretically, a larger K value often leads to a looser evaluation condition in student clustering and outlier detection. In this scenario, students with dissimilar health conditions may be grouped into the same cluster, leading to a reduction in the accuracy of both clustering and outlier detection.

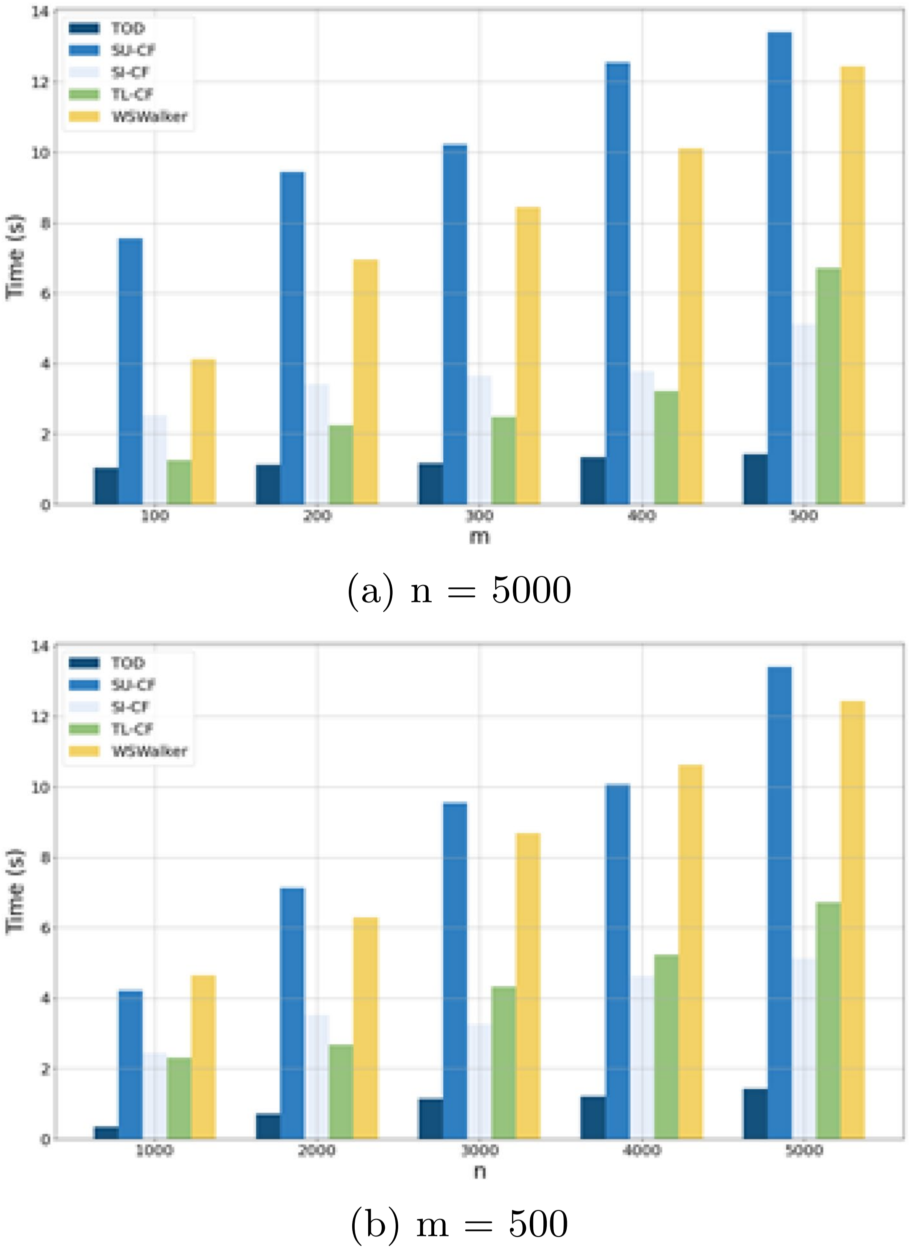


(a) n = 5000

(b) m = 500

Fig. 4 Outlier detection accuracy comparison results

Low latency is a common requirement in data-intensive applications with network communications [35–39]. As we analyzed before, TOD's performances including computational cost is highly related to parameters of k and K. Their relationships are demonstrated in Fig. 6(b), where the time cost of TOD approximately increase with the growth of both k and K. This finding can be explained as follows: as Eqs. (6)-(8) and Fig. 3 show, a larger k or K often means more clusters according to student health indices while each cluster needs to count the students who belong to the cluster. As a result, when k or K grows, the number of clusters as well as the statistics burden are increased accordingly. Therefore, the time cost of TOD rises with the growth of k or K. However, as Fig. 6(b) shows, the time cost of TOD is generally small. This low time complexity indicates that our proposal can be applied to general outlier detection scenarios even in the big data context.



(a) n = 5000

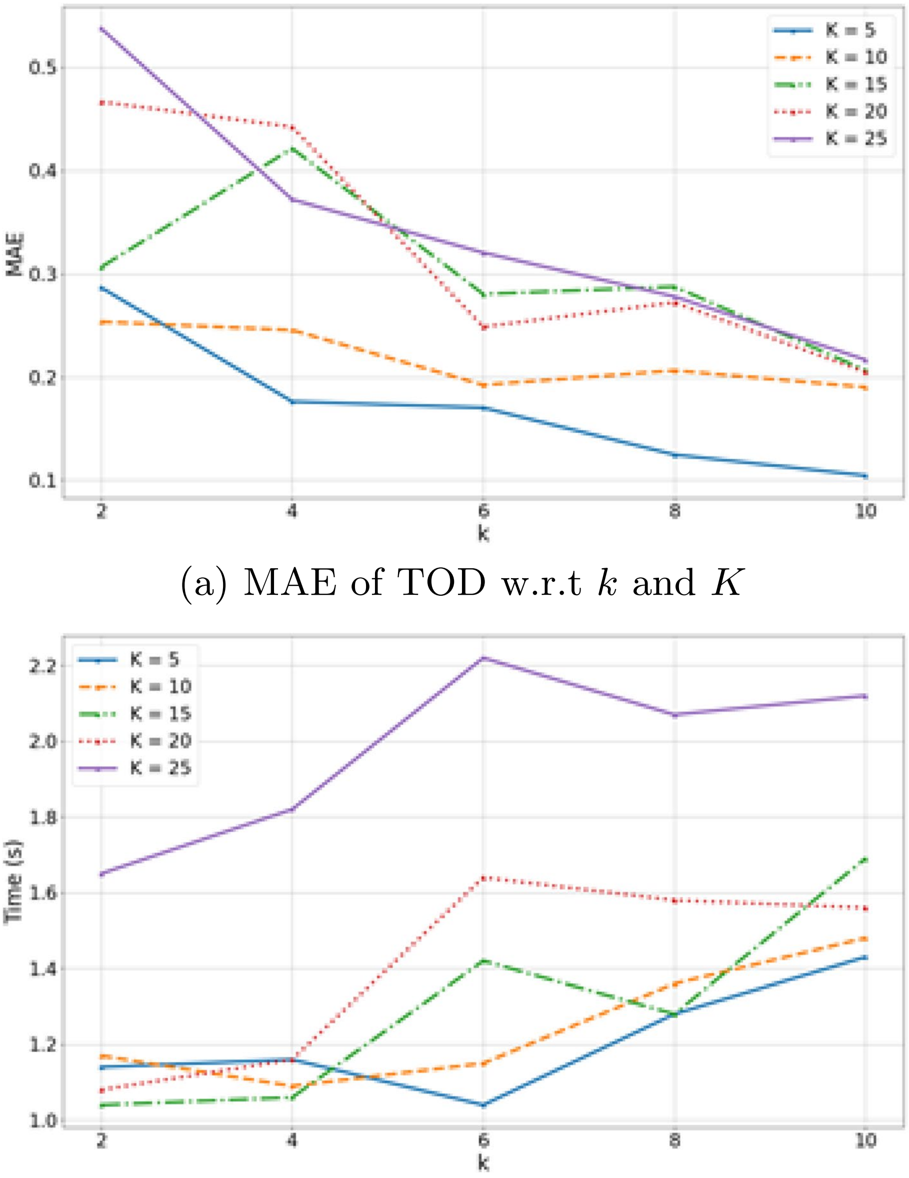
(b) m = 500

Fig. 5 Outlier detection efficiency comparison results

Conclusion and future work

The research presented in this paper successfully addresses the challenge of efficiently processing and analyzing massive physiological data in the context of sport education. Our proposed time-aware outlier detection method, TOD, leverages the principles of Locality-Sensitive Hashing to effectively condense extensive physiological data into a manageable set of student health indices. This approach not only eases the burden of data storage and processing but also enables the precise and rapid identification of potential health outliers among students. The simulated experiments conducted using the WS-DREAM dataset have demonstrated the practicality and superiority of TOD over existing methods. These findings underscore the potential of TOD in enhancing health monitoring and evaluation in educational settings, paving the way for more responsive and targeted interventions for student health and well-being.

However, there are still several shortcomings or limitations in TOD. First of all, TOD can only process the



(a) MAE of TOD w.r.t k and K

(b) Time cost of TOD w.r.t k and K

Fig. 6 Ablation study: TOD's performances w.r.t. *k* and *K*

time-aware user health data, without incorporating more complex health data with other valuable context information such as user location. In addition, Locality-Sensitive Hashing adopted in TOD is essentially a probability-based approximate neighbor search technique and can lead to false-positive or false-negative outlier detection results inevitably. In the future work, we will further improve our proposed TOD method for outlier detection by including more context factors such as the dynamic movement trajectory information of people for better outlier detection performances. In addition, Locality-Sensitive Hashing used in TOD method is still a probability-based fast data search technique; therefore, how to minimize the false-positive or false-negative probability existed in outlier detection is another research direction that calls for intensive study.

Abbreviations

TOD	Time-aware Outlier Detection
O-DRNN	Optimized Deep Recurrent Neural Network
CNN	Convolutional Neural Network
LEWIS	Low-cost Efficient Wireless Intelligent Sensor
SDN	Software-Defined Networking
OOD	Out-Of-Distribution;

STM	Student-Time health Matrix
SI-CF	Similar Items-based Collaborative Filtering
SU-CF	Similar Users-based Collaborative Filtering
TL-CF	Time-aware and Location-aware Collaborative Filtering
WSWalker	Random walk with Location-aware Collaborative Filtering

Acknowledgements
The authors show special thanks to the providers of WS-DREAM dataset for experiment validation.

Authors' contributions
Y. L: English writing and conceived idea. L. K: Developed research motivation and established the model. Z. L: Conceived experimental ideas and designed methodologies. F. J: Wrote the initial draft and executed the experimental plan. N. B: Reviewed and revised the initial draft, validated experimental results. T. D: Conducted a literature review for related work, providing comparative analysis. M. A: Shaped research ideas and enriched English writing.

Funding
Not applicable.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The coauthors all agree on the paper publication.

Competing interests
The authors declare no competing interests.

Received: 29 January 2024 Accepted: 15 March 2024
Published online: 25 March 2024

References

1. Diao G, Liu F, Zuo Z, Moghimi MK (2022) Privacy-aware and efficient student clustering for sport training with hash in cloud environment. J Cloud Comput 11(1):1–11
2. Zhou X, Ye X, Kevin I, Wang K, Liang W, Nair NKC, Shimizu S, Yan Z, Jin Q (2023) Hierarchical federated learning with social context clustering-based participant selection for internet of medical things applications. IEEE Trans Comput Soc Syst 10(4):1742–1751
3. Wu X, Zhou J, Zheng M, Chen S, Wang D, Anajemba J, Zhang G, Abdelhaq M, Alsaqour R, Uddin M (2022) Cloud-based deep learning-assisted system for diagnosis of sports injuries. J Cloud Comput 11(1):1–18
4. Liu Y, Zhou X, Kou H, Zhao Y, Xu X, Zhang X, Qi L (2023) Privacy-preserving point-of-interest recommendation based on simplified graph convolutional network for geological traveling. ACM Trans Intell Syst Technol https://doi.org/10.1145/3620677
5. Qi L, Liu Y, Zhang Y, Xu X, Bilal M, Song H (2022) Privacy-aware point-of-interest category recommendation in internet of things. IEEE Internet Things J 9(21):21398–21408
6. Xu X, Li H, Li Z, Zhou X (2022) Safe: synergic data filtering for federated learning in cloud-edge computing. IEEE Trans Ind Inform 19(2):1655–1665
7. Qi L, Xu X, Wu X, Ni Q, Yuan Y, Zhang X (2023) Digital-twin-enabled 6g mobile network video streaming using mobile crowdsourcing. IEEE J Sel Areas Commun 41(10):3161–3174
8. Wang F, Zhu H, Srivastava G, Li S, Khosravi MR, Qi L (2021) Robust collaborative filtering recommendation with user-item-trust records. IEEE Trans Comput Soc Syst 9(4):986–996
9. Xu X, Tang S, Qi L, Zhou X, Dai F, Dou W (2023) Cnn partitioning and offloading for vehicular edge networks in web3. IEEE Commun Mag 61(8):36–42
10. Zhou X, Yang Q, Zheng X, Liang W, Kevin I, Wang K, Ma J, Pan Y, Jin Q (2024) Personalized federation learning with model-contrastive learning for multi-modal user modeling in human-centric metaverse. IEEE J Sel Areas Commun 42(4):817–831
11. Zhang S, Tang Y, Wang D, Karia N, Wang C (2023) Secured sdn based task scheduling in edge computing for smart city health monitoring operation management system. J Grid Comput 21(4):1–14
12. Pavithra D, Nidhya R, Shanthi S, Priya P (2023) A secured and optimized deep recurrent neural network (drnn) scheme for remote health monitoring system with edge computing. Automatika 64(3):508–517
13. Stephen DK, Dr. VM, Manalang AR, et al (2023) IOT-based generic health monitoring with cardiac classification Using edge computing[J]. J Internet Serv Inf Secur 13(2):128–145
14. Gowrishankar V, Jayakumar T, Parameswaran S, Senthilkumar M, Lekashri S, Kumar R (2023) Patient health monitoring using fog and edge computing. In: 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), IEEE, pp 250–256
15. Deng J, He J, Duan X, et al (2022) Application research of advanced intelligent big data analysis based on intelligent sensor network in the design of personalized education management system and the construction of innovation system[J]. J Sens 2022:7113098
16. Sanei M, Atcitty S, Moreu F (2023) Low-cost efficient wireless intelligent sensor (lewis) for engineering, research, and education. arXiv preprint arXiv:2303.13688
17. Lamb R, Choi I, Owens T (2023) Artificial intelligence and sensor technologies the future of individualized and differentiated education[J]. Int J Psychol Neurosci 9(1):30–36
18. Alampalli S, Malek K, Mohammadkhorasani A, Moreu F (2023) Low-cost efficient wireless intelligent sensor (lewis) deployment for community driven decision making. Struct Health Monit
19. Shekhar S, Shah N, Akoglu L Fairod: Fairness-aware outlier detection[C]//Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021:210–220
20. Huang JW, Zhong MX, Jaysawal BP (2020) Tadilof: Time aware density-based incremental local outlier detection in data streams. Sensors 20(20):5829
21. Ahmad H, Dowaji S (2018) A novel framework for context-aware outlier detection in big data streams. J Digit Inf Manag 16(5):213–222
22. Yang P, Liang J, Cao J, He R (2023) Auto: Adaptive outlier optimization for online test-time ood detection. arXiv preprint arXiv:2303.12267
23. Park CG, Ahn H (2019) Temporal outlier detection and correlation analysis of business process executions. IEICE Trans Inf Syst 102(7):1412–1416
24. Wahyuni E, Surono S, Eliyanto J (2021) Outlier detection using k-means clustering with minkowski-chebyshev distances for inquiry-based learning results in students dataset. In: 2021 International Conference on Artificial Intelligence and Big Data Analytics, IEEE, pp 1–5
25. Kirielle N, Christen P, Ranbaduge T (2019) Outlier detection based accurate geocoding of historical addresses. In: Data Mining: 17th Australasian Conference, AusDM 2019, Adelaide, SA, Australia, December 2–5, 2019, Proceedings 17, Springer, pp 41–53
26. Yau SC, Bakar JA, Abdullah AA, Harun H, Rasli RM, Yang LZ, Mun ETY (2021) Detection of topic on health news in twitter data. Emerg Adv Integr Technol 2(2):23–29
27. Zhou H, Dai X, Lou L, Zhou C, Zhang W (2021) Association of sedentary behavior and physical activity with depression in sport university students. Int J Environ Res Public Health 18(18):9881
28. Wang H, Shen Z, Jiang S, Sun G, Zhang RJ (2021) User-based collaborative filtering algorithm design and implementation. In: Journal of Physics: Conference Series, IOP Publishing, vol 1757, p 012168
29. Kharita MK, Kumar A, Singh P (2018) Item-based collaborative filtering in movie recommendation in real time. In: 2018 first international conference on secure cyber computing and communication (ICSCCC), IEEE, pp 340–342
30. Yu C, Huang L (2016) A web service qos prediction approach based on time-and location-aware collaborative filtering. SOCA 10:135–149
31. Tang M, Dai X, Cao B, Liu J (2015) Wswalker: A random walk method for qos-aware web service recommendation. In: 2015 IEEE International Conference on Web Services, IEEE, pp 591–598

32. Kong L, Wang L, Gong W, et al (2022) LSH-aware multitype health data prediction with privacy preservation in edge environment[J]. World Wide Web 25:1793–1808
33. Hu C, Fan W, Zeng E, Hang Z, Wang F, Qi L, Bhuiyan MZA (2021) Digital twin-assisted real-time traffic data prediction method for 5g-enabled internet of vehicles. IEEE Trans Ind Inform 18(4):2811–2819
34. Qi L, Hu C, Zhang X, Khosravi MR, Sharma S, Pang S, Wang T (2020) Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment. IEEE Trans Ind Inform 17(6):4159–4167
35. Dai H, Xu Y, Chen G, Dou W, Tian C, Wu X, He T (2020) Rose: Robustly safe charging for wireless power transfer. IEEE Trans Mob Comput 21(6):2180–2197
36. Gu R, Chen Y, Liu S, Dai H, Chen G, Zhang K, Che Y, Huang Y (2021) Liquid: Intelligent resource estimation and network-efficient scheduling for deep learning jobs on distributed gpu clusters. IEEE Trans Parallel Distrib Syst 33(11):2808–2820
37. Dai H, Wang X, Lin X, Gu R, Shi S, Liu Y, Dou W, Chen G (2023) Placing wireless chargers with limited mobility. IEEE Trans Mob Comput 22(6):3589–3603
38. Gu R, Zhang K, Xu Z, Che Y, Fan B, Hou H, Dai H, Yi L, Ding Y, Chen G, et al (2022) Fluid: dataset abstraction and elastic acceleration for cloud-native deep learning training jobs. In: 2022 IEEE 38th International Conference on Data Engineering (ICDE), IEEE, pp 2182–2195
39. Dai H, Yu J, Li M, Wang W, Liu AX, Ma J, Qi L, Chen G (2023) Bloom filter with noisy coding framework for multi-set membership testing. IEEE Trans Knowl Data Eng 35(7):6710–6724

Publisher's Note