applied sciences

MDPI

Article

TSPADS: A Transformer-Based System for Proactive Student Physical Health Monitoring

Hanzi Zhu [1,2], Minghao Li [1,3], Xin Jiang [1,4], Qian Chen [1,3], Shiheng Ma [1,3], Xiaolei Zhang [5], Huiying Xu [1,3,*] and Xinzhong Zhu [1,3]

[1] Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University, Jinhua 321004, China; miazhz0817@outlook.com (H.Z.); liminghao@zjnu.edu.cn (M.L.); jiangxin10125@gmail.com (X.J.); chenqian2002@zjnu.edu.cn (Q.C.); 13172677363@163.com (S.M.); zxz@zjnu.edu.cn (X.Z.)
[2] Research Institute of Hangzhou Artificial Intelligence, Zhejiang Normal University, Hangzhou 311100, China
[3] College of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China
[4] College of Education, Zhejiang Normal University, Jinhua 321004, China
[5] Academic Affairs Office, Zhejiang Normal University, Jinhua 321004, China; znuzxl@zjnu.cn
* Correspondence: xhy@zjnu.edu.cn

Abstract

With the continuous growth in the scale of student physical health (SPH) monitoring data, annually sampled time-series records provide a valuable foundation for risk early warning. However, traditional models often fail to capture multi-year developmental trajectories of individuals, resulting in delayed intervention for students at potential health risk. This study aims to develop a Transformer-based Student Physical Anomaly Detection System (TSPADS), which is a dedicated intelligent software system to enable effective and timely anomaly detection in SPH data. The proposed TSPADS is built on the Transformer architecture and incorporates a novel masked anomaly-attention mechanism to learn implicit long-span dependencies in SPH data. A density-based clustering algorithm is then applied to distinguish anomalies and automatically generate hierarchical warning signals. Comprehensive experiments were conducted on a public multimodal movement and health dataset. The results demonstrate that TSPADS achieves high effectiveness and efficiency in both anomaly detection and classification tasks. The system shows strong potential to assist educational administrators and physical education teachers in providing timely, personalized health guidance, thereby addressing a critical gap in existing student health monitoring approaches.

Keywords: anomaly detection; time series analysis; student health monitoring; deep learning; educational data mining

Check for updates

Received: 15 February 2026
Revised: 8 March 2026
Accepted: 13 March 2026
Published: 16 March 2026

1. Introduction

The deep integration of computer technology and software tools has become a significant driving force in innovation in science, technology, engineering, and education. Students' physical health (SPH) has always been a concern in both the education and public health areas, as it is closely related to the comprehensive development of adolescents and the future reserve of national human resources. However, large-scale longitudinal monitoring data show that several key physical fitness measures have sharply declined, especially in adolescents' cardiorespiratory endurance in recent decades [1–4]. Evidence from [4] indicated that the pandemic-induced deterioration in cardiorespiratory endurance among adolescents persists, even after the removal of COVID-19 restrictions. Furthermore, the research findings in [1] demonstrate a consistent downward trend from 2010 to 2019 in boys' physical capacities, including flexibility, explosive power, and muscular strength. Those phenomena reveal potential downward risks to the future physical fitness of adolescents. Meanwhile, multidimensional physical fitness data of student populations are continuously collected and accumulated on an unprecedented scale Such rich multidimensional time-series data provide a solid foundation for in-depth analysis of changes in SPH. Nevertheless, traditional fitness evaluation and analysis methods are relatively outdated, mainly focusing on annual static scores and one-time pass/fail judgments [5,6]. These methods ignore individual temporal trends, making it difficult to quickly detect and warn potential anomalies. Therefore, in the dual background of national emphasis and data proliferation, it is necessary to explore more advanced time-series analysis methods to monitor the evolution of SPH.

The current SPH system easily ignores dynamic changes in data that fail to comprehensively reflect developmental pathways of students' physique. Typically, even if the performance of a student in a given year meets the passing standard, an abnormal decrease compared to previous years might be masked by static evaluation. Conversely, a student with weaker foundational skills who makes significant progress in a given year might still be categorized as failing under static criteria, failing to capture the extent of progress [7,8]. In addition, fitness is influenced by various factors, resulting in vastly different health change patterns between students. The authors of [9] verified both fat-free mass, chronological age, and age at peak height velocity as significant determinants of physical fitness performance in students. Similarly, as noted by refs. [10,11], behavioral and lifestyle factors such as learning motivation and living habits can also impact SPH. Fluctuations that fall within the normal range according to population-level standards may actually represent abnormal signals for specific individuals. However, traditional methods lack comprehensive interrelationships among multidimensional metrics, making it difficult to adopt the complexity and individualization of time-series SPH monitoring data [12,13]. These limitations have driven us to search for more intelligent and dynamic anomaly detection methods.

Deep learning models designed for time-series data have provided a new technical approach of monitoring SPH in recent years [14–17]. In the classic recurrent neural network (RNN) architecture, state information is recursively updated across a sequence, where the resulting prediction or reconstruction loss is used as evidence of anomalies. However, RNN is prone to vanishing gradients during the training process, making it difficult to stably retain information across long time intervals. Furthermore, their inherently sequential recursive computation limits parallelization, resulting in reduced training and inference efficiency in long-sequence scenarios. These drawbacks hinder its use in applications requiring the real-time monitoring of long-span SPH data. As a key architectural innovation, Transformer replaces recursive structures with a self-attention mechanism, thereby eliminating time-step recursion for sequence modeling. This mechanism not only facilitates parallelization, but also directly establishes connections between any two points in time. For multidimensional SPH data, multi-head attention can capture various interactions across different subspaces, enabling the model to learn both long-term trends and multidimensional covariations within a unified framework. The variant of Transformer for time-series tasks has also developed a relatively systematic research lineage, which provided a reusable technical foundation for modeling SPH data with long sequences and complex patterns.

Nevertheless, the application of Transformer-based technology to anomaly detection and early warning of SPH still faces two key challenges: (1) SPH data are usually collected annually, with the time span covering the entire academic period or even longer. Although the sequence length may not be large, the data generally suffer from problems such as long time intervals, missing measurements, and unevenness, which will weaken the model's ability to capture long-term dependencies and trends of slow variables. (2) Existing methods typically produce binary results (normal or abnormal), while frontline management require more detailed explanations regarding the types and causes of abnormalities. In response to these challenges, we introduce a Transformer-based Student Physical Anomaly Detection System (TSPADS). Through an improved attention mechanism, our TSPADS model captures meaningful patterns in long-span longitudinal SPH data better than existing deep learning and Transformer-based methods. To enable graded warnings for abnormal physical conditions, an anomaly discrimination and grading module that fuses DBSCAN clustering with maximum similarity matching, distinguishing between different severity levels and types of anomalies. This establishes a basis for proactive prevention and intervention of student health issues. In summary, this study makes the following innovations and contributions based on the technical framework mentioned above:

- We present a novel system named TSPADS for anomaly detection in time-series SPH data, which facilitates the analysis of long-sequence fitness signs and proactive health management.
- We construct an anomaly discrimination and classification module that fuses clustering analysis to realize fine-grained classification of anomalies. Compared to traditional single anomaly alerts, our system can further distinguish between different types or patterns, thereby enhancing the interpretability and practicality of the results.
- We explore the potential application scenarios of TSPADS in the management of sports science education. The results of the experiment on the public motion monitoring dataset validate its superior performance.

2. Related Work

We have divided the relevant research into two main sections. First, we explore the characteristics and challenges of anomaly detection data in physical education. Next, we discuss machine learning methods applied to anomaly detection.

2.1. Characteristics of SPH Data and Challenges in Anomaly Detection for Physical Education

In physical education (PE), students typically undergo physical health assessments annually, generating a continuous longitudinal data record that spans multiple years. Consequently, SPH data are commonly represented in the form of time series [18,19], exhibiting significant temporal dependencies as well as multidimensional complexity due to the diverse composition of fitness indicators. In PE practice, these longitudinal records are not only used for reporting, but also serve as evidence to understand students' development trajectories and to inform timely instructional support. The following features of this data pose substantial challenges for anomaly detection.

Firstly, the SPH data demonstrate strong temporal characteristics, which increase the difficulty of identifying anomalous patterns over the years. Specifically, physical fitness indicators exhibit certain regularities and trends as they age and progress through different stages of physical development [20]. If a student's performance in a particular year differs significantly from their historical trend, this can signal a potential anomaly. To successfully identify anomalous patterns that emerge over multiple years, models must have strong long-range memory capabilities [21].

Secondly, SPH data are multidimensional and interdependent, making traditional univariate methods inadequate for effectively characterizing complex abnormal patterns. SPH data typically contain multiple physical indicators that often show complex correlations among these indicators and specific combination patterns. Guan et al. [22] noted that under normal circumstances, different fitness indicators tend to develop in a coordinated manner or maintain a form of dynamic equilibrium. Consequently, anomalies may not only present as extreme values in individual indicators but are more likely to emerge as imbalances in coordination among multiple indicators. It is necessary to develop detection models capable of comprehensively analyzing and modeling relationships among multidimensional indicators.

Thirdly, SPH data inherently contain measurement noise and variability, which significantly amplify uncertainty and error risk in anomaly detection results. As a result, how to effectively utilize relevant indicators and eliminate irrelevant or noisy dimensions during anomaly detection has become a huge challenge [23]. This issue is corroborated by the research of King-Dowling et al. [24], who demonstrate that SPH data are inevitably affected by measurement errors and situational factors, such as the transient physical or psychological states of students during the testing. Therefore, anomaly detection methods must exhibit robustness, allowing them to effectively distinguish short-term, random fluctuations, and meaningful abnormal patterns.

Finally, category imbalance and group differences further intensify the difficulty of establishing a stable normal baseline. On the one hand, in real-world student populations, the majority of individuals exhibit normal ranges of physical development, while the proportion of students with genuine health abnormalities is relatively low [25]. On the other hand, there are significant individual differences between the student group in terms of age, sex, and regional background, leading to substantial variation in the normal fitness ranges between different subgroups. Therefore, an effective anomaly detection system needs to take into account both population-level modeling and individualized analysis. In other words, such systems should establish a reasonable global normal pattern while simultaneously detecting deviations relative to each individual's historical baseline [26].

2.2. Applications of Machine Learning Methods in Anomaly Detection

Before the rise of deep learning, numerous classical statistical and machine learning methods had already been widely applied to anomaly detection tasks [27]. Common methods include simple statistical threshold detection, principal component analysis (PCA), k-nearest neighbors (KNN), local outlier factor (LOF), and one-class support vector machines (OC-SVM) [28,29]. However, they suffer from several notable limitations.

In practical applications, researchers typically need to manually select or construct features based on experience to suit specific algorithms. However, in the context of human physical health, which is a highly complex and dynamically evolving system, artificially designed features can miss key information or introduce overly simplistic assumptions, limiting the model's ability to capture the real abnormal patterns [30]. The second limitation concerns the expressive capacity. Although traditional machine learning models are good at handling linear relationships, they are insufficient in their capacity to automatically model non-linear relationships and high-dimensional interactions [31]. The third limitation stems from the implicit assumption that the training data distributions remain approximately stationary over time. Once the data distribution changes, the decision boundaries of the old model can lead to false alarms or incorrectly classify true anomalies as normal instances [32]. Beyond these methodological constraints, the SPH data face a significant practical challenge: the severe scarcity of labeled anomalous samples. Therefore, when applied to SPH data, deep learning models hold considerable promise to improve both detection accuracy and robustness [33].

Early deep temporal models were predominantly based on recurrent neural networks (RNNs) and their variants, such as long short-term memory (LSTM) networks and gated recurrent units (GRUs). RNNs employ recurrent hidden-state connections to retain historical information, while LSTM networks integrate gating mechanisms to address gradient disappearance issues associated with long-term dependencies, enabling the learning of patterns over extended time spans [34]. Nevertheless, LSTMs also exhibit limitations. As series length increases, computational costs increase substantially, and training and inference speeds are gradually hindered by recursion. Moreover, the interdependence between the data is likely to be diluted and forgotten over extended sequences [35].

The introduction of the Transformer model marked a revolutionary shift in temporal modeling. The Transformer is built on self-attention mechanisms that enable pairwise interactions across all positions within a sequence. Furthermore, these mechanisms incorporate the multi-head mechanism and positional encoding. Therefore, the Transformer is more conducive to capture long-range dependencies and complex global patterns through more effective parallel training [36].

In anomaly detection, Transformer models have demonstrated remarkable performance and flexibility. Zhang et al. [37] employed a Transformer encoder for semi-supervised network traffic anomaly detection, using multi-head attention to simultaneously model temporal and structural patterns, achieving superior performance on multiple real-world datasets. Xu et al. [38] proposed the Anomaly Transformer model, introducing an attention mechanism to measure differences in temporal point correlations, which can learn the correlation maps of normal patterns from time series, thereby more accurately distinguishing anomalies. Tuli et al. [39] developed TranAD, which combines self-attention encoders with adversarial training to efficiently detect and diagnose anomalies in multivariate time series. These advances fully demonstrate that the evolution from traditional machine learning methods to Transformer models represents a major adaptive leap in time-series anomaly detection.

Despite these advantages, introducing Transformer also requires addressing some practical issues. One major issue concerns the mismatch between model complexity [40] and data availability. In addition, deep anomaly detection systems usually output continuous anomaly scores that must be converted into alarm signals via thresholding.

3. Method

In this section, we first introduce the foundational background about the transformer architecture. We subsequently present the framework of our transformer-based student physical anomaly detection system (TSPADS), followed by a thorough exposition of its detection model design.

3.1. Preliminaries

As a major breakthrough in natural language processing (NLP), the Transformer remains similar to previous neural sequence models. It continues to adopt an encoder-decoder architecture, the encoder extracts abstract features from the data, while the decoder reconstructs it. The objective is to learn the required mapping $f(x)$. The encoder maps an input sequence of symbols $(x^1, x^2, \ldots, x^n)$ to a sequence of continuous representations $(z^1, z^2, \ldots, z^n)$. The decoder then generates a new output sequence $(y^1, y^2, \ldots, y^n)$ based on these continuous representations. Unlike common RNN architectures, the Transformer replaces recurrent connections with self-attention mechanisms. This design allows the model to perform parallel computations, significantly improving the data processing efficiency [41]. Furthermore, the Transformer utilizes global information rather than just sequential data. The addition of Positional Embeddings enables the model to capture relationships between different positions within the sequence [42].

https://doi.org/10.3390/app16062851

3.2. System Architecture

Traditional models, including RNNs and their variants, struggle to capture long-span dependencies across multi-year developmental trajectories, often diluting or forgetting early health signals that are crucial for identifying gradual decline patterns. Moreover, existing methods are difficult to distinguish between different severity levels or types of anomalies, providing no basis for hierarchical response protocols. Thus, we propose an innovative intelligent SPH monitoring system based on the Transformer model. This system enables dynamic detection and early warning of health data, using time-series health data to accurately identify students with physical anomalies. Figure 1 illustrates the overall framework of the system.

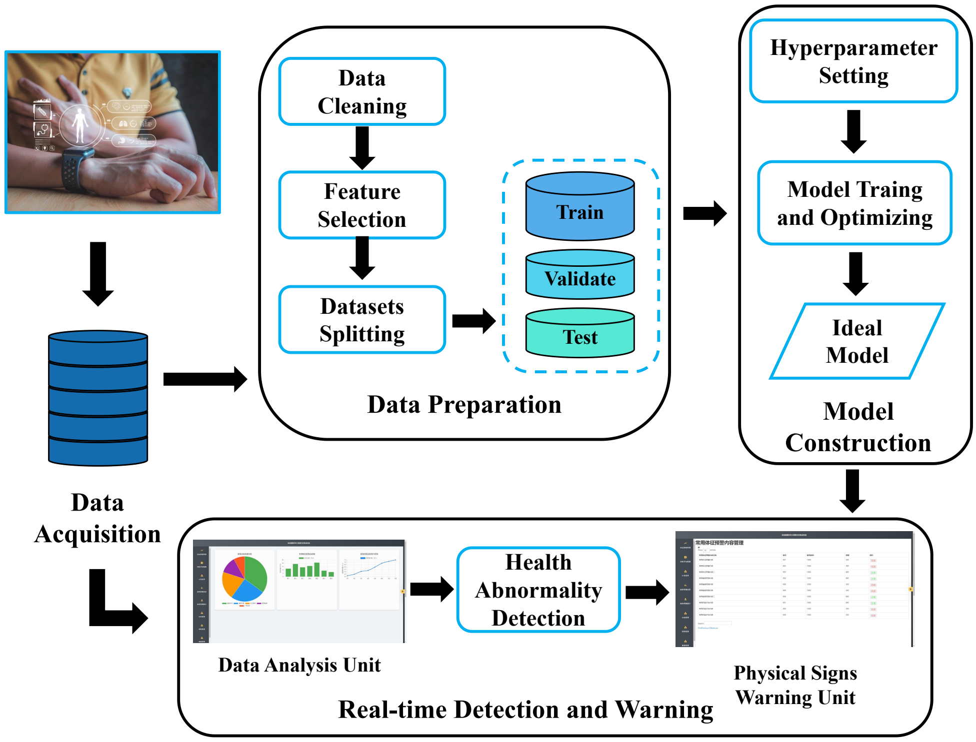


Figure 1. Framework of the physical health anomaly detection system. It consists of three parts: data preparation, model construction, and real-time detection and warning.

The system comprises three main components: the data preparation module, the model construction module, and the real-time detection and warning module. In the data preparation module, collected raw data undergoes three processes: data cleaning, feature extraction, and dataset splitting. A critical function of the data cleaning process is to address the challenges posed by missing or irregular data. We first detect missing values and irregular time intervals. For continuous variables, missing data are imputed using interpolation methods based on neighboring time points to preserve temporal trends. For irregularly sampled data, we apply a resampling and alignment procedure to standardize the time series to a fixed frequency, ensuring consistent input for the subsequent Transformer model. This robust data preparation ensures that the downstream model can effectively learn from incomplete or unevenly spaced real-world data. Subsequently, in the model construction module, we use the prepared data to train a time-series anomaly detection Transformer model. After parameter fine-tuning, the optimized model is applied to subsequent health anomaly detection. Finally, in the real-time detection and warning module, the data analysis unit is responsible for integrating and analyzing the real-time physical data. The health anomaly detection unit utilizes the adjusted model to detect abnormal signs and classify the anomalies. The physical signal warning unit then issues graded warnings to adolescents with abnormal health data. This system promotes a shift in health management from passive treatment to active prevention. It helps achieve core goals such as controlling adolescent obesity rates and improving physical fitness levels.

The time-series anomaly detection model serves as the core component of our system. It consists primarily of three parts: the input embedding and encoder stack, the masked multi-Head anomaly attention mechanism module, and the anomaly discrimination and grading Module. Specifically, the model first performs positional and token embeddings on the raw input data. Subsequently, the multi-head anomaly attention sub-layer performs deep feature extraction and reconstruction on the data. The features are then processed sequentially by a normalization sub-layer and a point-wise feed-forward sub-layer. This sequence concludes the forward propagation. In the anomaly discrimination and grading part, the reconstructed features are converted into anomaly scores by an anomaly discriminator. An initial separation between abnormal and non-abnormal data is achieved based on the DBSCAN algorithm [43]. The anomaly filter further selects anomaly samples and their corresponding scores from the training set. Subsequently, by maximizing the distribution similarity between the two, the grading of the anomaly results is completed. Algorithm 1 and Figure 2 present the pseudo-code description and the overall structure of the proposed time-series anomaly detection Transformer model.

Algorithm 1 TSPADS pseudocode

Input: training set (X, Y).
Output: binary labels Y^B, multi-levels labels Y^M and abnormal set B^A.

for $i \rightarrow 0$ until Number of Epochs **do**
    for each $x \in X$, **do**
        get vectorized x_e by Embedding(x).
        calculate hidden representation z by Encoder(x_e).
        calculate anomaly scores s by Criterion(z).
    end for
    optimize reconstruction loss.
    get Y^B by DBSCAN(S).
    filter B^A by Abnormal_Filter(S,Y).
    get Y^M by Max_Similarity(S,B^A).
end for

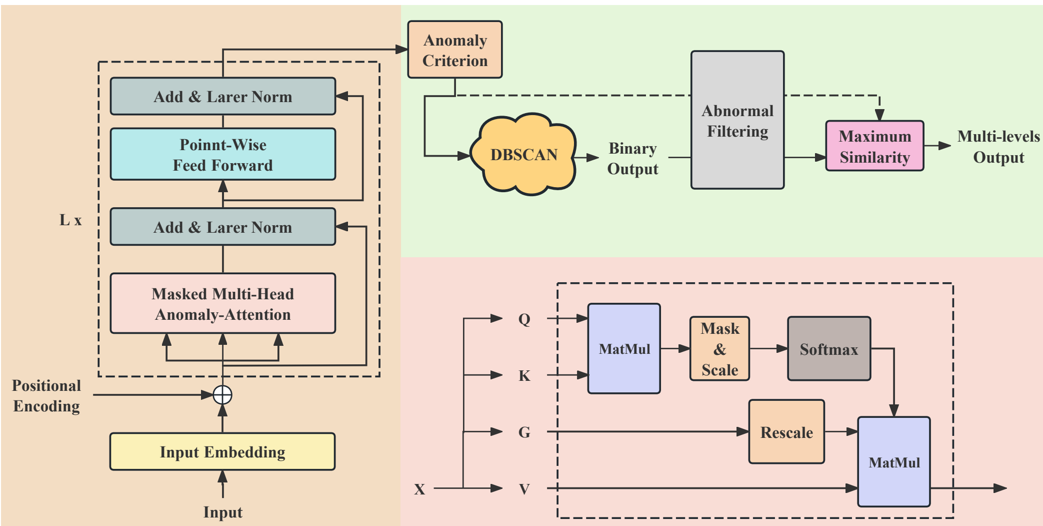


Figure 2. Overall structure of the proposed time-series anomaly detection Transformer model. Input data first passes through the input embedding layer. Subsequently, it enters the encoder stack. Finally, the anomaly discrimination and grading module outputs the anomaly detection results.

3.3. Model Design

3.3.1. Input Embedding

Before raw data enters the anomaly-attention layer, we must first embed it in vector form. Given the sequential nature of time-series data, the model incorporates additional positional embeddings alongside standard token embeddings to fully utilize location information. Since both resulting vectors share the same dimensions, they are directly added together before being inputted into the subsequent anomaly-attention layer. To enable the model to process sequences longer than those used in training while maintaining generalization, we employ sine and cosine functions of varying frequencies for encoding. In the context of student physical health data, this positional encoding serves a crucial role: it injects temporal order information into the model, allowing it to distinguish between health records from different academic years. The sine and cosine formulations ensure that the model can generalize to students with longer assessment histories than those seen during training. Specifically, cosine functions are used for odd positions, while sine functions are used for even positions. For the given channel dimension d_{model}, the specific encoding formula for the pos-th position is shown below:

$$PE(pos, 2i) = sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right). \tag{1}$$

$$PE(pos, 2i+1) = cos\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right), \tag{2}$$

3.3.2. Encoder Stack and Masked Multi-Head Anomaly-Attention Mechanism

The encoder stack comprises L encoder modules. Each module consists of a masked multi-head attention layer, a pointwise feed-forward network (FNN), residual connections, and normalization layers. The dimension of the FNN serves as a hyperparameter tunable during training. Meanwhile, residual connections and normalization facilitate information propagation and further refine sub-layer outputs. To achieve optimal performance, we set $L = 3$ and the FNN neurons to 512. For input time-series data $\mathbf{X} \in \mathbb{R}^{N \times d}$, where $\boldsymbol{N}$ represents the sequence length (e.g., the number of academic years a student has been assessed) and $\boldsymbol{d}$ represents the feature dimension (e.g., the multimodal health signals including acceleration, ECG, and gyroscope data), the normalization layer and the FNN layer are expressed below:

$$\begin{aligned} Feed\_Forward(\mathbf{X}^l) &= f\Big(Conv1d\Big(\mathbf{X}^l\Big)\Big), \\ Layer\_Norm(\mathbf{X}) &= \frac{\alpha(x-\mu)}{\delta} + \beta, \end{aligned} \tag{3}$$

where $Conv1d(\cdot)$ signifies a one-dimensional convolution operation, μ and δ represent the mean and standard deviation. α and β serve as trainable affine transformation parameters. The Layer Normalization operation adaptively scales and shifts the health features, μ and δ are the mean and standard deviation computed across the feature dimension for each time point. The trainable parameters α and β allow the model to restore expressive power after normalization, learning the appropriate relative importance of each health indicator.

Then, the overall formulation for the l-th layer can be described as follows:

$$\begin{aligned} \mathbf{Z}^l &= Layer\_Norm(Anomaly\_Attention(\mathbf{X}^{l-1}) + \mathbf{X}^{l-1}), \\ \mathbf{X}^l &= Layer\_Norm(Feed\_Forward(\mathbf{Z}) + \mathbf{Z}), \end{aligned} \tag{4}$$

where $\mathbf{X}^l, \mathbf{Z}^l \in \mathbb{R}^{N \times d_{model}}$ refers to the output of the l-th layer, and the hidden representation within the l-th layer, respectively. We define $\mathbf{X}_0 = Embedding(\mathbf{X})$ as the embedded representation of the raw input.

To simultaneously mine both prior and series associations within the time series, we employ a multi-branch attention mechanism supplemented by mask scaling to enhance generalization. For the prior association, we utilize a learnable Gaussian kernel to model the prior distribution of relative temporal distances. Leveraging its unimodal property, this design focuses on structural information within adjacent time horizons. The learnable scale parameter σ enables the prior association to adaptively fit various patterns, effectively capturing anomalies of varying lengths. In parallel, the series association employs a multi-head masked self-attention mechanism to identify effective dependencies for each time point. Unlike pointwise representations that focus on isolated moments, these associations preserve temporal dependencies and encapsulate richer information. Furthermore, while the prior branch reflects theoretical neighborhood concentration, the series branch encodes actual learned associations. The discrepancy between them serves as a robust feature for distinguishing normal from abnormal patterns. Given weight matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d_{model} \times \frac{d_{model}}{h}}$, $\mathbf{W}_O \in \mathbb{R}^{d_{model} \times d_{model}}$ and $\mathbf{W}_\sigma \in \mathbb{R}^{d_{model} \times h}$, the masked multi-head mechanism $Anomaly\_Attention(\cdot)$ for the l-th layer is formulated in two parts.

To understand how the model captures temporal health patterns, we explain the role of each matrix in the context of SPH data. The Query matrix $\mathbf{Q} \in \mathbb{R}^{N \times d_{model}}$ represents the "current" health state being analyzed at each time point. The Key matrix $\mathbf{K} \in \mathbb{R}^{N \times d_{model}}$ represents the "historical" health patterns available at each time point. The Value matrix $\mathbf{V} \in \mathbb{R}^{N \times d_{model}}$ carries the actual health indicator information to be aggregated. The mask matrix $\mathbf{M} \in \mathbb{R}^{N \times N}$ plays a critical role in unsupervised learning. This ensures that when reconstructing the health state at year i, the model cannot peek into future years $i+1, i+2, \ldots$. In SPH terms, the model must learn to understand a student's current health based solely on their past assessments, just as a human expert would.

First, the series association is defined as follows:

$$\begin{aligned} S^l &= Multihead(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Concat(head_1, \ldots, head_n)\mathbf{W}_O, \\ head_i &= MaskedAttention(\mathbf{Q}\mathbf{W}_Q^i, \mathbf{K}\mathbf{W}_K^i, \mathbf{V}\mathbf{W}_V^i), \\ MaskedAttention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= softmax(\mathbf{M} + \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{model}}})\mathbf{V}. \end{aligned} \tag{5}$$

In this context, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d_{model}}$ refers to the self-attention query, key, and value matrices, respectively. $\mathbf{M} \in \mathbb{R}^{N \times N}$ stands for the mask matrix. Then, the prior correlation is formulated as follows:

$$\mathbf{G}^l = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right) \quad i, j \in \{1, 2, \ldots, N\}. \tag{6}$$

$$\mathbf{P}^l = Rescale(\mathbf{G}^l) \tag{7}$$

In Equations (6) and (7), the correlation weights between the i-th and j-th time points are derived via a Gaussian kernel, σ_i is a learnable parameter associated with time point i. Physically, σ_i captures the expected temporal stability or volatility of health indicators at that stage. By learning different σ values for different time points and health contexts, the model adaptively determines how much temporal smoothness to expect, enabling it to distinguish between normal developmental fluctuations and genuine anomalies. Subsequently, the rescale function $Rescale(\cdot)$ is used to convert these weights into a discrete distribution, and the softmax function $Softmax(\cdot)$ is applied as normalization to the last output dimension.

https://doi.org/10.3390/app16062851

3.3.3. Anomaly Discrimination and Grading

To quantify the association, we formalize it as the symmetric Kullback–Leibler (KL) divergence between the prior and series associations. Subsequently, we average the association degrees across multiple layers to integrate them into a more informative metric for anomaly discrimination. This metric effectively captures the information gain between the prior and series association distributions, denoted as $ASSD(\mathbf{P}, \mathbf{S}; \mathbf{X})$, carries a specific physical interpretation in the SPH context. The prior association $\mathbf{P}$ represents where the model expects to find relevant information. The series association $\mathbf{S}$ represents where the model actually finds relevant information when analyzing the data. When a student's health follows a normal developmental trajectory, these two distributions should align closely. However, when an anomaly occurs, $\mathbf{S}$ will diverge from $\mathbf{P}$. The symmetric KL divergence captures this divergence, serving as a strong signal for anomaly detection. This mechanism allows TSPADS to detect anomalies not just based on absolute values, but based on disruptions in expected temporal patterns.

$$\begin{aligned} &ASSD(\mathbf{P}, \mathbf{S}; \mathbf{X}) = \frac{1}{L} \sum_{l=1}^{L} \left(KL\left(\mathbf{P}^{l}_{i,:} \parallel \mathbf{S}^{l}_{i,:}\right) + KL\left(\mathbf{S}^{l}_{i,:} \parallel \mathbf{P}^{l}_{i,:}\right) \right), \\ &i \in \{1, 2, \ldots, N\}. \end{aligned} \tag{8}$$

Specifically, $KL(\cdot \parallel \cdot)$denotes the KL divergence between discrete distributions associated with the rows of $\mathbf{P}^{l}$ and $\mathbf{S}^{l}$. The term $ASSD(\mathbf{P}, \mathbf{S}; \mathbf{X})$ represents the point-wise association discrepancy of $\mathbf{X}$ relative to these distributions, and its smaller value means a lower association.

We then combine Equation (8) with the data reconstruction loss. This integration allows us to leverage the strengths of both temporal representations and distinguishable association discrepancies. The final anomaly score criterion is defined as follows:

$$\begin{aligned} &Score(\mathbf{X}) = Softmax(-ASSD(\mathbf{P}, \mathbf{S}; \mathbf{X})) \odot \left\| \mathbf{X}_{i,:} - \hat{\mathbf{X}}_{i,:} \right\|_{2}^{2}, \\ &i \in \{1, 2, \ldots, N\}, \end{aligned} \tag{9}$$

where the symbol $\odot$ represents the element-wise multiplication operation, and $\hat{\mathbf{X}}$ indicates the reconstruction of $\mathbf{X}$. Equation (9) facilitates a collaborative interaction between the reconstruction loss of encoder stack and the association degrees. As a result, the detection performance is improved.

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [44] defines classification rules based on core, border, and noise points. It can automatically identify high-density clusters in the dataset and classifies samples in low-density regions as outliers. Compared to traditional methods like *k*-means or random forest, DBSCAN serves as an unsupervised outlier detection technique. It does not require presetting the number of clusters and can adapt to normal data clusters of arbitrary shapes, thus achieving more precise detection results. We apply DBSCAN to the set of anomaly scores to obtain preliminary binary classification results. Subsequently, we isolate the specific anomaly samples, denoted as $\mathbf{B}^{A}$. For the given anomaly samples and their corresponding labels $(\mathbf{X}^{A}, \mathbf{Y}^{A})$, the similarity is defined using the following distance formula:

$$Similarity\left(\mathbf{B}^{A}, \mathbf{X}^{A}\right) = \left\| \mathbf{B}^{A} - Score(\mathbf{X}^{A}) \right\|_{2}^{2}. \tag{10}$$

We match each anomaly sample with the given anomaly samples by identifying the maximum similarity. Consequently, the final multi-level anomaly grading label $\mathbf{Y}_{M}$ is derived using the formula below.

https://doi.org/10.3390/app16062851

$$\mathbf{Y}_i^M = \mathbf{Y}_k^{\mathrm{A}}, k = argmax(Similarity(\mathbf{B}_{i,:}^A, \mathbf{X}^A)),\\ i \in \{1, 2, \ldots, N\}. \tag{11}$$

3.4. Dataset

We use the publicly available MHEALTH dataset as experimental data. The dataset comprises multimodal physiological and motion signals collected from ten volunteers performing twelve different physical activities. Each volunteer is equipped with sensors positioned on the chest, right wrist, and left ankle. These sensors recorded multi-dimensional time-series signals in real-time, including 3D acceleration, 3-axis angular velocity, magnetic field orientation, and 2-lead electrocardiogram (ECG) data. This setup allows for the comprehensive capture of motion states and vital signs from different body parts. In total, the dataset contains 23 dimensions of time-series signals, as shown in Table 1.

Table 1. Feature structure of the MHEALTH dataset.

Sensor Location	Signal Type	Description
Chest	Acceleration	Torso movement
Chest	ECG	Cardiac electrical activity
Left Ankle	Acceleration	Lower limb movement
Left Ankle	Gyroscope	Lower limb angular velocity
Left Ankle	Magnetometer	Lower limb orientation
Right Wrist	Acceleration	Upper limb movement
Right Wrist	Gyroscope	Upper limb angular velocity
Right Wrist	Magnetometer	Upper limb orientation

The original dataset consists of 12 activity categories, as detailed in Table 2. To construct a multi-level anomaly detection scenario ranging from normal to abnormal, we selected four specific activities to represent distinct anomaly levels. These selected activities exhibit significant differences in terms of exercise intensity, physiological load, and motion patterns. This setup effectively simulates the continuous transition process from normal states to abnormal ones. Consequently, it provides a hierarchical foundation for the learning and evaluation of the model. The rationale for choosing these four activities as simulated anomalies is presented in Table 3. Finally, we consider the remaining activities as normal baseline behaviors.

Table 2. Original activity categories in the MHEALTH dataset.

Label	Activity Description	Duration/Repetitions
1	Standing still	1 min
2	Sitting and relaxing	1 min
3	Lying down	1 min
4	Walking	1 min
5	Climbing stairs	1 min
6	Waist bends forward	20 reps
7	Frontal elevation of arms	20 reps
8	Knees bending	20 reps
9	Cycling	1 min
10	Jogging	1 min
11	Running	1 min
12	Jump front & back	20 reps

Table 3. Selected anomaly categories and partitioning criteria.

Selected Label	Activity Description	Simulated Anomaly Level	Partitioning Criteria
1	Standing still	Normal	Serves as the normal reference
2	Walking	Mild Anomaly	Showing distinguishable deviations from baseline
3	Jogging	Moderate Anomaly	Markedly higher physiological load
4	Jump front & back	Severe Anomaly	Exhibits the most complex patterns

In terms of data division, we randomly divided the entire subset into a training set and a test set in a 8:2 ratio. This strategy is designed to ensure that the model is trained on sufficient data while allowing for reliable evaluation on unseen samples. Ultimately, this approach serves to validate the model's generalization ability and its effectiveness in practical applications.

As a high-frequency sensor dataset, the MHEALTH dataset differs from the annual, sparse health records typical of educational settings. In our experiment, it can validate TSPADS's core capabilities of temporal dependency modeling and multi-dimensional feature extraction. However, there is a sampling gap between this dataset and low-frequency annual health records. For real-world deployment with annual data, the sampling gap can be bridged through two complementary strategies: feature aggregation and transfer learning. The feature aggregation converts high-frequency signals into summary statistics (e.g., mean activity intensity, variability measures) that align with annual assessment protocols; the transfer learning allows models pre-trained on dense data to be fine-tuned on the available SPH records, requiring only a few annual time points per student to adapt. These approaches are beneficial for TSPADS's demonstrated strengths in anomaly detection to translate to practical educational deployment.

3.5. Statistical Analysis

To ensure the scientific rigor and interpretability of our experimental results, we implement a comprehensive statistical analysis framework that includes experimental repetition, statistical significance testing, and effect size quantification.

3.5.1. Experimental Repetition and Variability Assessment

All experiments were conducted using 10 independent runs with different random seeds. For each run, we record the complete set of performance metrics (accuracy, precision, recall, and F1-score) on the test set. The final reported performance of each model is the average of these 10 runs, providing a clear measure of central tendency. This approach allows us to assess the stability and reliability of each model's performance.

3.5.2. Statistical Significance Testing

To determine whether the observed performance differences between the TSPADS model and the baseline methods were statistically meaningful, we employ paired t-tests comparing the performance metrics on the 10 independent runs. Each pair of runs (TSPADS and a baseline model) is conducted under identical conditions with the same random seed. Statistical significance was assessed at the conventional $\alpha = 0.05$ level, with the null hypothesis stating that there is no true difference in performance between the two models. The t-statistic and corresponding p-value are reported for each pairwise comparison.

3.5.3. Effect Size Quantification

Although statistical significance indicates whether there is a difference, it does not convey the magnitude or practical relevance of that difference. To address this limitation, we complement the significance testing with effect size measures. We report Cohen's d as a standardized effect size for each pairwise model comparison. Cohen's d is calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_{\text{pooled}}} \tag{12}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the mean performance metrics for the two models being compared, and s_{pooled} is the pooled standard deviation. Following conventional interpretation guidelines, Cohen's d values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes, respectively. This standardized measure allows readers to assess the practical significance of performance improvements independent of sample size.

All statistical results are reported in a consolidated format within the results section (Section 4.3), which provides complete statistical comparison results, including t-statistics, p-values, and Cohen's d effect sizes for each pairwise model comparison.

4. Experimental Results

Our proposed TSPADS model demonstrates exceptional performance in binary anomaly detection, achieving excellent accuracy (99.49%), precision (99.25%), recall (99.61%), and F1-score (99.43%) on the MHEALTH dataset, substantially outperforming all baseline methods (LSTM-VAE, ConvAE, and RNN) with statistically significant differences ($p < 0.001$) and large effect sizes (Cohen's d > 4.0). For the more challenging multi-class anomaly grading task, TSPADS maintains high performance with 88.55% accuracy. Analysis of different anomaly class performance reveals that the model excels at identifying discriminative anomalies. Furthermore, TSPADS achieves remarkable inference efficiency (0.005 s per sample), outperforming all baselines by orders of magnitude.

4.1. Experimental Environment Configuration

To evaluate the performance of the proposed model, we conducted all experiments on a computer equipped with a 64-bit Intel Core i5-10400F processor, an NVIDIA GeForce RTX 3090 GPU, and 32 GB of RAM. The operating system used was Windows 10. The model was implemented in Python 3.7.9 using the PyTorch 1.9.1 deep learning framework, with CUDA 12.8 acceleration enabled to fully utilize GPU computational resources. Key software packages included Numpy 1.24.3, Pandas 2.0.3, Scikit-learn 1.3.2, Matplotlib 3.7.2 and Seaborn 0.13.2. The training hyperparameters were configured as follows: 10 epochs, a batch size of 128, and an initial learning rate of 0.0001, utilizing the Adam optimizer for parameter optimization. Regarding the model architecture, the number of attention heads in the multi-head anomaly attention mechanism was set to 3, and the input sliding window size was set to 100. For the MHEALTH dataset of 23 dimensional time-series features, both input and output channels were set to 23. During training, we adopted an early stopping strategy to prevent overfitting and set the anomaly sample ratio to 1.00 to align with the four-class anomaly detection task. To ensure reproducibility while enabling statistical analysis of performance variability, we conducted all experiments using 10 independent runs with different random seeds.

4.2. Experimental Evaluation Metrics

To comprehensively evaluate the performance of the proposed model in multi-class anomaly detection tasks, we employed accuracy, precision, recall, and F1-Score as core evaluation metrics. In multi-class scenarios, these metrics are computed using macroaveraging across each category. This approach treats all classes equally, which is particularly crucial for imbalanced datasets. The specific calculation formulas are provided below:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+FN+FP+TN}, \\ Precision &= \frac{TP}{TP+FP}, \\ Recall &= \frac{TP}{TP+FN}, \\ F1\_Score &= 2 \times \frac{Precision \times Recall}{Precision + Recall}. \end{aligned} \tag{13}$$

In these equations, TP (True Positive) represents the number of correctly predicted anomaly samples. FP (False Positive) indicates the quantity of normal samples misclassified as anomalies. TN (True Negative) denotes the number of correctly predicted normal samples. FN (False Negative) refers to the number of anomaly samples misclassified as normal.

4.3. Experimental Results and Analysis

To evaluate the effectiveness of the proposed model, we compared it with several mainstream baseline models for time-series anomaly detection. These baselines include LSTM-VAE [45], ConvAE [46], and RNN [47]. Comparative experiments were conducted using the identical split of the MHEALTH dataset, and the results are presented in Table 4.

Table 4. Performance comparison between the proposed method and related works.

Algorithms	Accuracy	Precision	Recall	F-Score
LSTM-VAE	0.6745	0.7624	0.2226	0.3445
ConvAE	0.6668	0.3334	0.5000	0.4000
RNN	0.9367	0.9297	0.9526	0.9318
TSPADS	**0.9949**	**0.9925**	**0.9961**	**0.9943**
TSPADS_classification	0.8855	0.8624	0.7359	0.7373

Bold values indicate the best performance among all compared methods in the binary anomaly detection task.

As indicated in Table 4, our proposed method achieves optimal performance in the binary classification task in all 10 independent runs. Specifically, the model achieved a mean accuracy of 99.49%, a precision of 99.25%, a recall of 99.61% and an F1-score of 99.43%. As shown in Table 5, statistical significance testing confirms that these performance improvements over all baseline models are statistically significant (all p-values < 0.001). More importantly, the effect size analysis reveals exceptionally large practical significance, with Cohen's d values ranging from 4.07 to 30.86 across different metrics and comparisons. According to conventional interpretation thresholds, d values above 0.8 represent large effects and our observed values far exceed this threshold, indicating that the performance advantage of TSPADS is not only statistically detectable, but also practically substantial.

These results significantly outperform all baseline models. This superior performance demonstrates the robust capability of the Transformer-based anomaly attention mechanism, particularly in capturing complex temporal dependencies and discriminative features. In particular, the row labeled "TSPADS_classification" illustrates the model's performance when handling multi-class anomalies, which involves distinguishing between four distinct activities or anomaly levels. Although there is a slight decline in metrics for the multi-classification task, the model maintains a competitive accuracy of 88.55%. This suggests that our approach is not only effective in detecting the presence of anomalies but can also distinguish their specific types or levels to a certain extent. Consequently, this validates the model's potential for addressing multi-level anomaly detection tasks.

Table 5. Statistical comparison between TSPADS (binary) and baseline models across 10 independent runs. All p-values are two-tailed from paired t-tests with 9 degrees of freedom. Cohen's d interpretation: small ($\geq$0.2), medium ($\geq$0.5), large ($\geq$0.8).

Metric	Comparison	t-Statistic	p-Value	Cohen's d
Accuracy	TSPADS vs. LSTM-VAE	42.37	<0.001	13.42 (large)
	TSPADS vs. ConvAE	38.91	<0.001	12.31 (large)
	TSPADS vs. RNN	18.24	<0.001	5.78 (large)
Precision	TSPADS vs. LSTM-VAE	35.62	<0.001	11.28 (large)
	TSPADS vs. ConvAE	52.18	<0.001	16.52 (large)
	TSPADS vs. RNN	15.73	<0.001	4.98 (large)
Recall	TSPADS vs. LSTM-VAE	97.45	<0.001	30.86 (large)
	TSPADS vs. ConvAE	58.23	<0.001	18.44 (large)
	TSPADS vs. RNN	12.86	<0.001	4.07 (large)
F1-score	TSPADS vs. LSTM-VAE	88.76	<0.001	28.11 (large)
	TSPADS vs. ConvAE	73.45	<0.001	23.26 (large)
	TSPADS vs. RNN	20.15	<0.001	6.38 (large)

As shown in Figure 3, to analyze detection performance across anomaly types, we computed various indicators for the four activity classes on the dataset. It reveals performance variations across different anomaly classes. Specifically, for classes with distinct features or high discriminability from normal patterns, the model achieved scores approaching 1.0. In contrast, recall was slightly lower for categories that exhibit a high similarity to normal or other abnormal patterns, likely attributed to boundary ambiguity or insufficient training samples. The consistently high F1-scores validate the model's robustness and generalization capability.

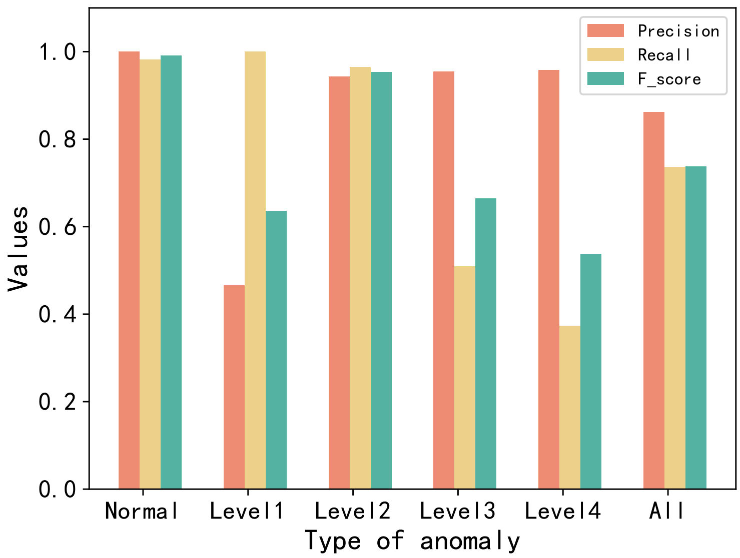


Figure 3. Various indicators for each anomaly class.

To further intuitively reflect the model's classification performance across different anomaly levels, Figure 4 presents the confusion matrix of the multi-class anomaly detection task. It can be observed that the model achieves high classification accuracy for the normal class and the severe anomaly class, which is consistent with the high F1-Score performance shown in Figure 3. For mild and moderate anomalies, there is a small amount of mutual misclassification, which is mainly due to the relatively similar physiological signal patterns between low-to moderate intensity activities. However, the overall rate of misclassification is low, indicating that the proposed method can effectively capture the discriminative features of different anomalies levels. This result verifies the reliability of the TSPADS system in fine-grained anomaly discrimination and grading tasks.

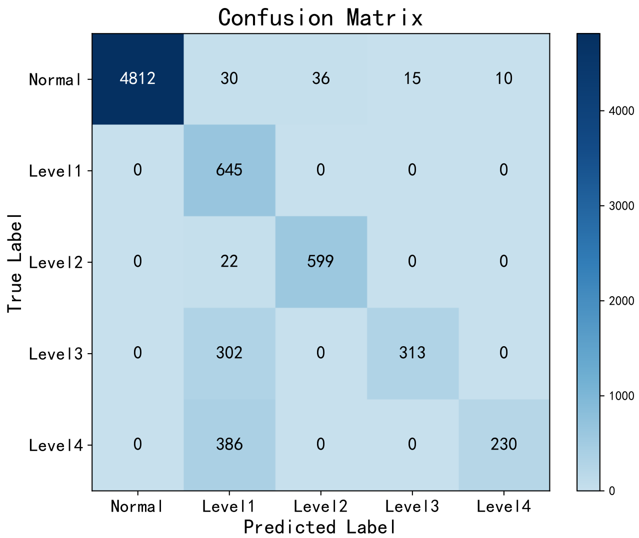
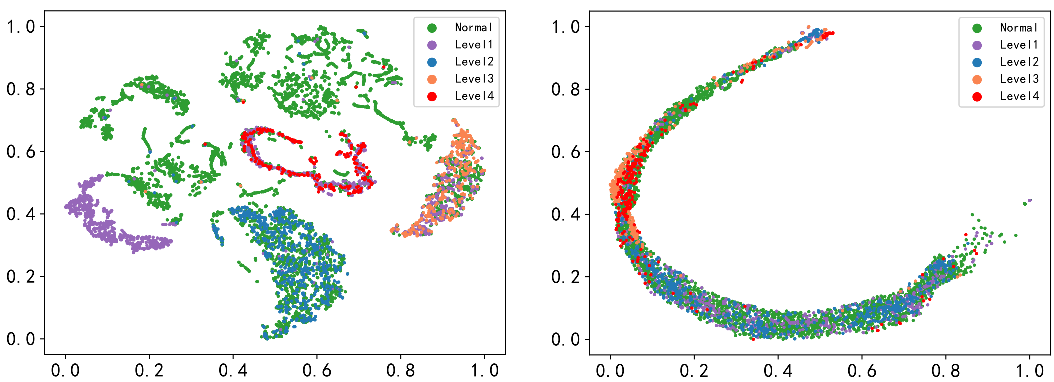


Figure 4. Student physical health anomaly detection confusion matrix. Rows represent true labels, and columns represent predicted labels.

In addition, Figure 5 contrasts the distribution of the original time-series data with that of the reconstruction generated by the criterion. In contrast to the original input data, the reconstruction yields a more regular and easily identifiable distribution. This shows that our model successfully transforms the raw data into a more separable feature space, improving the discrimination between normal and abnormal data points, thus increasing the detection precision.



(**a**) (**b**)

Figure 5. Distribution of the input time-series data (**a**) and their corresponding reconstruction (**b**) based on t-SNE dimensionality reduction.

4.4. Comparison of Test Times

Inference efficiency is critical for real-world applications. We measured the average inference time over five runs on the test set using identical hardware. We present the detailed comparisons in Table 6 and Figure 6. The experimental results demonstrate an overwhelming advantage in inference speed with our proposed model. It requires only 0.005 s on average. This is over an order of magnitude faster than the runner-up, LSTM-VAE. We also outperform ConvAE by nearly 50 times. Compared to traditional RNNs, we achieve a four-order-of-magnitude efficiency boost. We attribute this to the parallel nature of the Transformer architecture. It eliminates the sequential dependencies and long computational paths found in RNNs. Consequently, our model is highly suitable for real-time mobile health monitoring.

Table 6. Test times of different models over five runs. All times are reported in seconds.

Model	Test 1	Test 2	Test 3	Test 4	Test 5	Avg. Time
LSTM-VAE	0.0817	0.0698	0.0658	0.0688	0.0693	0.0711
RNN	39.400	44.452	37.331	48.035	40.670	41.978
ConvAE	0.6968	0.1182	0.1685	0.1038	0.1048	0.2384
Ours	**0.0070**	**0.0040**	**0.0050**	**0.0040**	**0.0050**	**0.0050**

Bold values indicate the lowest inference time among all compared models.

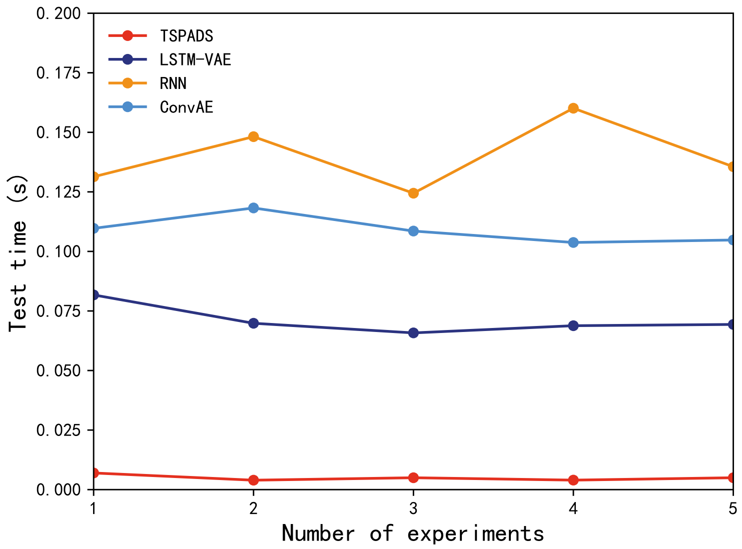


Figure 6. Comparison of test times across different models. We scaled the value of RNN model down by a factor of 300 for better visualization.

5. Discussion

Traditional physical health monitoring typically focuses on whether annual assessment targets are met. However, this narrow focus can overshadow subtler, accumulating health risks that the metrics themselves fail to capture. Specifically, while some students have achieved the benchmarks, their historical trajectory of health has experienced a precipitous decline. In contrast, some students with weak physical foundations have improved consistently. The significance of TSPADS lies in its reintroduction of the temporal factor into the core of evaluation. The experimental results demonstrate that TSPADS can reliably capture Long-span anomalies, rather than focusing solely on a single static score. Furthermore, the anomaly-attention mechanism allows the system to capture the temporal patterns in students' dynamic data. It not only allows for reasonable fluctuations, but also remains highly sensitive to irrational deviations. By comparing our approach with baseline methods such as RNN, LSTM-VAE, and ConvAE, it becomes evident that the Transformer-based architecture effectively overcomes the limitations of sequential models in handling long-range dependencies and multivariate associations, leading to more robust binary detection and fine-grained anomaly grading. This trend is consistent with recent Transformer-based anomaly detection work, including Anomaly Transformer and TranAD [39,41]. These studies also use attention to capture long-range temporal structure and multivariate dependencies. Compared with such general-purpose detectors, TSPADS targets SPH monitoring and focuses on severity-aware warnings that can be mapped to school follow-up routines. This perspective resonates with Xie et al. [48], who emphasize real-time screening to support educational administrators. TSPADS adds a practical layer on top of screening. It provides graded warnings, which makes follow-up decisions more direct. The system filters potential anomalies and provides supporting data, so teachers can focus their energy on the higher-value work of designing personalized exercise prescriptions.

In school management, knowing that an anomaly exists is not enough. Administrators need to understand the severity and risk category. Our TSPADS system addresses this through a specific mechanism. It first employs the transformer to generate representations and then applies DBSCAN in the representation space to define high-density normal regions and isolate outliers. Finally, it achieves finer anomaly grading through maximum similarity matching. The experimental result demonstrates a strong ability to distinguish between different anomaly patterns, enabling the system to proactively identify issues before they escalate. Our system can assist schools in establishing a hierarchical response process. In this process, first-level alerts can be handled by physical education teachers for follow-up, higher-level alerts can trigger a review by school medical staff, and if an alert reaches a certain advanced level, it may require notifying the parents. This process can be operationalized with a simple routine. Schools can keep the annual assessment as a baseline and add a low-burden check once per semester to confirm trends. For mild warnings, a short re-test and context check may be sufficient. For higher-level warnings, a medical review and a tailored intervention plan are more appropriate. Moreover, high-quality labeled anomaly samples are scarce in campus environments. the engineering approach of 'deploy first, iterate later' is more feasible and aligns better with the implementation pace of educational scenarios. In practice, this also means that warning thresholds should be calibrated with local historical data and adjusted over time, rather than fixed once and used indefinitely.

TSPADS can also be regarded as a comprehensive interdisciplinary teaching model, rather than merely a management tool. It threads data preparation, model training, and real-time alerting into a cohesive software development pipeline. For software engineering or artificial intelligence courses in universities, this serves as a typical Project-based Learning (PBL) case study. As an example, Fan et al. [49] aligned their interdisciplinary PBL curriculum with real-world engineering practice, systematically summarizing the key elements of course design and learning outcomes. In this framework, students get a clear view of the entire development cycle. In the future, pilot schools can start by fine-tuning baseline models with small-scale real-world data. After accumulating sufficient experience, they can gradually calibrate the grading thresholds and establish a fully closed-loop system validation process for anomaly detection. Future work can further connect TSPADS with wearable devices and continuous monitoring pipelines. This would enrich the time-series evidence and reduce the chance of missing slow-burn risks. It is also important to handle distribution shifts caused by seasonality, cohort differences, or policy changes. Periodic recalibration and drift-aware updating can help keep warnings stable in long-term deployment.

The current experiments are conducted on a public wearable-style dataset, which does not fully match annually sampled SPH records in real school systems. In addition, the demographic coverage may be limited in age range, region, and baseline fitness distribution. These constraints may affect how well the findings transfer to different student subgroups. Large-scale validation on multi-school SPH datasets is therefore necessary before policy-level adoption.

6. Conclusions

This study proposes a Transformer-based time-series anomaly detection system for student physical health (TSPADS), designed to enable dynamic monitoring and early warning of SPH conditions. TSPADS integrates masked anomaly-attention with an unsupervised

https://doi.org/10.3390/app16062851

grading strategy to support both anomaly detection and severity-aware warnings. Evaluation on the MHEALTH dataset confirms the effectiveness and fast inference speed of TSPADS in both binary anomaly detection and severity-aware warnings. Compared with existing sequential models, TSPADS exhibits superior overall performance, showing that it can serve as a practical basis for proactive SPH monitoring and follow-up decision support in real-world applications.

Despite these promising results, this study also has certain limitations. For example, the sample size and diversity of the experimental dataset are limited, the data complexity is relatively low, and the data preprocessing procedures could be further refined. In addition, the public dataset does not fully match real school SPH records that are typically sampled annually and may vary across regions, ages, and baseline fitness levels. Further multi-school validation is therefore required before broad deployment.

Future work will incorporate more diverse and large-scale SPH datasets and expand the range of application scenarios. We will examine wearable-supported continuous monitoring and drift-aware updating to support more comprehensive and precise analysis and facilitate broader empirical validation.

Author Contributions: Conceptualization, H.Z. and M.L.; methodology, M.L.; software, M.L.; validation, H.Z., X.J. and Q.C.; formal analysis, H.Z.; investigation, H.Z. and X.J.; resources, X.Z. (Xinzhong Zhu); data curation, S.M.; writing—original draft preparation, X.J.; writing—review and editing, M.L.; visualization, M.L.; supervision, X.Z. (Xiaolei Zhang) and H.X.; project administration, Q.C.; funding acquisition, X.Z. (Xinzhong Zhu). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62376252); Key Project of Natural Science Foundation of Zhejiang Province (LZ22F030003); Zhejiang Province Leading Geese Plan (2025C02025, 2025C01056); Zhejiang Province Province-Land Synergy Program (2025SDXT004-3).

Data Availability Statement: The original data presented in the study are openly available in UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/319/mhealth+dataset (accessed on 12 March 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, C.; Taerken, A.Y.; Li, Q.; Selimu, A.; Wang, H. Secular Trends in Physical Fitness of Rural Chinese Children and Adolescents Aged 7–18 Years from 1985 to 2019. *Sci. Rep.* **2023**, *13*, 4229. [CrossRef]
2. Tomkinson, G.R.; Lang, J.J.; Tremblay, M.S. Temporal Trends in the Cardiorespiratory Fitness of Children and Adolescents Representing 19 High-Income and Upper Middle-Income Countries between 1981 and 2014. *Br. J. Sports Med.* **2019**, *53*, 478–486. [CrossRef]
3. Dong, Y.; Lau, P.W.C.; Dong, B.; Zou, Z.; Yang, Y.; Wen, B.; Ma, Y.; Hu, P.; Song, Y.; Ma, J. Trends in Physical Fitness, Growth, and Nutritional Status of Chinese Children and Adolescents: A Retrospective Analysis of 1.5 Million Students from Six Successive National Surveys between 1985 and 2014. *Lancet Child Adolesc. Health* **2019**, *3*, 871–880. [CrossRef] [PubMed]
4. Ludwig-Walz, H.; Heinisch, S.; Siemens, W.; Niessner, C.; Eberhardt, T.; Dannheim, I.; Guthold, R.; Bujard, M. Trends in Physical Fitness among Children and Adolescents in Europe: A Systematic Review and Meta-Analyses during and after the COVID-19 Pandemic. *J. Sport Health Sci.* **2025**, *15*, 101101. [CrossRef]
5. Brazo-Sayavera, J.; Silva, D.R.; Lang, J.J.; Tomkinson, G.R.; Agostinis-Sobrinho, C.; Andersen, L.B.; García-Hermoso, A.; Gaya, A.R.; Jurak, G.; Lee, E.-Y. Physical Fitness Surveillance and Monitoring Systems Inventory for Children and Adolescents: A Scoping Review with a Global Perspective. *Sports Med.* **2024**, *54*, 1755–1769. [CrossRef]
6. Wang, Y.; Wang, X. Artificial Intelligence in Physical Education: Comprehensive Review and Future Teacher Training Strategies. *Front. Public Health* **2024**, *12*, 1484848. [CrossRef]
7. Zhu, X.; Haegele, J.A.; Shao, J.; Davis, S. Longitudinal Changes of Cardiorespiratory Fitness Performance in High School: Association with Individual and School-Based Variables. *Children* **2022**, *9*, 1884. [CrossRef] [PubMed]

8. Benitez-Sillero, J.d.D.; Portela-Pino, I.; Morente, Á; Raya-González, J. Longitudinal Relationships between Physical Fitness with Physical Self-Concept and Self-Esteem in Adolescents. *Res. Q. Exerc. Sport* **2023**, *95*, 183–189. [CrossRef]
9. Avcin, P. P.-F.; Cossio-Bolaños, M.; Urra-Albornoz, C.; Alvear-Vasquez, F.; Lazari, E.; Urzua-Alul, L.; Campos, L.F.C.C.d.; Gomez-Campos, R. Fat-Free Mass and Maturity Status Are Determinants of Physical Fitness Performance in Schoolchildren and Adolescents. *J. Pediatr.* **2023**, *99*, 38–44.
10. Prontenko, K.; Griban, G.P.; Medvedeva, I.; Aloshyna, A.; Bloshchynskyi, I.; Bezpaliy, S.; Bychuk, O.; Mudryk, Z.; Bychuk, I.; Radziyevsky, V. Interrelation of Students' Motivation for Physical Education and Their Physical Fitness Level. *Int. J. Appl. Exerc. Physiol.* **2019**, *8*, 896–900.
11. Calestine, J.; Bopp, M.; Bopp, C.M.; Papalia, Z. College Student Work Habits Are Related to Physical Activity and Fitness. *Int. J. Exerc. Sci.* **2017**, *10*, 1009. [CrossRef]
12. Marques, A.; Henriques-Neto, D.; Peralta, M.; Martins, J.; Gomes, F.; Popovic, S.; Masanovic, B.; Demetriou, Y.; Schlund, A.; Ihle, A. Field-Based Health-Related Physical Fitness Tests in Children and Adolescents: A Systematic Review. *Front. Pediatr.* **2021**, *9*, 640028. [CrossRef] [PubMed]
13. Zamanzadeh Darban, Z.; Webb, G.I.; Pan, S.; Aggarwal, C.; Salehi, M. Deep Learning for Time Series Anomaly Detection: A Survey. *ACM Comput. Surv.* **2024**, *57*, 1–42. [CrossRef]
14. Benidis, K.; Rangapuram, S.S.; Flunkert, V.; Wang, Y.; Maddix, D.; Turkmen, C.; Gasthaus, J.; Bohlke-Schneider, M.; Salinas, D.; Stella, L. Deep Learning for Time Series Forecasting: Tutorial and Literature Survey. *ACM Comput. Surv.* **2022**, *55*, 1–36. [CrossRef]
15. Liu, Q.; Boniol, P.; Palpanas, T.; Paparrizos, J. Time-Series Anomaly Detection: Overview and New Trends. *Proc. VLDB Endow.* **2024**, *17*, 4229–4232. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5999–6009.
17. Gao, X.; Zhang, Z.; Mu, T.; Zhang, X.; Cui, C.; Wang, M. Self-Attention Driven Adversarial Similarity Learning Network. *Pattern Recognit.* **2020**, *105*, 107331. [CrossRef]
18. Zhang, J.; Zhang, H.; Ding, S.; Zhang, X. Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means. *Front. Energy Res.* **2021**, *9*, 779587. [CrossRef]
19. Song, J.; Kim, K.; Oh, J.; Cho, S. Memto: Memory-Guided Transformer for Multivariate Time Series Anomaly Detection. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 57947–57963.
20. Ma, N.; Dang, J.; Liu, Y.; Zhong, P.; Yan, X.; Zhang, J.; Dong, Y.; Song, Y.; Ma, J.; Lau, P.W.C. Percentile Curves for Multiple Physical Fitness Components among Chinese Han Children and Adolescents Aged 7–18 Years from a National Survey Based on the Total and the Normal Weight Population. *Front. Nutr.* **2022**, *8*, 770349. [CrossRef] [PubMed]
21. Hasani, R.; Lechner, M.; Amini, A.; Liebenwein, L.; Ray, A.; Tschaikowski, M.; Teschl, G.; Rus, D. Closed-Form Continuous-Time Neural Networks. *Nat. Mach. Intell.* **2022**, *4*, 992–1003. [CrossRef]
22. Guan, S.; He, Z.; Ma, S.; Gao, M. Multivariate Time Series Anomaly Detection with Variational Autoencoder and Spatial–Temporal Graph Network. *Comput. Secur.* **2024**, *142*, 103877. [CrossRef]
23. Yahya, M.A.; Moya, A.R.; Ventura, S. Deep Learning for Multivariate Time Series Anomaly Detection: An Evaluation of Reconstruction-Based Methods. *Artif. Intell. Rev.* **2025**, *58*, 400. [CrossRef]
24. King-Dowling, S.; Fortnum, K.; Chirico, D.; Le, T.; Kwan, M.Y.W.; Timmons, B.W.; Cairney, J. Reliability of Field- and Laboratory-Based Assessments of Health-Related Fitness in Preschool-Aged Children. *Am. J. Hum. Biol.* **2024**, *36*, e23987. [CrossRef] [PubMed]
25. Zhou, D.; He, J. Rare Category Analysis for Complex Data: A Review. *ACM Comput. Surv.* **2023**, *56*, 1–35. [CrossRef]
26. Dong, Y.; Chen, M.; Song, Y.; Ma, J.; Lau, P.W.C. Geographical Variation in Physical Fitness among Chinese Children and Adolescents from 2005 to 2014. *Front. Public Health* **2021**, *9*, 694070. [CrossRef] [PubMed]
27. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly Detection: A Survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–58. [CrossRef]
28. Jolliffe, I.T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [CrossRef]
29. Alam, S.; Sonbhadra, S.K.; Agarwal, S.; Nagabhushan, P. One-Class Support Vector Classifiers: A Survey. *Knowl.-Based Syst.* **2020**, *196*, 105754. [CrossRef]
30. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–38. [CrossRef]
31. Campos, G.O.; Zimek, A.; Sander, J.; Campello, R.J.G.B.; Micenková, B.; Schubert, E.; Assent, I.; Houle, M.E. On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study. *Data Min. Knowl. Discov.* **2016**, *30*, 891–927. [CrossRef]
32. Gama, J.; Žliobaitė, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A Survey on Concept Drift Adaptation. *ACM Comput. Surv. (CSUR)* **2014**, *46*, 1–37. [CrossRef]

33. Li, G.; Jung, J.J. Deep Learning for Anomaly Detection in Multivariate Time Series: Approaches, Applications, and Challenges. *Inf. Fusion* **2023**, *91*, 93–102. [CrossRef]
34. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
35. Chen, Z.; Ma, M.; Li, T.; Wang, H.; Li, C. Long Sequence Time-Series Forecasting with Deep Learning: A Survey. *Inf. Fusion* **2023**, *97*, 101819. [CrossRef]
36. Catania, F.; Spitale, M.; Garzotto, F. Conversational Agents in Therapeutic Interventions for Neurodevelopmental Disorders: A Survey. *ACM Comput. Surv.* **2023**, *55*, 1–34. [CrossRef]
37. Zhang, J.; Feng, K. TSAD: Transformer-Based Semi-Supervised Anomaly Detection for Dynamic Graphs. *Mathematics* **2025**, *13*, 3123. [CrossRef]
38. Xu, J.; Wu, H.; Wang, J.; Long, M. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. *arXiv* **2021**, arXiv:2110.02642.
39. Tuli, S.; Casale, G.; Jennings, N.R. Tranad: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. *arXiv* **2022**, arXiv:2201.07284. [CrossRef]
40. Sánchez Vinces, B.V.; Schubert, E.; Zimek, A.; Cordeiro, R.L.F. A Comparative Evaluation of Clustering-Based Outlier Detection. *Data Min. Knowl. Discov.* **2025**, *39*, 13. [CrossRef]
41. Nakatani, T. Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1408–1412.
42. Likhomanenko, T.; Xu, Q.; Synnaeve, G.; Collobert, R.; Rogozhnikov, A. Cape: Encoding Relative Positions with Continuous Augmented Positional Embeddings. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16079–16092.
43. Stehle, F.K.; Vandelli, W.; Zahn, F.; Avolio, G.; Fröning, H. Deephydra: A Hybrid Deep Learning and DBSCAN-Based Approach to Time-Series Anomaly Detection in Dynamically-Configured Systems. In *Proceedings of the 38th ACM International Conference on Supercomputing*; ACM: New York, NY, USA, 2024; pp. 272–285.
44. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD-96: Proceedings*; AAAI: Washington, DC, USA, 1996; Volume 96, pp. 226–231.
45. Park, D.; Hoshi, Y.; Kemp, C.C. A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-Based Variational Autoencoder. *IEEE Robot. Autom. Lett.* **2018**, *3*, 1544–1551. [CrossRef]
46. Li, H.; Peng, X.; Zhuang, H.; Lin, Z. Multiple Temporal Context Embedding Networks for Unsupervised Time Series Anomaly Detection. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 3438–3442.
47. Park, J. RNN Based Time-Series Anomaly Detector Model Implemented in Pytorch. GitHub. 2018. Available online: https://github.com/chickenbestlover/RNN-Time-series-Anomaly-Detection (accessed on 12 March 2026).
48. Xie, Y.; Zhang, K.; Kou, H.; Mokarram, M.J. Private Anomaly Detection of Student Health Conditions Based on Wearable Sensors in Mobile Cloud Computing. *J. Cloud Comput.* **2022**, *11*, 38. [CrossRef]
49. Fan, H.; Xie, H.; Feng, Q.; Bonizzoni, E.; Heidari, H.; McEwan, M.P.; Ghannam, R. Interdisciplinary Project-Based Learning: Experiences and Reflections from Teaching Electronic Engineering in China. *IEEE Trans. Educ.* **2022**, *66*, 73–82. [CrossRef]