



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Construction and empirical study of a multi-source data fusion model for adolescent health literacy assessment

Man Guo

DOI: [10.1504/IJICT.2026.10076764](https://doi.org/10.1504/IJICT.2026.10076764)

Article History:

Received:	06 November 2025
Last revised:	03 December 2025
Accepted:	09 December 2025
Published online:	18 March 2026

Construction and empirical study of a multi-source data fusion model for adolescent health literacy assessment

Man Guo

School of Physical Education,
Hanshan Normal University,
Chaozhou, 521041, China
Email: happy20246688@163.com

Abstract: To overcome limitations of traditional adolescent health literacy assessment – single data source, insufficient accuracy and interpretability – a multi-source data fusion model is developed. This study has four key phases: first, extract and combine features from four data types (questionnaire, physiological sensor, behaviour record, environment); second, propose a deep learning model integrating a two-path feature encoder and random forest to enhance fitting ability; third, use SHAP and LIME for global and local interpretability analysis to assess key feature weight changes across groups; finally, verify model performance through multiple empirical experiments. Experiments show the integrated model achieves an F1-score of 0.873 and test-set AUC of 0.921, outperforming comparative models significantly while maintaining stable performance across age and gender heterogeneous groups. This paper verifies the validity of multiple data fusion strategies and provides an extensible algorithm framework with an interpretable mechanism.

Keywords: health literacy assessment; multi-source data fusion; deep learning; feature importance analysis; cohort analysis.

Reference to this paper should be made as follows: Guo, M. (2026) ‘Construction and empirical study of a multi-source data fusion model for adolescent health literacy assessment’, *Int. J. Information and Communication Technology*, Vol. 27, No. 23, pp.1–16.

Biographical notes: Man Guo obtained her Doctoral degree from José Rizal University and is currently a faculty member at Hanshan Normal University. Her research focuses on public sports management and adolescent health promotion.

1 Introduction

With the continuous expansion of digital health ecosystems, adolescent health literacy – defined as the ability to access, comprehend, and apply health information – has become a crucial determinant of public health development. Recent evidence indicates that inadequate health literacy among youth affects not only personal well-being but also long-term behavioural patterns, making accurate assessment increasingly important. Meanwhile, digital health policies worldwide emphasise the role of data-driven tools in supporting youth health monitoring and preventive intervention programs.

Traditional assessment methods, which rely primarily on questionnaires or self-reported scales, face well-documented limitations. As noted by Chou et al. (2023) adolescent self-reports are highly vulnerable to social desirability effects and peer-influenced self-presentation bias, reducing the reliability of results. This is particularly problematic for youth, whose health behaviours and cognitive maturity vary widely. Consequently, researchers have begun incorporating objective indicators – such as physiological signals and environmental data – to mitigate subjective distortions and enhance assessment validity.

Existing HQCA models, however, often depend on single-source information and fail to capture the multidimensional nature of adolescent health behaviours, cognition, and environmental context. Although Zhu (2025) introduced a multi-source fusion framework that integrates behavioural, physiological, and environmental streams, challenges remain. These include heterogeneous sampling frequencies, temporal alignment difficulty, and cross-platform deployment barriers, which collectively reduce the adaptability and precision of assessment systems in real-world school or community environments.

To address these limitations, this study develops an integrated evaluation model that fuses questionnaire data with behavioural, environmental, and physiological sources. By combining deep temporal modelling with ensemble learning and incorporating SHAP- and LIME-based interpretability, the proposed framework enhances accuracy, cross-platform robustness, and decision transparency. Furthermore, subgroup analyses based on age and gender provide evidence of the model’s applicability across diverse adolescent populations, contributing an interpretable and scalable foundation for personalised health literacy assessment.

2 Data collection and integration strategy

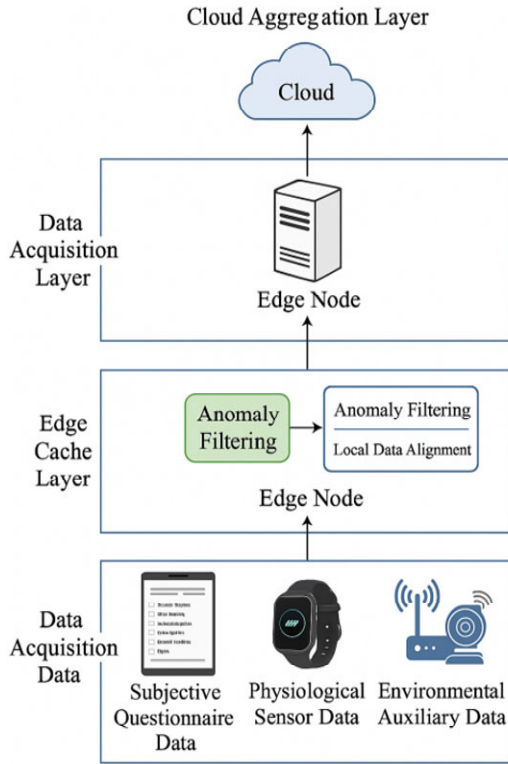
2.1 Data sources and collection methods

Based on the data collection stage, we have designed a multi-dimension data collection system, including subjective questionnaire, physiological sensor and environment assistant data. This method allows comprehensive perception of cognitive levels, behavioural habits, physiological states, and social support environments. Data of subjective questionnaire were collected with the help of an electronic survey platform. The questionnaire employed a 5-point Likert scale covering eight dimensions – including cognition, information acquisition, and self-evaluation of health behaviours – totalling 42 items. Use wearable devices to gather physiological data, such as heart rate, skin conductivity, sleep duration, and activity intensity. The apparatus has a sampling rate of 1 Hz, and the data is transferred to the mobile terminal using the Bluetooth low energy (BLE) protocol (Zhu and Deng, 2025; Chen and Li, 2024). The internet of things platform provides basic information about air quality (PM and CO₂ concentrations), classroom illumination and noise levels. The data is collected every 10 minutes, and the MQTT is applied to edge server in real time. In addition, the system’s data collection framework consists of three layers: front collection, edge caching, and cloud aggregation, as shown in Figure 1. To increase the real-time processing efficiency and bandwidth utilisation, a light-weight container based on KubeEdge is used in the edge nodes to carry out initial filtering of anomalous data and local data alignment. The overall data flow path of the collection system can be abstracted as equation (1):

$$D_f = F_{trans} \left(\bigcup_{i=1}^N S_i \circ T_i \right) \quad (1)$$

Here, D_f denotes the final merged dataset, S_i represents the i -th data source (questionnaire, physiological, or environmental), T_i indicates the corresponding sampling sequence function, F_{trans} is the transformation operator for data transmission and alignment operations, and N is the total number of data sources. Through this multi-source data acquisition architecture design, the system achieves synchronous acquisition and preliminary fusion of heterogeneous data, providing a high-quality, temporally consistent input foundation for subsequent evaluation model construction.

Figure 1 Data acquisition and upload architecture diagram (see online version for colours)



2.2 Heterogeneous data preprocessing

In order to solve the problem of structural inconsistency and quality variation in multi-source heterogeneous data before fusion, a five-step preprocessing process was developed: data structure analysis, missing value attribution, time alignment, numerical normalisation, and standardisation transformation. The data structure parsing stage unifies intermediate representation standards, mapping JSON-formatted questionnaire data, CSV-formatted physiological signals, and MQTT streaming environmental data into

a standardised tensor format: $\mathfrak{R}^{n \times m}$, where n denotes the time step and m represents the feature dimension. Missing value handling employs a K-nearest neighbour interpolation-based imputation function, as shown in equation (2).

$$\hat{x}_i = \frac{1}{K} \sum_{j \in N_K(i)} x_j \quad (2)$$

where $\hat{x}_i$ denotes the imputed value for the i -th sample, $N_K(i)$ represents its K nearest neighbour set, and x_j indicates the nearest neighbour values. Time alignment uses a sliding window resampling mechanism with a 10-second window and a 5-second stride for synchronisation of data between different sampling rates (Zhang et al., 2023). Numerical normalisation incorporates Z-score standardisation, mapping features of varying scales to the standard normal distribution range via the processing function:

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

In equation (3), x is the original feature, μ is the mean, and σ is the characteristic standard deviation. All of the processing steps are pre-executed at the edge nodes, and are cached locally in an interim database with a unified timestamp format. This ensures the consistency of the data, the time integrity, and the feasibility of modelling during the subsequent fusion process.

2.3 Multi-source fusion framework design

To effectively integrate subjective questionnaire data, physiological signals, and environmental monitoring inputs, we propose a multi-source fusion framework composed of edge-side preprocessing, timestamp alignment, and attention-based fusion encoding. The framework follows a distributed architecture illustrated in Figure 2, designed for edge-cloud collaboration.

At the edge layer, each data source undergoes standardisation and outlier filtering independently. Questionnaire data, physiological signals, and environmental streams – denoted as X_q , X_s , and X_e – are normalised and cleaned. MQTT protocol handles environmental data via asynchronous streaming (Goodfellow et al., 2023).

Following edge preprocessing, a timestamp alignment module aligns all inputs into a unified temporal scale using a mapping function $t_{align} \cdot T_i \rightarrow T'$ is the original sampling sequence and T' is the reference timeline. A sliding window mechanism (e.g., 10 s window with 5 s stride) ensures synchronisation across heterogeneous frequencies.

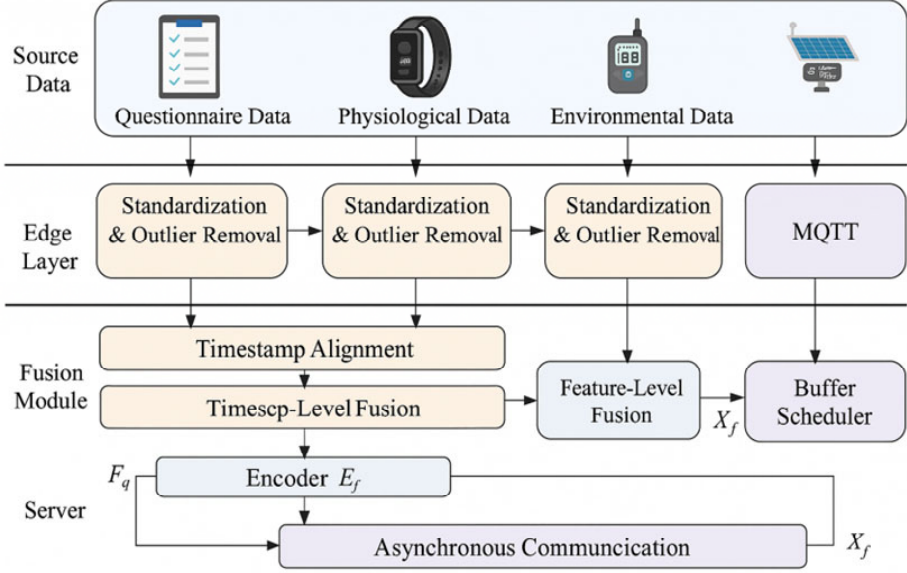
Once temporally aligned, a feature-level fusion module employs a parallel nested strategy to project X_q , X_s , and X_e into a shared latent space. These are concatenated and passed to the fusion encoder E_f , producing a joint representation $\mathbf{X}_f$. The process is mathematically described in equation (4):

$$\mathbf{X}_f = E_f \left(F_{concat} (X'_q, X'_s, X'_e) \right) \quad (4)$$

where F_{concat} represents the concatenation function, and E_f contains an attention mechanism that dynamically adjusts the fusion weights based on feature relevance to the prediction task. The feature stream F_v is then buffered through a Global Buffer Scheduler (WSP) to prevent congestion and ensure consistent data availability.

All fused features are transmitted via RESTful API to the server asynchronously. The system's lightweight deployment, combined with MQTT and REST, ensures efficient operation across cloud-edge platforms (Curran et al., 2023). This fusion pipeline supports scalable health literacy assessment across varied data sources and real-time environments.

Figure 2 Multi-source fusion framework architecture (see online version for colours)



2.4 Fusion data quality assurance and privacy security mechanism

In order to ensure the high quality, reliability and compliance of multi-source fusion data, a dual-layer mechanism of quality control and privacy is proposed. In order to ensure the data quality, the system uses three kinds of authentication strategies: Firstly, a hash function called $H(x)$ is used to verify the integrity of the data block and make sure that it will not be changed during transmission. Secondly, a sliding window verification scheme is proposed for detecting the time loss of samples (Vasli et al., 2024). Finally, it incorporates a residual-based feature anomaly detection formula:

$$r_i = \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2 \quad (5)$$

In equation (5), $\mathbf{x}_i$ represents the original feature vector, $\hat{\mathbf{x}}_i$ denotes the standard model estimate, and r_i indicates the residual for the i -th sample.

If $r_i > \theta$, the sample is marked as anomalous. As far as privacy is concerned, the system adopts end-to-end AES-256 encryption and two-way TLS channels to secure data transfer. A Role-Based Access Control (RBAC) model is implemented on the server side, which imposes granular restrictions on the scope of user operations and authorisation boundaries. To prevent the risk of re-identification, all raw identity data will undergo pseudonymisation during collection, and the unique ID will be mapped to irreversible

hashes and stored in a special de-tagged index database (Renwick et al., 2024). The Integrated Platform supports the regular invocation of the K-Anonymous Module to verify the identity of the published data. This increases the protection of the group's privacy and guarantees the security and reliability of the evaluation system.

3 Design of health literacy assessment model

3.1 Indicator system construction

For the purpose of comprehensive and objective evaluation of youth health literacy, we established a multi-dimension comprehensive assessment system, including cognition, behaviour, environment and information. A combination of self-assessment scales and behavioural simulation tests was employed to quantify cognitive scores (Sun et al., 2023). Secondly, the behavioural dimension assessed the performance of daily health activities, including diet, exercise, and tobacco/alcohol consumption. The evaluation tools designed for this purpose included frequency scales and lifestyle habits questionnaires. The environmental support dimension focused on the extent of support for health behaviours within families, schools, and communities. Indicators included parental modelling of health behaviours, school health education programs, and public health facilities availability. This section has been complemented by data from local authorities and educational authorities (Marinucci et al., 2023). Last but not least, the Information Access Dimension assesses how individuals acquire health-related information, including Internet, media, and family health care providers. The indicator system is constructed with multiple sources of data, and each dimension is weighted to produce a composite health literacy score. The composite score is calculated using the equation (6):

$$S_{health} = w_1 \times S_{cognitive} + w_2 \times S_{behaviour} + w_3 \times S_{support} + w_4 \times S_{information} \quad (6)$$

where S_{health} represents the comprehensive health literacy score, w_1, w_2, w_3, w_4 denotes the weighting for each dimension, and $S_{cognitive}, S_{behaviour}, S_{support}, S_{information}$ respectively represent the scores for knowledge, behaviour, environmental support, and information access. The weighting was based on expert scores and analytic hierarchy process (AHP), which ensured that the indicator system had a balance of scientific rigor and practicality. The health literacy assessment indicator system was shown in Table 1.

3.2 Model structure and algorithm selection

To handle heterogeneous and multi-scale features in adolescent health data, we propose a hybrid model that integrates ensemble learning (Random Forest, RF) and deep temporal learning (Long Short-Term Memory, LSTM). The fusion architecture is illustrated in Figure 3 and consists of two encoding paths: structured data (e.g., questionnaire scores, behaviour logs) is processed through RF, while temporal physiological signals (e.g., heart rate, sleep time) are fed into a stacked LSTM network. The Random Forest module aggregates predictions across multiple decision trees and is defined as Equation (7).

$$\hat{y} = \frac{1}{M} \sum_{m=1}^M T_m(x) \quad (7)$$

where $\hat{y}$ is the final forecast output of the model, $T_m(x)$ is the forecast result for the m -th tree, and M is the sum of the trees. RF handles high-dimensional, categorical, and ordinal features efficiently and serves as a strong baseline for population-level health literacy signals (Yeo et al., 2024). For dynamic behaviour modelling, the LSTM module captures sequential dependencies in physiological time series using the following core cell update, as shown in equation (8).

$$h_t = f(W_f x_t + U_f h_{t-1} + b_f) \quad (8)$$

where h_t is the hidden state at time step t , x_t is the current input, and W_f , U_f , and b_f denotes the weight matrix and bias of the forget gate, respectively. The LSTM enables memory retention for long-range dependencies, crucial for interpreting physiological trends in youth behaviour.

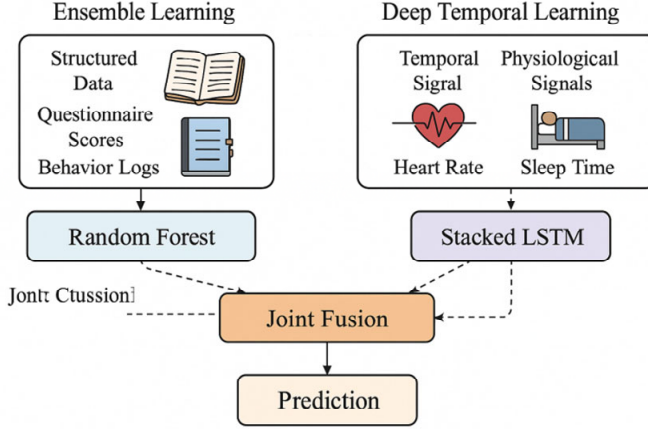
Table 1 Health literacy assessment indicator system

<i>Dimension</i>	<i>Indicator</i>	<i>Scoring Method</i>	<i>Weight</i>
Cognitive Dimension	Health knowledge mastery	Self-assessment scale, behavioural simulation test	0.25
	Disease prevention knowledge	Based on questionnaire survey, multiple-choice scoring	
	Nutrition and exercise knowledge	Questionnaire-based assessment, multiple-choice scoring	
Behavioural dimension	Health behaviour implementation	Daily health behaviour frequency scale	0.30
	Dietary habits	Self-assessment, log recording	
	Physical activity level	Weekly activity duration, activity intensity assessment	
	Tobacco and alcohol use	Frequency of tobacco and alcohol use assessment	
Environmental support dimension	Role of family health behaviour role models	Parental health behaviour assessment form	0.20
	School health education coverage rate	School health education curriculum arrangement survey	
	Accessibility of public health facilities	Community-based survey to assess distribution and utilisation of public health facilities	
Information access dimension	Health information access channels	Information source evaluation form (internet, media, family medical personnel)	0.25
	Health information utilisation	Evaluated via health information effectiveness questionnaire	

After both modules produce latent vectors, a joint fusion layer F_v applies attention weighting to align their contribution and feed the combined vector into the final prediction head. This hybrid design improves both prediction accuracy and interpretability, especially when paired with SHAP-LIME analysis.

Cross-validation with five folds is employed to optimise hyperparameters and prevent overfitting. Evaluation metrics – AUC and F1-score – are used to assess model stability across subgroups and time spans (Lee et al., 2023).

Figure 3 Hybrid model architecture combining ensemble and deep temporal learning for health literacy assessment (see online version for colours)



3.3 Model training and parameter configuration

The model training pipeline is structured into three stages: dataset partitioning, hyperparameter optimisation, and performance evaluation. A 7:2:1 ratio was used to divide the dataset into training, validation, and test subsets. All model parameters were adjusted on the training set and verified against the validation set to avoid overfitting and ensure generalisation (Ortega et al., 2023).

Hyperparameter tuning was performed independently for the random forest (RF) and LSTM modules. Grid Search was applied to RF, exploring combinations of tree count (`n_estimators`), maximum depth (`max_depth`), and node constraints such as `min_samples_split` and `min_samples_leaf`. For the LSTM module, a Random Search approach was used to optimise learning rate (`learning_rate`), dropout rate, and hidden unit size. The search space was guided by prior empirical studies and summarised in Table 2.

An early stopping mechanism was implemented based on validation loss. If no improvement was observed over 10 consecutive epochs, training was halted to avoid overfitting. To further enhance robustness, five-fold cross-validation was adopted across the entire dataset.

During training, the mean squared error (MSE) loss function was used:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (9)$$

In equation (9), where MSE denotes mean squared error, $\hat{y}_i$ represents the model prediction, y_i indicates the true label, and N signifies the sample size. Gradient clipping was applied to LSTM with a norm threshold of 1.0 to avoid gradient explosion. The

Adam optimiser was selected for its adaptive learning rate and convergence stability (Pereira et al., 2023). These settings were selected based on a tradeoff between model performance and computational efficiency, ensuring deployability on edge devices. Evaluation metrics include AUC and F1-score, capturing both ranking quality and class-specific prediction balance.

Table 2 Model parameter configuration

<i>Model name</i>	<i>Parameter name</i>	<i>Value/setting</i>	<i>Parameter description</i>
Random forest model	n_estimators	150	Number of decision trees
	max_depth	12	Maximum depth per tree
	min_samples_split	4	Minimum number of samples for splitting internal nodes
	min_samples_leaf	2	Minimum number of samples for leaf nodes
	max_features	'sqrt'	Number of features considered when finding the optimal split
	bootstrap	True	Whether to use bootstrap sampling to construct subsamples
LSTM model	input_dim	30	Input feature dimension
	hidden_units	128	Number of hidden layer neurons per LSTM unit
	num_layers	2	Number of stacked LSTM layers
	dropout_rate	0.3	Inter-layer Dropout Rate
	batch_size	64	Number of training samples per batch
	learning_rate	0.001	Initial learning rate of the optimiser
	optimiser	Adam	Optimisation algorithm used
	clip_norm	1.0	Gradient clipping threshold
	Early stopping	Enabled (tolerance rounds = 10)	Stop training early if validation set loss does not decrease within 10 epochs

3.4 Model evaluation metrics and performance measures

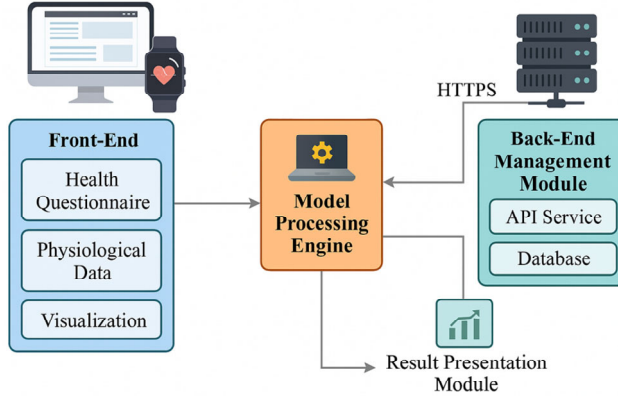
In order to evaluate the performance of the HQES in classification and regression tasks, we designed an integrated approach to evaluate the accuracy and the error measure. For classification tasks, four key metrics are used to measure the performance of the model: precision, accuracy, recall, and FR (Thongnopakun et al., 2023). Among these, the F1-score comprehensively evaluates the model's ability to identify positive samples, calculated as equation (10):

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (10)$$

where $P = \frac{TP}{TP + FP}$ denotes precision, $R = \frac{TP}{TP + FN}$ denotes recall, and TP , FP , FN represent the number of true positive, false positive, and false negative samples,

respectively. In order to increase the reliability of the assessment of the classification ability of a model across different thresholds, the ROC curve and its area measure, area under curve (AUC), are used as additional discrimination criteria. Mean squared error (MSE) is used as an error measure for regression tasks such as the prediction of health literacy scores [for calculation see equation (9)] (Khanal et al., 2023). To ensure the stability and generalisability of all metrics, we apply cross-validation strategies for performance statistics. Based on these measures, the model structure can be adjusted and the parameters can be optimised to guarantee the effectiveness and adaptability of the assessment results.

Figure 4 System functional architecture diagram (see online version for colours)



3.5 System implementation or prototype design

To implement the deployment of HCI model, we have developed a prototype system based on a front-end-back separation architecture. The system is composed of four parts: data collection subsystem, model processing engine, result display module, and back-end management module. The Vue Framework is used to create a responsive user interface, which can be used to input health questionnaire, to upload physiological status data, and to visualise model results. The back-end uses Python to deploy model API services using the flask framework, which allows efficient data storage and state management using Redis and MySQL databases (Zhao et al., 2025). The model service exposes RESTful interfaces supporting input of standardised feature vectors $\mathbf{X} = \mathbb{R}^{n \times m}$, where n denotes sample count and m represents feature dimensions. Health literacy predictions are computed via the model function $f_{\theta}(\mathbf{X})$, as shown in equation (11).

$$\hat{\mathbf{y}} = f_{\theta}(\mathbf{X}) = \sigma(\mathbf{W} \times \mathbf{X} + \mathbf{b}) \quad (11)$$

where $\hat{\mathbf{y}}$ represents the predicted output value, $\mathbf{W}$ denotes the model weight matrix, $\mathbf{b}$ indicates the bias term, and σ signifies the activation function (e.g., Sigmoid). The system is based on the Alibaba Cloud ECS, which includes a 4-core CPU, 16 GB memory, and an instance of GPU-accelerated artificial intelligence computing (Zhamantayev et al., 2023). As illustrated in Figure 4, the system uses HTTPS protocol to guarantee the security of communication, and supports concurrent multi-device access, and makes the

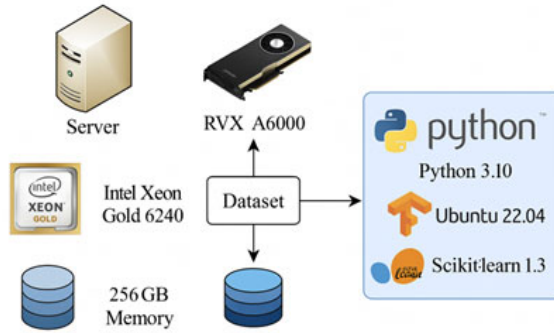
results traceable. The prototype has proved its strong scalability and cross-platform adaptability, and has provided the basis for regional public health administration.

4 Empirical research and results analysis

4.1 Experimental setup

The aim of this study is to evaluate its applicability to real data situations, as illustrated in Figure 5, which comprises an Intel Xeon Gold 6240 (18 cores, 2.6 GHz), RAM of 256 GB, and NVIDIA RTX A6000 GPU. The operating system was Ubuntu 22.04, with Python 3.10 as the development environment. The primary frameworks included TensorFlow 2.13 and Scikit-learn 1.3 (Mancone et al., 2024). The data was collected from a sample of local high schools, which included health surveys, physiological sensors, and school environment data from 2,000 teenagers. After the integration, a unified structured data set of 42 features was created. The dataset was divided into 7: 2: 1 for training, for validation, and for testing. Five times cross validation was adopted to guarantee the generalisation capability. Model training was carried out with a batch size of 64 and a time step of 30. To avoid overfitting, we set up a learning rate adjustment strategy to optimise the convergence efficiency. All the training and assessment tasks were performed separately on the isolated GPU nodes to ensure that the models were comparable and the results were consistent.

Figure 5 Experimental platform architecture diagram (see online version for colours)



4.2 Model performance comparison

To evaluate the adaptability and predictive capability of the proposed multi-source data fusion framework, three baseline models – support vector machine (SVM), Random forest (RF), and LSTM – were trained under identical preprocessing, feature construction, and cross-validation settings. Performance metrics included Accuracy, Precision, Recall, F1-score, AUC, and MSE (Wu, 2023). All experiments were conducted using a unified hardware and software environment, and the final values represent the mean results obtained from five-fold cross-validation. The detailed comparison results are presented in Table 3.

Across all metrics, the LSTM model achieved the highest overall performance, with an Accuracy of 0.889 and an AUC of 0.921. This advantage is attributable to its ability to capture long-term dependencies in physiological time-series data and to leverage the temporally aligned features generated in the fusion module. The RF model performed competitively on structural features but exhibited weaker recall (0.851), suggesting that ensemble-based tree models may have difficulty detecting subtle transitional patterns contained in sequential physiological variables. In contrast, SVM demonstrated the lowest performance across most metrics, particularly in MSE (0.062), reflecting its sensitivity to heterogeneous feature distributions and its reliance on manually crafted kernels, which are less effective for high-dimensional fused representations.

The higher recall (0.882) and F1-score (0.873) observed in the LSTM model indicate that dynamic behavioural variations – such as changes in sleep continuity or activity cycles – were captured more reliably. This aligns with the multi-source architecture’s emphasis on time alignment and temporal fusion. However, LSTM also incurs higher computational cost, which may affect its suitability for certain lightweight deployments.

Overall, the results in Table 3 illustrate that models capable of exploiting temporal continuity and fused feature interactions tend to produce superior predictions in adolescent health literacy assessment, while static models show limitations in handling multimodal dynamics.

Table 3 Model performance metrics comparison

<i>Model name</i>	<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1 score</i>	<i>AUC</i>	<i>MSE</i>
LSTM	0.889	0.865	0.882	0.873	0.921	0.034
Random forest	0.864	0.843	0.851	0.847	0.891	0.049
Support vector machine	0.836	0.812	0.825	0.818	0.871	0.062

4.3 Feature importance and interpretability analysis

To improve model transparency and understand the role of individual features in health literacy assessment, both global and local interpretability methods were employed – SHAP for population-level insights and LIME for instance-specific reasoning (Ma et al., 2025).

4.3.1 Global feature attribution via SHAP

Shapley additive explanations (SHAP), based on cooperative game theory, calculates the marginal contribution of each feature to the final prediction. Unlike traditional variable importance, SHAP considers feature interaction and contextual impact. As shown in Figure 6 (left), the most influential features globally include health knowledge, physical activity, and dietary habits. These factors consistently show positive SHAP values, contributing to increased prediction confidence for high health literacy scores.

In contrast, features such as screen time (> 4 hours), support system quality, and sleep duration often exhibit negative SHAP values, indicating inhibitory effects. Moreover, feature interactions – e.g., high physical activity mitigating the negative impact of poor sleep duration – were observed, reinforcing the benefit of a multi-source fusion approach that captures cross-domain behaviour patterns.

4.3.2 Local sample explanation via LIME

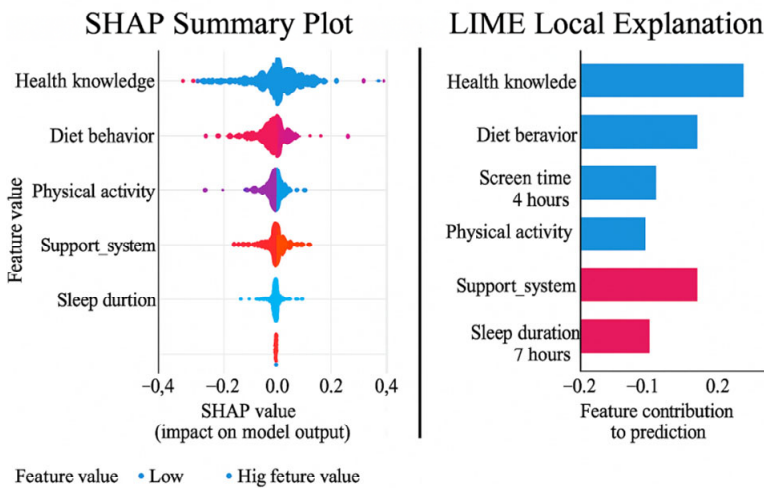
To interpret individual-level predictions, local interpretable model-agnostic explanations (LIME) was used to construct linear approximations around each data instance. For selected adolescents, LIME identified health knowledge and diet behaviour as key positive predictors, whereas excessive screen time and weak family support emerged as negative drivers.

The local explanation graph in Figure 6 (right) visualises LIME's ranking of input features for a specific prediction, demonstrating how short-term behavioural attributes dominate individual classification decisions. LIME's sensitivity to local feature distributions complements SHAP's global trends.

4.3.3 Comparative insights and model feedback

By combining SHAP and LIME, the model achieves both global coherence and individual interpretability. This dual-perspective analysis enables dynamic feature reweighting – for instance, penalising overlong screen exposure or enhancing the learning signal from consistent physical routines. These insights guide future adjustments in model architecture and feature embedding strategies.

Figure 6 SHAP value visualisation (see online version for colours)



4.4 Heterogeneous group analysis

To examine the generalisability of the proposed health literacy assessment model across different subpopulations, a heterogeneous group analysis was conducted based on age, gender, and health status. The dataset was first stratified into two age groups (12–15 and 16–18 years) and by gender (male and female). The predictive performance for each subgroup is summarised in Table 4.

The older group (16–18 years) achieved a higher accuracy of 0.89 compared with 0.82 for the younger group. This gap is consistent with the SHAP global feature analysis, in which variables such as Health Knowledge, Physical Activity, and Dietary Habits

demonstrated stronger stability and lower variance among older adolescents. As older youths typically exhibit more regular routines and more complete physiological cycles, the temporal encoder (LSTM) benefits from clearer sequential patterns, thereby enhancing prediction accuracy.

Female participants showed an accuracy of 0.87, approximately 3% higher than the 0.84 observed in males. LIME-based local explanations revealed that females demonstrated more consistent patterns in health-related behaviours, particularly in dietary habits and sleep routines, reducing feature fluctuation and improving model reliability. This aligns with the higher positive SHAP contributions assigned to behaviour-related variables in the fusion model.

Further comparison between individuals with and without chronic health conditions indicated larger prediction errors in the chronic disease group. The increased error is primarily attributed to higher intra-individual variability in physiological signals (e.g., irregular heart rate cycles) and inconsistent behavioural patterns, which challenge the temporal modelling capability of LSTM. Additionally, missing data rates were higher in this subgroup due to incomplete device usage, leading to weaker feature embedding quality.

These findings suggest that subgroup-specific calibration could enhance predictive robustness. Examples include adjusting feature weights for high-variance groups, incorporating domain adaptation modules for chronic disease populations, or applying stratified fine-tuning based on age and gender characteristics. Such adaptive strategies may further improve the model’s applicability across diverse youth populations.

Table 4 Comparison of health literacy assessment model accuracy across groups

<i>Group category</i>	<i>Age group 12–15</i>	<i>Age group 16–18</i>	<i>Male</i>	<i>Female</i>
Accuracy	0.82	0.89	0.84	0.87

5 Conclusions

This study presents a comprehensive framework for adolescent health literacy assessment through multi-source data fusion, integrating questionnaire responses, physiological signals, and environmental variables. The proposed hybrid architecture, which combines Random Forest for structured data processing and LSTM for temporal modelling, demonstrates robust predictive capabilities across age, gender, and health condition subgroups. The incorporation of SHAP and LIME ensures model interpretability at both global and local levels, enhancing the transparency and practical applicability of the assessment system.

Despite its strengths, the study has certain limitations. The dataset primarily consists of youth participants from a specific region, which may introduce population bias and limit the model’s generalisability. In addition, the computational complexity of the fusion model, particularly the LSTM components, poses challenges for large-scale or real-time deployment. The model also assumes consistent signal quality, which may not hold in community-based or sensor-sparse settings. Addressing these issues will require improvements in model compression, signal preprocessing robustness, and domain adaptation mechanisms.

Future research will explore lightweight architectures to support deployment on edge devices, as well as transfer learning techniques to adapt the model to other regions or populations. Further, integrating adaptive fusion strategies that dynamically adjust to feature availability and quality may improve performance in data-incomplete environments. These extensions can facilitate the model's broader implementation in digital public health initiatives, youth health monitoring platforms, and personalised health education systems.

Funding

This work was supported by Doctoral and Professor Startup Project Fund of Hanshan Normal University (Reference No. QD2024116).

Declarations

The author declares that this article is free of any conflicts of interest.

References

- Chen, P. and Li, J. (2024) 'Construction of a physical health monitoring platform for adolescents driven by big data', *2024 4th International Conference on Information Technology and Contemporary Sports (TCS)*, IEEE, Guangzhou, China, pp.120–123.
- Chou, A.K., Liao, C.H. and Chen, D.R. (2023) 'Measuring adolescent health literacy in Taiwan: validation of the health literacy assessment scale for adolescents', *BMC Public Health*, Vol. 23, No. 1, p.2409.
- Curran, T., Ito-Jaeger, S., Perez Vallejos, E. and Crawford, P. (2023) 'What's up with everyone?': The effectiveness of a digital media mental health literacy campaign for young people', *Journal of Mental Health*, Vol. 32, No. 3, pp.612–618.
- Goodfellow, C., Macintyre, A., Knifton, L. and Sosu, E. (2023) 'Associations between dimensions of mental health literacy and adolescent help-seeking intentions', *Child and Adolescent Mental Health*, Vol. 28, No. 3, pp.385–392.
- Khanal, S.P., Budhathoki, C.B., Okan, O., van Teijlingen, E., Sharma, M.K., Acharya, J. et al. (2023) 'Systematic review of health literacy and health promotion in school-aged adolescents', *Journal of Education and Community Health*, Vol. 10, No. 1, pp.49–57.
- Lee, J.E., Goh, M.L. and Yeo, S.F. (2023) 'Mental health awareness of secondary schools students: mediating roles of knowledge on mental health, knowledge on professional help, and attitude towards mental health', *Heliyon*, Vol. 9, No. 3, p.e14512.
- Ma, X., Zhou, P. and He, X. (2025) 'Advances in multi-source navigation data fusion processing methods', *Mathematics*, Vol. 13, No. 9, p.1485.
- Mancone, S., Corrado, S., Tosti, B., Spica, G. and Diotaiuti, P. (2024) 'Integrating digital and interactive approaches in adolescent health literacy: a comprehensive review', *Frontiers in Public Health*, Vol. 12, p.1387874.
- Marinucci, A., Grove, C. and Allen, K.A. (2023) 'A scoping review and analysis of mental health literacy interventions for children and youth', *School Psychology Review*, Vol. 52, No. 2, pp.144–158.

- Ortega, F.B., Leskošek, B., Gil-Cosano, J.J., Mäestu, J., Tomkinson, G.R., Ruiz, J.R. et al. (2023) 'European fitness landscape for children and adolescents: updated reference values, fitness maps and country rankings based on nearly 8 million test results from 34 countries gathered by the FitBack network', *British Journal of Sports Medicine*, Vol. 57, No. 5, pp.299–310.
- Pereira, A.F., Escola, J.J.J., Almeida, C.M.T. and Rodrigues, V.M.C.P. (2023) 'Health education provided by nurses to children and young people: parents' assessment', *BMC Nursing*, Vol. 22, No. 1, p.287.
- Renwick, L., Pedley, R., Johnson, I., Bell, V., Lovell, K. and Bee, P. (2024) 'Mental health literacy in children and adolescents in low-and middle-income countries: a mixed studies systematic review and narrative synthesis', *European Child & Adolescent Psychiatry*, Vol. 33, No. 4, pp.961–985.
- Sun, X., Lv, K., Wang, F., Ge, P., Niu, Y., Yu, W. et al. (2023) 'Validity and reliability of the Chinese version of the health literacy scale short-form in the Chinese population', *BMC Public Health*, Vol. 23, No. 1, p.385.
- Thongnopakun, S., Manwong, M., Visanuyothin, S., Wattanaburanon, A., Suwannarat, W. and Magteppong, W. (2023) 'Development and validation of a sexual health literacy measurement tool for pregnancy prevention among adolescents (SHL-PPA)', *Journal of Multidisciplinary Healthcare*, Vol. 16, pp.2417–2430.
- Vasli, P., Shekarian-Asl, Z., Zarmehrparirouy, M. and Hosseini, M. (2024) 'The predictors of COVID-19 preventive health behaviors among adolescents: the role of health belief model and health literacy', *Journal of Public Health*, Vol. 32, No. 1, pp.157–166.
- Wu, Y.C. (2023) 'Evaluation method of physical education students' mental health based on multi-source and heterogeneous data', *International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 10, pp.138–146.
- Yeo, G.H., Reich, S.M., Liaw, N.A. and Chia, E.Y.M. (2024) 'The effect of digital mental health literacy interventions on mental health: systematic review and meta-analysis', *Journal of Medical Internet Research*, Vol. 26, p.e51268.
- Zhamantayev, O., Nukeshtayeva, K., Kayupova, G., Daulekaliyeva, Z., Bolatova, Z., Takuadina, A. et al. (2023) 'Mapping the terrain: a comprehensive exploration of health literacy among youth', *Journal of Clinical Medicine of Kazakhstan*, Vol. 20, No. 6, pp.12–22.
- Zhang, X., Yue, H., Hao, X., Liu, X. and Bao, H. (2023) 'Exploring the relationship between mental health literacy and psychological distress in adolescents: a moderated mediation model', *Preventive Medicine Reports*, Vol. 33, p.102199.
- Zhao, Z., Zhu, L., Liao, J., Xia, J. and Pu, X. (2025) 'Parental mediation, digital media usage, and health literacy: an exploration among Chinese elementary school students', *Health Communication*, Vol. 40, No. 6, pp.1144–1156.
- Zhu, D. (2025) 'Knowledge creation in vocational education using multi-source data fusion under big data environment', *International Journal of Information and Communication Technology*, Vol. 26, No. 5, pp.69–83.
- Zhu, Y. and Deng, J. (2025) 'Multi source data fusion and graph neural network based joint identification algorithm for adolescent internet addiction and psychological disorders', *2025 2nd International Conference on Digital Media, Communication and Information Systems (DMCIS)*, IEEE, Nanjing, China, pp.100–104.