Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 34(9): 6286–6314.
DOI: 10.3934/era.2026275
Received: 20 November 2025
Revised: 01 July 2026
Accepted: 09 July 2026
Published: 17 July 2026

Research article

Semi-supervised hyperbolic graph deep clustering for estimation of pediatric reference intervals

Jianguo Zheng[1,2,†], Fanxia Zeng[3,†], Yongqiang Tang[1,2,*], Xiaoxia Peng[4], Ruohua Yan[4], Jun Zhao[4], Yaguang Peng[4,*] and Wensheng Zhang[1,2,*]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

[3] School of Artificial Intelligence, Shenzhen Polytechnic University, Shenzhen 518055, China

[4] National Center for Clinical Laboratories, National Center for Children's Health, Beijing Children's Hospital, Capital Medical University, Beijing 100045, China

† The authors contributed equally to this work.

* **Correspondence:** Email: yongqiang.tang@ia.ac.cn, plwumi@hotmail.com, zhangwenshengia@hotmail.com.

Abstract: Pediatric biomarker reference intervals (RIs) are crucial for clinical diagnosis and health assessment, but the establishment of RIs is often limited by issues such as scarce labeled data and the inadaptability of traditional methods to implicit hierarchical data distributions. To overcome these challenges, we proposed a semi-supervised model integrating partial supervised information guidance and hyperbolic space representation for efficient prediction of pediatric biomarker reference intervals. By introducing hyperbolic space exponential mapping, we mapped the preprocessed high-dimensional features from Euclidean space to the hyperbolic space of the Poincaré ball. Leveraging the negative curvature property of hyperbolic space, the model more accurately characterizes the hierarchical distribution patterns of biomarkers. The partially labeled samples guide hyperbolic representation learning through a supervised loss, and the structural information of unlabeled samples guide the reconstruction of unsupervised loss. This enables the collaborative use of a small number of labeled samples and a large number of unlabeled samples, balancing model performance and data collection cost. Analytical and comparative experiments were conducted on pediatric datasets using two biomarkers, validating the proposed method as a proof of concept.

Keywords: indirect estimation; reference interval; deep clustering; graph neural network; semi-supervised learning

1. Introduction

As a core benchmark for clinical decision-making, the biomarker reference interval (RI) defines the normal range of physiological indicators for specific populations [1]. In pediatric clinical settings, it serves as a critical link between laboratory test results and clinical diagnosis. First, it assists clinicians in rapidly distinguishing pathological abnormalities from age-related physiological fluctuations during children's growth and development [2]. Second, it provides a basis for the longitudinal monitoring of healthy children, facilitating the tracking of health growth and the identification of subtle abnormality indicators preceding overt health issues [3]. Given the dual value of pediatric RIs in emergency diagnosis and preventive medicine, establishing accurate RIs that align with the characteristics of the pediatric population holds significant clinical importance. Figure 1 illustrates the process of establishing RIs using healthy populations. The "gold standard" [4] for establishing biomarker RIs is the direct method recommended by the Clinical and Laboratory Standards Institute (CLSI) [5] international guidelines. This method establishes RIs through a rigorous process: first, strict inclusion/exclusion criteria are developed to identify the healthy population; second, biological samples from this population are collected and target biomarkers are measured; and finally, RIs are calculated based on the distribution of measured values (typically the 2.5th to 97.5th percentiles).

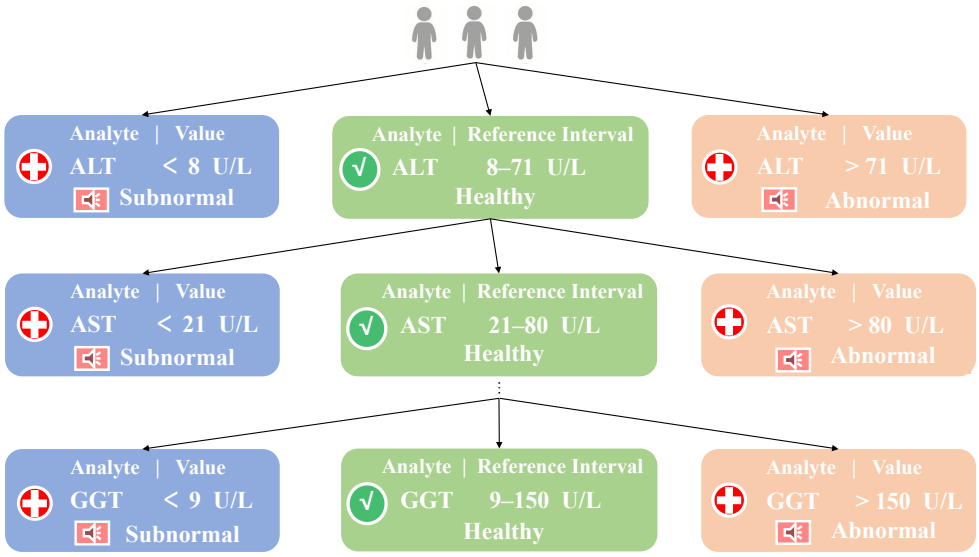

Figure 1. Process of establishing reference intervals using healthy populations.

Although the direct method enables the generation of highly reliable RIs by leveraging explicitly labeled healthy samples, its application in the pediatric field is constrained by notable challenges. To begin with, the definition of health in children is shaped by multiple factors including age and gender, making the dataset complex and hard to represent effectively. This renders it difficult to develop unified inclusion criteria for the reference population. Moreover, the labeling of samples as healthy or nonhealthy demands professional clinical assessments, and it is both resource-intensive and time-consuming [6]. What is more, the collection of pediatric samples is subject to ethical restrictions. These factors collectively tend to result in insufficient labeling of the reference population

or poor representativeness of the sampled group [7]. These limitations have driven the exploration of the indirect method—which performs clustering on unlabeled data—as an alternative approach for establishing pediatric RIs.

While the indirect method demonstrates significant advantages in the utilization efficiency of hospitals' laboratory information system (LIS) data, the lack of labels makes it difficult to rigorously verify the clinical validity of RIs derived through clustering methods. So clinicians generally find it hard to trust RIs that have not undergone correlation verification with the "gold standard" [4], which in turn significantly undermines the application value of RIs derived from the indirect method in practical clinical decision-making.

To address this dilemma, we leverage a semi-supervised learning (SSL) framework to avoid informational loss from discarding either labeled or unlabeled data. The abundant unlabeled data covering a broader pediatric population implies the underlying data structure, while the scarce healthy-labeled data shows credibility in anchoring real physiological benchmarks. Notably, state-of-the-art SSL methods are characterized by strong adaptability to limited annotations and heterogeneous data, as they can enhance model robustness under partial labeling [8] and realize effective utilization of multi-site heterogeneous unlabeled data [9]. Aligning with semi-supervised learning (SSL) [10–12], this work exploits the synergy between sparse annotations and large-scale unlabeled distributions to enhance the performance of pediatric reference interval estimation.

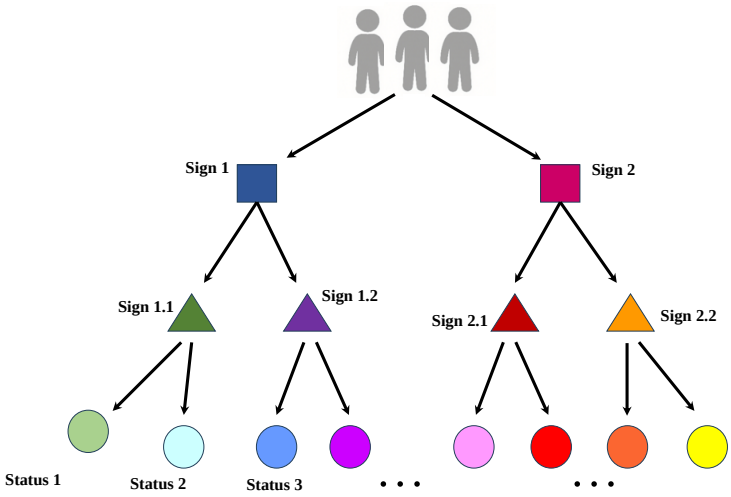

Figure 2. Hierarchical structure diagram for diagnosing a child's health status using multiple biomarkers.

Although mainstream SSL (such as label propagation [13] and constrained K-means semi-supervised clustering [14]) have demonstrated potential in biomedical data analysis, Euclidean space is suboptimal for the encoding of the hierarchical structure. Figure 2 illustrates the diagnostic process based on multiple biomarkers, which is a tree-like decision-making framework. Following the diagnostic tree from the root node, the health status of each child patient can be assessed through the hierarchical assessment of biomarkers. In particular, the hyperbolic space is more suitable to encode hierarchies than the Euclidean space [15, 16]. Mathematically defined, hyperbolic space is a type of

n-dimensional Riemannian manifold [17] with constant negative curvature [18]. Its core characteristic is its volume growing exponentially with the radius, which distinguishes it from Euclidean space and spherical space [17]. Hyperbolic space enables the independent expansion of each branch through low-dimensional manifolds without sacrificing structural resolution. This stands in sharp contrast to the limitations of spherical space (branch compression) and Euclidean space (branch flattening) [19, 20].

Based on the above analysis, we propose a semi-supervised hyperbolic graph autoencoder (SS-HGAE) for the establishment of pediatric RIs. First, in the encoding layer of the graph autoencoder, samples are mapped to the constrained Poincaré ball via exponential mapping to capture the implicit hierarchical physiological characteristics of the evaluated alanine aminotransferase (ALT) and aspartate aminotransferase (AST) data, which improves the representation of the dataset. Second, by integrating the angular-based and distance-based loss between samples in hyperbolic space, supervised hyperbolic contrastive losses with partial labels are constructed to align these embedded representations. This ensures the clusters are clinically valid. Finally, after the model converges during training, the unsupervised Gaussian mixture model [21] is adopted for the clustering, ultimately establishing RIs. To verify the effectiveness of the proposal, two types of biomarkers, ALT and AST, were selected from the LIS database for experimental validation. Experiment datasets are from five groups with different age periods and genders: "28 days to < 1 year", "1 to < 2 years", "2 to < 13 years", "13 to < 18 years (male)", and "13 to < 18 years (female)". Our experimental results demonstrate that the proposed framework achieves favorable performance across all datasets as a proof of concept. The contributions of this work can be summarized as follows:

- Operating under strictly low-resource settings, we address the real-world characteristics of pediatric clinical data, which features scarce labels alongside ample unlabeled samples. With the dual impetus of the graph reconstruction loss and supervised hyperbolic contrastive loss, the framework realizes the coordination of feature excavation from unlabeled samples and a limited number of labeled samples.
- According to the tree-like hierarchical characteristics of the evaluated pediatric biomarker data, we integrate hyperbolic geometry into the embedding space to capture better representation.
- As a proof of concept, extensive experiments on pediatric ALT and AST datasets covering different age groups and genders consistently demonstrate the feasibility and effectiveness of the proposal.

2. Related works

Indirect estimation methods in the medical field can be categorized into two major technical branches: probability distribution model-based methods and neural network-based methods. As the mainstream approach in the early stage, probability distribution model-based methods have evolved around the optimization of distribution assumption adaptability and the enhancement of practicality, while gradually exposing inherent limitations at the technical level.

In the initial stage of technical development, indirect estimation methods based on probability distribution models [22] were centered on the Gaussian distribution assumption: they determined the RI boundaries by analyzing the central smooth peak characteristics of the data distribution and combining probability with linear regression tools. Later, a graphical analysis approach [23] was further developed, which derived RIs by identifying Gaussian peaks in data histograms and using the

population with the largest peak as the reference.

However, such methods are constrained by the strict Gaussian distribution assumption. In contrast, real-world medical data often exhibit asymmetric, skewed, and other non-Gaussian features. Meanwhile, the operation mode relying on manual visual inspection not only leads to strong subjectivity in results but also fails to meet the needs of automated analysis, resulting in obvious shortcomings in the technical applicability. So research shifted toward adapting to non-Gaussian distributions, and new methods introduced data transformation and iterative optimization strategies. For example, the Box-Cox transformation [24] was used to process the raw data, and an iterative algorithm was proposed to locate the optimal truncation interval of the reference value distribution, enabling parameter estimation in non-Gaussian scenarios. Alternatively, the truncated minimum chi-square (TMC) approach [25] improved the parameter estimation and normality test logic of the previous method [24], thereby enhancing result reliability. Nevertheless, such methods still face practical bottlenecks: their implementation is limited to Excel, making it impossible to integrate into automated analysis workflows and to support large-scale data-processing requirements. Although subsequent studies [21, 26] promoted further optimization of method practicality by developing command-line tools to establish analysis pipelines and designing reverse modeling algorithms to improve estimation efficiency and quality, indirect estimation methods [27] based on probability distribution models never overcame the core limitation of fixed distribution assumptions. The restrictive mathematical distribution cannot characterize the nonlinear correlations among features in high-dimensional medical data space, nor can they capture the synergistic or inhibitory relationships among multiple biomarkers. In medical data scenarios with multi-dimensional and complex interactions, the model's adaptation to the inherent laws of data is insufficient, and technical limitations have not been fundamentally resolved.

In contrast, deep clustering methods [28], relying on the nonlinear representation capability of neural networks, can automatically learn the mapping of high-dimensional features while preserving the inherent correlations among features. This provides an effective approach [29, 30] for modeling complex interactions in multi-dimensional medical data. In the early development stage, deep clustering was mainly done with a dimension-reduction-based approach [31]: after extracting low-dimensional features of data via autoencoders, traditional clustering algorithms such as K-means [32] were exploited. Later, hybridizing feature learning and clustering together, studies realized the synchronous optimization of the two by defining clustering-oriented objective functions. For instance, an objective function was constructed based on Kullback-Leibler divergence [33], enabling the deep model to complete a sample clustering assignment while learning the mapping from the observation space to the low-dimensional latent space. Huang et al. proposed DeepCluE [34], which integrates instance-level and cluster-level contrastive learning with multi-layer ensemble clustering to fully exploit diverse semantic information from different network layers and enhance clustering robustness. Ju et al. [35] proposed a fuzzy deep clustering method that combines self-attention fuzzy feature extraction and dual-granularity contrastive learning to address data uncertainty and suboptimal negative sample selection in traditional contrastive learning. However, such methods still focused on the intrinsic feature representation of individual samples and failed to capture the correlation information among samples, making it difficult to adapt to the structural characteristics of sample dependence in medical data.

To improve the structural information modeling, graph autoencoders (GAEs) [36] were proposed. By combining graph convolutional encoders with inner-product decoders, the GAE converts corre-

lations among samples into graph structures for learning, realizing the joint modeling of feature and structure information. In medical scenario applications, studies further constructed density maps [37] to characterize the similarity of analyte measurements between patients, used a GAE to learn patient embedding representations and clustering, and finally achieved RI establishment. This promoted the in-depth integration of deep clustering and medical tasks. With technological iteration, deep clustering frameworks have gradually developed toward multi-module integration [38] and performance optimization [39]. On one hand, a unified framework, the structural deep clustering network (SDCN) [40], was built by integrating autoencoders (AEs) with graph convolutional networks, and a dynamic information fusion module [41] was designed to perform refined processing of attribute and structure information, enhancing the comprehensiveness and accuracy of feature representation. On the other hand, to address the problem of representation collapse that is common in previous frameworks, the deep graph clustering via dual correlation reduction method (DCRN) [42] introduced information correlation reduction mechanisms from both sample and feature levels, effectively alleviating this defect. The diverse-bellwether deep graph clustering method (DDGC) [43] proposed a dual-module framework with diverse centroid learning and minimum entropy-based assignment sharpening to balance intra-cluster compactness and inter-cluster separation while mitigating over-smoothing. Deep graph clustering through effective distance and density (DeepEDD) [44] designed a nonparametric framework integrating a graph-masking autoencoder and effective distance-density decoder to automatically determine cluster numbers without prior knowledge. Despite the significant progress of deep clustering, it faces a core limitation in the task of establishing pediatric biomarker RIs: such methods rely on an unsupervised paradigm. However, pediatric biomarker data generally lack real labels for healthy samples, making it difficult to verify their clinical effectiveness through empirical methods, which in turn limits the application value of the methods in actual clinical scenarios. Therefore, the introduction of a semi-supervised learning paradigm is of great necessity and practical significance.

Semi-supervised learning aims to optimize models through the collaboration of a small number of labeled samples and a large number of unlabeled samples, and its development has always centered on the core issue of how to effectively utilize unlabeled data. The early semi-supervised methods [10] were mainly based on traditional machine learning frameworks, and relied on manually designed features and simple assumptions. Self-training [45, 46] generates pseudo-labels through high-confidence predictions of the model for iterative updates; although it is prone to being affected by incorrect pseudo-labels, it is widely used in scenarios such as text classification and image recognition due to its simplicity. Co-training [47] trains two models to transmit pseudo-labels to each other by leveraging the independence of multiple data views, which requires satisfying the assumption of sufficient and conditionally independent views. Later, single-view co-training [48] expanded its scope of application by constructing pseudo-views through random feature projection and different hyperparameters. Graph-based methods are a core branch of transductive learning. Label propagation [13] updates node labels iteratively to make node labels approach the labels of neighboring nodes; it is computationally efficient but depends on the quality of the graph structure. Gaussian random fields [49] treat label prediction as a probabilistic inference problem and use Laplacian graph regularization to ensure smooth prediction, laying the foundation for manifold regularization. A semi-supervised support vector machine (SVM) [50] extends a supervised SVM by optimizing the margin of unlabeled data to the decision boundary, but it faces the problem of non-convex optimization. Subsequent studies improved

its practicality through convex relaxation and gradient descent approximation. With the rise of deep learning, semi-supervised methods have shifted toward combining with neural networks, focusing on solving the problem of deep models' dependence on labeled data. Semi-supervised variational autoencoders (SSVAEs) [51] treat class labels as latent variables and jointly optimize generation and classification objectives through variational inference, which is suitable for scenarios with extremely limited labeled data. Guan et al. [52] proposed S^2Match, which adopts a linear policy-based self-paced sampling strategy to avoid repeated learning of incorrect pseudo-labels in data-limited scenarios and mitigate the risk of falling into local minima. Weng et al. [8] proposed a semi-supervised information fusion framework that integrates pseudo-labeling, consistency regularization, and other methods with data, feature, and decision fusion technologies to enhance the robustness of medical image analysis models. Xu et al. [9] proposed a separated collaborative learning framework, which realizes effective utilization of multi-site heterogeneous unlabeled data through local distribution fitting and external information-mining based on mutual information and adversarial stability learning.

However, semi-supervised methods based on Euclidean space have limitations in handling the hierarchical structure inherent in pediatric biomarker data, making it difficult to achieve optimal classification performance. Notably, there exists an inherent compatibility between such hierarchical structures and the geometric properties of hyperbolic space. This characteristic provides a potentially feasible pathway for optimizing the classification scheme of pediatric biomarker data. From a mathematical definition, hyperbolic space is a type of n-dimensional Riemannian manifold with constant negative curvature [18]. Its core characteristic that distinguishes it from Euclidean space and spherical space is that its volume grows exponentially with radius [53]. This unique geometric property makes it naturally suitable for encoding tree-like nested structures. Nickel et al. [54] created a method based on Poincaré embedding through Riemannian optimization to describe the similarity and hierarchical relationships between objects. Subsequently, they developed a hyperbolic method from object similarity using the Lorentz model, which improved numerical stability compared with the Poincaré model [55]. Existing studies have shown that neural networks such as the Visual Geometry Group network (VGG) and ResNet inherently exhibit hyperbolic properties when learning from ImageNet and the Canadian Institute for advanced research dataset (CIFAR) [17]. Ganea et al. [15] first designed a hyperbolic deep learning version by extending computational units such as multinomial logistic regression in hyperbolic space. Different from previous shallow networks, hyperbolic attention networks were introduced into deep networks, which utilize an efficient version of the hyperbolic attention mechanism [56]. A more general version of hyperbolic neural networks was proposed by constructing parameter-efficient multinomial logistic regression, convolutional layers, and attention mechanisms in the Poincaré ball model [57]. A large number of studies have proven that hyperbolic space has strong expressive capabilities in different tasks, such as natural language processing [15], graph analysis [56], recommendation systems [58], image classification [17], and generalized category discovery [59].

Although existing hyperbolic methods are not designed for medical datasets and cannot be directly applied in our context, their success implies the powerful expression of hyperbolic embedding. Inspired by the above analysis, this study proposes a semi-supervised method based on hyperbolic space, investigates its performance on pediatric blood biomarker concentration data across different ages and genders, and examines the impact of label information with different annotation ratios on the proposed method.

3. Materials and methods

3.1. Dataset and preprocessing method

The core variables of the data used in this study include test ID, age, gender, biochemical analytes, test units, and test sample type. To better clarify the correlation between diseases or discomfort symptoms and the biomarkers under study, this study extracted the outpatient impression diagnosis variable for reference. This study collected test information on 13 pediatric biomarkers, specifically including total protein (TP), albumin (ALB), creatinine (Crea), urea (Urea), alanine aminotransferase (ALT), aspartate aminotransferase (AST), γ-glutamyl transferase (GGT), alkaline phosphatase (ALP), calcium (Ca), inorganic phosphorus (IP), potassium (K), sodium (Na), and chlorine (Cl). All biochemical analysis data were measured by the Central Laboratory of Beijing Children's Hospital using the Cobas C702 testing instrument (Roche Diagnostics GmbH, Mannheim, Germany); the internal quality control of biomarkers complied with the requirements of the National Standard of the National Health Commission of the People's Republic of China (WS/T 780-2021) [60]. Clinical age and gender stratifications differed drastically across blood biomarkers. Therefore, we selected ALT and AST as core validation targets to explore our semi-supervised method, driven by their distinct and challenging statistical trajectories.

During the data-cleaning process, samples with missing values in core variables (such as age, gender, and biomarker concentration) were directly excluded instead of performing missing value imputation. The age division standard in this study refers to the national industry standards commonly used in Chinese pediatric clinical practice; therefore, children aged < 28 days or ≥ 18 years were not included, and only serum samples were retained as the sample type. Using the test ID as the unique identifier, the study converted the long-format data of test results into wide-format data to ensure that each piece of data corresponds to an individual child. In addition, considering that samples with multiple tests have a higher potential for pathological conditions, this study carefully screened child samples with repeated test records: for individuals with 2 or more laboratory test records within 1 year, only their earliest test record was retained.

The cleaned dataset originates from two independent sources: 63,915 strictly unlabeled samples extracted from the hospital's LIS, and 11,714 labeled healthy samples obtained from the Pediatric Reference Interval in China (PRINCE) study [4]. The unlabeled LIS data were derived from the biochemical test results of the Outpatient Laboratory Department of Beijing Children's Hospital from January 1, 2021, to December 31, 2021. The labeled data were collected by obtaining children's blood samples through direct sampling and labeling health information. In accordance with national industry standards [60], and considering the differences in ALT and AST levels among children of different age groups and genders, the study subjects were divided into five groups, namely: "28 days to < 1 year", "1 to < 2 years", "2 to < 13 years", "13 to < 18 years (male)", and "13 to < 18 years (female)". Table 1 presents the detailed data distribution of different groups. The dataset sizes of each group are 3243, 4962, 58,731, 4492, and 4201, respectively. The quantities of labeled data included in each group are 463, 500, 8,006, 1,228, and 1,517, respectively. To investigate the impact of the label ratio and strictly prevent data leakage, 10%, 20%, 30%, 50%, and 70% of the labeled PRINCE samples (11,714 samples) were randomly selected as supervised information for ALT and AST, respectively, in Section 4.4.

Table 1. Original distribution of ALT and AST across different age and gender groups.

Biomarkers	Age / sex Partition	Total Number	Number of Labels	Mean (U/L)	SD (U/L)	Median (U/L)	$P_{2.5}$ (U/L)	$P_{97.5}$ (U/L)	Proportion in RIs (%)	WS/T (U/L)
ALT	28 days to < 1 year	3243	463	33.7	34.23	25.8	10.5	100.3	93.71	8~71
	1 to < 2 years	4962	500	21.1	25.86	17.2	8.9	52.9	95.24	8~42
	2 to < 13 years	58,731	8006	17.7	30.91	13.3	7.2	50.6	91.99	7~30
	13 to < 18 years male	4492	1228	19.6	22.29	13.6	6.3	78.6	89.67	7~43
	13 to < 18 years female	4201	1517	15.4	24.02	11.2	5.7	50.8	90.91	6~29
AST	28 days to < 1 year	3243	463	51.0	40.36	44.3	26.9	112.1	93.68	21~80
	1 to < 2 years	4962	500	43.1	25.61	40.1	27.1	70.1	94.68	22~59
	2 to < 13 years	58,731	8006	30.4	22.61	27.9	16.8	51.6	94.85	14~44
	13 to < 18 years male	4492	1228	23.0	33.21	20.1	13.2	45.7	94.59	12~37
	13 to < 18 years female	4201	1517	20.2	16.25	17.8	12.1	38.5	95.02	10~31

3.2. Problem definition

The task of pediatric reference-interval prediction based on semi-supervised learning aims to develop a model that can accurately cluster unlabeled samples by learning from a small number of labeled samples. Let the unlabeled dataset be denoted as $D_u = \{(\mathbf{x_{u_i}}, y_{u_i})\} \in \mathcal{X} \times \mathbf{Y}_u$ and the labeled dataset be represented as $D_l = \{(\mathbf{x_{l_i}}, y_{l_i})\} \in \mathcal{X} \times \mathbf{Y}_l$, where $\mathbf{Y}_u$ and $\mathbf{Y}_l$ denote the label sets corresponding to the two types of datasets, respectively. The unlabeled dataset contains both unlabeled and labeled categories, which satisfies $\mathbf{Y}_l \subset \mathbf{Y}_u$. After obtaining the clustering results, the largest cluster is selected as the healthy population according to the principle that the proportion of healthy individuals is the highest, and its 2.5th percentile and 97.5th percentile are taken as the lower and upper bounds of the reference interval, respectively.

3.3. Overall architecture

This study proposes a novel semi-supervised hyperbolic graph autoencoder (SS-HGAE). In this subsection, the overall architecture of the proposed model is first introduced, and its detailed framework is presented in Figure 3. The SS-HGAE model consists of five modules, including three network structure modules and two optimization algorithm modules. Specifically, the graph autoencoder, hyperbolic space exponential representation learning, and hidden layer representation clustering together form the overall network structure of the SS-HGAE model; furthermore, in the optimization process, supervised hyperbolic contrastive loss and reconstruction loss are adopted to enhance the overall representational capability of the model. Detailed descriptions of each module and the optimization process will be elaborated in subsequent subsections.

3.3.1. Graph autoencoder

Given the unlabeled dataset $D_u = \{(\mathbf{x_{u_i}}, y_{u_i})\} \in \mathcal{X} \times \mathbf{Y}_u$ and the labeled dataset $D_l = \{(\mathbf{x}_{l_i}, y_{l_i})\} \in \mathcal{X} \times \mathbf{Y}_l$, the node feature matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times F}$ is composed of individual sample features (including both labeled and unlabeled samples) $\mathbf{x}_i$, where $i = 1, 2, \ldots, N$ and F represents the number of biomarkers. We first construct an adjacency matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{N \times N}$ to characterize pairwise correlations between samples. Specifically, we employ the K-nearest neighbor (KNN) method [61]

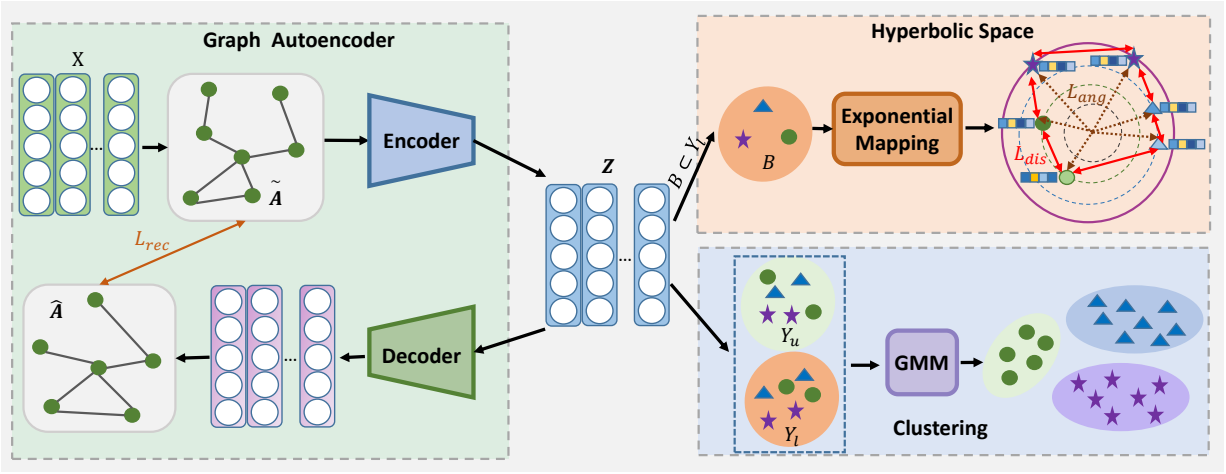

Figure 3. The overall architecture of the SS-HGAE model. The sample data $\mathbf{X}$ is first used to construct the normalized adjacency matrix $\tilde{\mathbf{A}}$, and then both $\mathbf{X}$ and $\tilde{\mathbf{A}}$ are input into the encoder to derive the hidden representation $\mathbf{Z}$. The update of $\mathbf{Z}$ is jointly determined by two types of loss functions: one is the loss function constructed under hyperbolic exponential mapping, and the other is the inherent reconstruction loss of the graph autoencoder.

to build this matrix: for each sample i, we compute its Euclidean distance to all other samples in the original feature space and then select its K most similar samples to form local connections. The adjacency matrix element a_{ij} is set to 1 if sample j is among the K-nearest neighbors of sample i, and is 0 otherwise. This KNN-based graph construction captures the local manifold structure of the data, ensuring that each sample is only connected to its most relevant peers, which helps the graph encoder to focus on meaningful pairwise relationships while avoiding noisy global connections. Subsequently, the node feature matrix $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$ are fed into a graph encoder to obtain the hidden-layer Euclidean feature representation $\mathbf{z}_i$ corresponding to sample i. The full embedding matrix of the node is denoted as $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N\times P}$, where P denotes the dimension of the latent space after applying the graph encoder. The graph encoder operates based on two stacked graph convolutional layers, with its mathematical implementation governed by the following expressions: First, the adjacency matrix $\mathbf{A}$ is normalized using the degree matrix $\mathbf{Q}$ to yield the normalized adjacency matrix $\tilde{\mathbf{A}}$:

$$\tilde{\mathbf{A}} = \mathbf{Q}^{-\frac{1}{2}}\mathbf{A}\mathbf{Q}^{-\frac{1}{2}}, \tag{3.1}$$

where $\mathbf{Q} = \mathrm{diag}(q_1, q_2, \ldots, q_N) \in \mathbb{R}^{N\times N}$, $q_i = \sum_{j=1}^{N} a_{ij}$, and a_{ij} denotes the element of $\mathbf{A}$ at the i-th row and j-th column.

The normalized adjacency matrix $\tilde{\mathbf{A}}$ and node feature matrix $\mathbf{X}$ then serve as inputs to these two graph convolutional layers. The first layer applies the rectified linear unit activation function ReLU$(\cdot)$ to the linear transformation of $\mathbf{X}$ with learnable parameter $\mathbf{W}_1 \in \mathbb{R}^{F\times F_1}$, where F_1 denotes the output dimension of the first graph convolutional layer. This activation function ensures non-negative output to introduce nonlinearity, and the result is subsequently aggregated via $\tilde{\mathbf{A}}$; the second layer further transforms the output with learnable parameter $\mathbf{W}_2 \in \mathbb{R}^{F_1\times P}$ to generate the final embedding matrix of node $\mathbf{Z}$:

$$\mathbf{Z} = \tilde{\mathbf{A}}\mathrm{ReLU}\left(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}_1\right)\mathbf{W}_2. \tag{3.2}$$

Additionally, the graph encoder incorporates a graph reconstruction branch, where the embedding matrix of node $\mathbf{Z}$ is used to reconstruct the adjacency matrix via the sigmoid activation function $\sigma(\cdot)$. This activation function maps values to the range $[0, 1]$ for similarity measurement, yielding the reconstructed adjacency matrix $\hat{\mathbf{A}} \in \mathbb{R}^{N\times N}$:

$$\hat{\mathbf{A}} = \sigma\left(\mathbf{Z}\mathbf{Z}^{\top}\right). \tag{3.3}$$

As the core optimization objective of graph autoencoders, the reconstruction loss directly defines the difference between reconstructed results and original inputs. In this study, the mean squared error (MSE) loss is employed as the reconstruction loss. By minimizing this reconstruction loss, the encoder is compelled to learn embeddings that preserve key information of the original graph, such as node connection relationships and feature similarities. Its mathematical expression is as follows:

$$L_{rec} = \frac{1}{N^2}\|\tilde{\mathbf{A}} - \hat{\mathbf{A}}\|_F^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\tilde{a}_{ij} - \hat{a}_{ij}\right)^2, \tag{3.4}$$

where $\tilde{a}_{ij}$ and $\hat{a}_{ij}$ are elements of matrices $\tilde{\mathbf{A}}$ and $\hat{\mathbf{A}}$, respectively.

3.3.2. Hyperbolic space exponential representation learning

In this section, we focus on projecting the Euclidean-space representation $\mathbf{z}$ into the hyperbolic space, where similarity computation for positive and negative sample pairs (constructed based on label information) is performed. This design is motivated by the intrinsic structural characteristics of pediatric biomarker data—particularly critical for the task of reference interval establishment.

In the scenario of pediatric biomarker reference interval establishment, a synergistic health-status determination mechanism using multiple biomarkers [60] inherently endows the data with a hierarchical structure. This structure is explicitly manifested in the two-stage clinical decision-making process: the first stage involves single-index normal range determination (e.g., ALT 8–71 U/L, AST 21–80 U/L, GGT 9–150 U/L; any index exceeding its range initially signals potential health abnormalities), while the second stage requires synergistic normal determination using multiple indices (only when all core biomarkers fall within their respective single-index ranges can the health status be finally classified as healthy).

Existing semi-supervised methods [62] primarily focus on optimizing clustering performance of unlabeled samples in Euclidean or spherical spaces. However, the geometric properties of these spaces (e.g., polynomial volume growth in Euclidean space) limit their capacity to effectively capture such hierarchical relationships. In contrast, hyperbolic space—characterized by exponential volume growth with respect to radius—naturally aligns with the hierarchical nature of pediatric biomarker data, making it a more suitable representation space for this task.

There are five isometric models available for hyperbolic space with constant negative curvature [63]. Considering computational feasibility, we adopt the widely used Poincaré ball model [53, 54, 59]. Given the negative curvature $-c^2$ $(c > 0)$, the n-dimensional Poincaré ball is denoted as $\mathbb{D}_c^n = \left\{\mathbf{u} \in \mathbb{R}^n \mid c\|\mathbf{u}\|^2 < 1\right\}$, which is equipped with the Riemannian metric: $g_{\mathbf{u}}^c = (\lambda_{\mathbf{u}}^c)^2 g^{\mathbb{E}}$, where $\lambda_{\mathbf{u}}^c = \frac{2}{1-c\|\mathbf{u}\|^2}$, and $g^{\mathbb{E}}$ is the Euclidean metric. This metric scales local distances by $\lambda_{\mathbf{u}}^c$, which diverges near the ball's boundary, imparting exponential expansion properties to hyperbolic space.

The basic operation in Euclidean space cannot be directly used in hyperbolic space such as vector addition, which consequently prohibits the use of standard Euclidean vector addition. To address this, our approach adopts the gyrovector framework [64]. This formalism introduces Möbius addition of any two points $\mathbf{u}, \mathbf{v} \in \mathbb{D}_c^n$ that faithfully respect the hyperbolic geometry:

$$\mathbf{u} \oplus_c \mathbf{v} = \frac{\left(1 + 2c\langle \mathbf{u}, \mathbf{v}\rangle + c\|\mathbf{v}\|^2\right)\mathbf{u} + \left(1 - c\|\mathbf{u}\|^2\right)\mathbf{v}}{1 + 2c\langle \mathbf{u}, \mathbf{v}\rangle + c^2\|\mathbf{u}\|^2\|\mathbf{v}\|^2}, \tag{3.5}$$

where $\langle \mathbf{u}, \mathbf{v}\rangle$ denotes the Euclidean inner product of vectors $\mathbf{u}$ and $\mathbf{v}$ (quantifying the linear correlation between the two vectors), and $\|\mathbf{u}\|$ represents the Euclidean norm (i.e., L_2 norm) of vector $\mathbf{u}$, calculated as the square root of the sum of the squares of its components, which measures the magnitude of the vector. Based on Möbius addition, the hyperbolic distance between $\mathbf{u}$ and $\mathbf{v}$ is calculated as:

$$\mathcal{D}_{\mathbb{H}}(\mathbf{u}, \mathbf{v}) = \frac{2}{\sqrt{c}} \operatorname{arctanh}\left(\sqrt{c}\,\|-\mathbf{u} \oplus_c \mathbf{v}\|\right). \tag{3.6}$$

It is obvious that $\lim_{c\to 0} \mathcal{D}_{\mathbb{H}}(\mathbf{u}, \mathbf{v}) = 2\|\mathbf{u} - \mathbf{v}\|$. This convergence ensures the consistency with Euclidean geometry in the limit.

Thus, the addition operation and distance operation can be conducted in hyperbolic space according to Eqs (3.5) and (3.6), respectively. To map Euclidean feature embeddings $\mathbf{z}$ into hyperbolic space $\mathbb{H}^n$, we utilize an exponential mapping [18], facilitating the effective capture of hierarchy. For base point $\mathbf{o}$, the mathematical formulation of projecting a tangent vector $\mathbf{z}$ from $\mathbb{E}^n$ to $\mathbb{H}^n$ is:

$$\mathcal{H}(\mathbf{z}) = \exp_{\mathrm{o}}^c(\mathbf{z}) = \mathbf{o} \oplus_c \left(\tanh\left(\sqrt{c}\frac{\lambda_{\mathrm{o}}^c\|\mathbf{z}\|}{2}\right)\frac{\mathbf{z}}{\sqrt{c}\|\mathbf{z}\|}\right), \tag{3.7}$$

where Möbius addition is defined as Eq (3.5). To enforce the constraint that the feature vector $\mathbf{z}$ lies within the Poincaré ball and to ensure numerical stability, we apply a clipping operation before the exponential mapping: $C(\mathbf{z}) = \min\left\{1, \frac{\beta}{\|\mathbf{z}\|}\right\} \cdot \mathbf{z}$, where β is a constant clipping value. For a Euclidean feature vector $\mathbf{z}_i$, its hyperbolic counterpart is thus $\mathcal{H}(\mathbf{z}_i) = \exp_{\mathbf{o}}^c(C(\mathbf{z}_i))$, adhering to hyperbolic geometric operation norms.

To investigate how the number of labeled samples affects model performance, we adopt a subset of labeled samples from the label set with a fixed ratio to provide supervision for contrastive learning. Let $B \subseteq \mathbf{Y}_l$ denote the subset of labels sampled from the labeled label set $\mathbf{Y}_l$ at a sampling ratio ρ ($0 < \rho \leq 1$). Thus, we define the unified form of supervised contrastive loss as follows:

$$L^s = \frac{1}{|B_l|}\sum_{i\in B_l}\frac{1}{|N_i|}\sum_{q\in N_i} -\log\frac{\exp\left(\mathcal{S}\left(\mathcal{H}(\mathbf{z}_i), \mathcal{H}\left(\mathbf{z}_q\right)\right)/\tau_r\right)}{\sum_{j\neq i}\exp\left(\mathcal{S}\left(\mathcal{H}(\mathbf{z}_i), \mathcal{H}\left(\mathbf{z}_j\right)\right)/\tau_r\right)}, \tag{3.8}$$

where $B_l \subset B$ is the labeled mini-batch (i.e., a subset of the sampled label set B), N_i is the set of samples sharing the same label as $\mathbf{z}_i$ in B_l, τ_r is the temperature parameter, and $\mathcal{S}$ represents the similarity function for the hyperbolic counterparts of two feature vectors.

For pairwise sample similarity computation $\mathcal{S}$, we propose a hybrid loss strategy integrating distance-based and angle-based components—consistent with prior findings that such integration enhances model optimization in hyperbolic space [59, 65]. The distance-based similarity is defined as a negative hyperbolic distance:

$$\mathcal{S}_d = -\mathcal{D}_{\mathbb{H}}. \tag{3.9}$$

This is similar to a negative Euclidean distance of prior methods [66]. The angle-based similarity is defined as cosine similarity, which is preserved across Euclidean and hyperbolic spaces due to their conformal equivalence:

$$\mathcal{S}_{ang}\left(\mathcal{H}\left(\mathbf{z}_i\right), \mathcal{H}\left(\mathbf{z}_j\right)\right) = \frac{\mathcal{H}\left(\mathbf{z}_i\right) \cdot \mathcal{H}\left(\mathbf{z}_j\right)}{\left\|\mathcal{H}\left(\mathbf{z}_i\right)\right\| \cdot \left\|\mathcal{H}\left(\mathbf{z}_j\right)\right\|}. \tag{3.10}$$

Substituting Eqs (3.9) and (3.10) into Eq (3.8), we obtain supervised contrastive losses L^s_{dis} and L^s_{ang} in the form of distance and angle, respectively. The final supervised hyperbolic contrastive loss combines these two components with a loss coefficient α_d:

$$L^s_{hyp} = \alpha_d L^s_{dis} + (1 - \alpha_d) L^s_{ang}. \tag{3.11}$$

The loss coefficient α_d increases linearly from 0 to 1.0. This gradual weighting strategy is designed to first emphasize the angular component L^s_{ang} at the early training stage, which helps to capture the global hierarchical similarity structure in the hyperbolic space. As training progresses, the increasing α_d gradually shifts the focus to the distance component L^s_{dis}, which refines the fine-grained discriminative power of the model by enforcing precise distance constraints between samples with different labels.

Finally, the model's total loss integrates the supervised hyperbolic contrastive loss and the graph autoencoder's reconstruction loss. The supervised hyperbolic contrastive loss captures hierarchical similarity, while the graph autoencoder's reconstruction loss ensures feature fidelity. The total loss is defined as:

$$L_{total} = L^s_{hyp} + \gamma L_{rec}, \tag{3.12}$$

where γ denotes the weight hyperparameter for the graph autoencoder's reconstruction loss. To highlight the primacy of label information and the supervised hyperbolic contrastive loss, we set the weight hyperparameter γ for the reconstruction loss to 0.5.

3.3.3. Latent clustering

To cluster the learned sample representations, we select the classical and most widely used clustering method based on the Gaussian mixture model (GMM) [21]. The probability density function of the GMM consists of a weighted sum of K Gaussian distributions, which is formally defined as:

$$p(\mathbf{z} \mid \Theta) = \sum_{k=1}^{K} \alpha_k \cdot \mathcal{N}\left(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{3.13}$$

where $\Theta = \{\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ denotes the set of model parameters, and each parameter satisfies the following conditions: α_k is the weight of the k-th component, which satisfies $\sum_{k=1}^K \alpha_k = 1$ and $0 \leq \alpha_k \leq 1$, representing the proportion of this component in the mixture model, and $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the probability density function (PDF) of the d-dimensional Gaussian distribution, which is given by:

$$\mathcal{N}\left(\mathbf{z} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \frac{1}{(2\pi)^{d/2} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{z} - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{z} - \boldsymbol{\mu}_k\right)\right), \tag{3.14}$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is the mean vector of the k-th component, $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ is a positive definite covariance matrix, and $|\boldsymbol{\Sigma}_k|$ denotes the determinant of the covariance matrix. Guided by clinical standards and

previous research [21, 28, 67], we set the number of clusters to $K = 3$, representing three physiological states: low, normal, and high.

To formalize the clustering mechanism, we introduce a latent variable $z_{ik} \in \{0, 1\}$ indicating whether the i-th sample $\mathbf{z}_i$ is generated by the k-th Gaussian component, with the constraint $\sum_{k=1}^{K} z_{ik} = 1$ (each sample belongs to exactly one component). The posterior probability γ_{ik}, quantifying the confidence that $\mathbf{z}_i$ originates from the k-th component, is derived using Bayes' theorem as:

$$\gamma_{ik} = p(z_{ik} = 1 \mid \mathbf{z}_i, \Theta) = \frac{\alpha_k \cdot \mathcal{N}(\mathbf{z}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma} k)}{\sum_{j=1}^{K} \alpha_j \cdot \mathcal{N}(\mathbf{z}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}, \tag{3.15}$$

where $\gamma_{ik} \in [0, 1]$ and $\sum_{k=1}^{K} \gamma_{ik} = 1$, reflecting the probabilistic membership of $\mathbf{z}_i$ across all components.

Since the latent variables z_{ik} are unobserved, direct maximum likelihood estimation of Θ is infeasible. Instead, we employ the expectation-maximization (EM) [68] algorithm to iteratively optimize the parameters. This iterative process continues until parameter convergence is achieved. After discovering convergent parameters, each data point $\mathbf{z}_i$ is assigned to the component with the maximum posterior probability, which is expressed as:

$$\text{label}\,(\mathbf{z}_i) = \arg \max_{k \in \{1,\ldots,K\}} \gamma_{ik}, \tag{3.16}$$

where label($\mathbf{z}_i$) denotes the cluster label of $\mathbf{z}_i$, which ultimately allows the cluster partitioning of the dataset.

4. Results and discussion

4.1. Baseline models

To validate the advancement of the proposed model, we select the following models as baseline models: Gaussian mixture model (GMM) [21], K-means clustering [32], semi-supervised Gaussian mixture model (SSGMM) [69], semi-supervised K-means (Semi-K-means) [14], semi-supervised graph attention network (Semi-GAT) [70], semi-supervised graph convolutional network (Semi-GCN) [71], and semi-supervised variational autoencoder (SSVAE) [51]. Detailed descriptions of these models are provided below:

GMM [21]: The GMM is a probability-based generative model. Its core idea is to assume that the probability distribution of data is formed by the weighted combination of multiple Gaussian distributions with different parameters, where each sub-component corresponds to a type of potential pattern that may exist in the dataset.

K-means [32]: K-means is a classic unsupervised machine learning algorithm, whose core objective is to partition a given dataset into a pre-specified number of K non-overlapping clusters. This partitioning ensures that data samples within each cluster are as similar as possible to one another, while the differences between samples from different clusters are as large as possible.

SSGMM [69]: The SSGMM is a probabilistic generative model improved by incorporating semi-supervised learning ideas into the traditional unsupervised GMM. It inherits the GMM's core framework of assuming the data distribution is composed of a weighted combination of multiple Gaussian sub-components but introduces class constraints from labeled samples during parameter

estimation (such as iteratively updating the mean, covariance, and weights of each sub-component via the EM algorithm).

Semi-K-means [14]: Semi-K-means inherits the iterative framework of K-means, which consists of initializing cluster centers, assigning samples, and updating centers. However, it introduces label constraints during the sample assignment: labeled samples are enforced to be fixedly assigned to the cluster corresponding to their annotated class, while unlabeled samples are still assigned according to the nearest distance principle; in the center update phase, all assigned samples (including both labeled and unlabeled ones) jointly participate in calculating new cluster centers, thereby optimizing the cluster boundaries.

SSVAE [51]: The SSVAE is a semi-supervised generative model improved based on the variational autoencoder framework. It takes the output of the approximate posterior distribution of latent variables as classifier features to enhance feature robustness. Its core lies in constructing a joint generative process by introducing categorical variables and continuous latent variables, enabling it to model data distribution using both a small number of labeled samples and a large number of unlabeled samples simultaneously.

Semi-GAT [70]: The Semi-GAT is a graph-neural-network model that embeds the semi-supervised learning paradigm into the unsupervised graph attention network. It inherits the GAT's core logic of aggregating neighbor node features based on attention weights—by calculating attention coefficients between target nodes and their neighbors, it performs weighted aggregation of neighbor features to update node representations. Meanwhile, it incorporates semi-supervised constraints during training, optimizing model parameters solely based on the classification loss of a small number of labeled nodes, enabling unlabeled nodes to indirectly learn category information through graph connectivity.

Semi-GCN [71]: The Semi-GCN integrates semi-supervised learning into the unsupervised graph convolutional network. It abandons the reliance on graph Laplacian regularization in traditional graph-based semi-supervised learning; instead, by making the model depend on both the node feature matrix and the graph adjacency matrix, it enables the propagation of gradient information from the supervised loss. This allows the model to be optimized using only the supervised information from a small number of labeled nodes, thereby realizing the classification of a large number of unlabeled nodes.

4.2. Evaluation metrics

To comprehensively evaluate the prediction accuracy of various reference interval establishment methods, we adopt five commonly used evaluation metrics in the field [67, 72], including lower-bound error rate e_{lower}, upper-bound error rate e_{upper}, overall error rate e_{overall}, lower-bound bias rate BR_{LL}, and upper-bound bias rate BR_{UL}. The definitions and calculation methods of each metric are as follows, where x_{lower} and x_{upper} represent the lower and upper bounds of the reference interval predicted by the model, respectively, and $x_{\text{actual, lower}}$ and $x_{\text{actual, upper}}$ represent the lower and upper bounds of the ground-truth reference interval, respectively.

The error rate metrics directly reflect the accuracy of RI establishment by quantifying the relative deviation between the predicted bounds and the ground-truth bounds. Specifically, they include three categories: lower-bound error rate, upper-bound error rate, and overall error rate.

The lower-bound error rate e_{lower} is used to evaluate the degree of deviation between the lower bound of the predicted reference interval and the ground-truth lower bound, with the calculation formula as

follows:

$$e_{\text{lower}}\% = \frac{x_{\text{lower}} - x_{\text{actual, lower}}}{x_{\text{actual, lower}}}. \tag{4.1}$$

Similarly, the upper-bound error rate e_{upper} measures the deviation between the predicted upper bound (tail) and its ground truth, with the formula:

$$e_{\text{upper}}\% = \frac{x_{\text{upper}} - x_{\text{actual, upper}}}{x_{\text{actual, upper}}}. \tag{4.2}$$

The overall error rate e_{overall} is designed to comprehensively reflect the model's overall estimation accuracy for both the lower and upper bounds of the interval. It adopts the average of the absolute error rates of the lower and upper bounds as the comprehensive indicator, with the calculation formula as follows:

$$e_{\text{overall}}\% = \frac{1}{2}\left(\frac{\left|x_{\text{lower}} - x_{\text{actual, lower}}\right|}{x_{\text{actual, lower}}} + \frac{\left|x_{\text{upper}} - x_{\text{actual, upper}}\right|}{x_{\text{actual, upper}}}\right). \tag{4.3}$$

This metric takes into account the estimation quality of both ends of the interval. The closer its value is to 0, the better the overall prediction performance of the model.

The bias rate metrics eliminate the impact of the ground-truth interval width on bias evaluation through normalization, making them more suitable for cross-scenario comparison. Specifically, they include two categories: lower-bound bias rate and upper-bound bias rate. The lower-bound bias rate BR_{LL} measures the relative offset of the predicted lower bound against the width of the ground-truth reference interval, calculated as:

$$BR_{\text{LL}}\% = \frac{x_{\text{lower}} - x_{\text{actual, lower}}}{x_{\text{actual, upper}} - x_{\text{actual, lower}}}. \tag{4.4}$$

Similarly, the upper-bound bias rate BR_{UL} quantifies the relative offset of the predicted upper bound using the same ground-truth interval width as the reference, with the formula:

$$BR_{\text{UL}}\% = \frac{x_{\text{upper}} - x_{\text{actual, upper}}}{x_{\text{actual, upper}} - x_{\text{actual, lower}}}. \tag{4.5}$$

Compared with the upper-bound error rate, this metric is not affected by the absolute size of the ground-truth interval and is more suitable for evaluating the robustness of the model. Values of the lower-bound error rate e_{lower} and lower-bound bias rate BR_{LL} closer to 0 indicate better alignment between the predicted and ground-truth lower bound. Similarly, values of the upper-bound error rate e_{upper} and upper-bound bias rate BR_{UL} nearer to 0 signify better matching between the predicted and ground-truth upper bounds.

To align with the practical requirements of RI establishment for medical data, we employ the following evaluation rules: first, regarding the meaning of metric signs, a positive metric value indicates that the predicted reference interval is wider than the ground-truth reference interval, while a negative value signifies that the predicted reference interval is narrower than the ground-truth counterpart; second, as the core evaluation criterion, the absolute value of the metric serves as the primary evaluation standard, with a smaller absolute value denoting higher prediction accuracy; and third, following the safety-first rule, when different methods yield metrics with equal absolute values, the negative result (corresponding to a narrower predicted interval) is deemed superior to the positive result (corresponding to a wider predicted interval) for medical safety considerations—specifically, to mitigate the risk of missed judgment arising from an excessively narrow predicted interval.

4.3. Experimental setup

We conduct a comprehensive evaluation of the proposed method on multiple datasets encompassing diverse ages, genders, and biomarkers. Specifically, the evaluated cohorts include five age and gender groups: "28 days to < 1 year", "1 to < 2 years", "2 to < 13 years", "13 to < 18 years (male)", and "13 to < 18 years (female)", along with two biomarkers (ALT and AST). For each group, which includes both labeled and unlabeled samples, we randomly select a subset of the labeled part as supervised samples, while the remaining labeled samples are treated as unlabeled. This design constructs a one-class semi-supervised setting. For the proposed model, all experiments adopt the graph autoencoder as the backbone network structure. The Adam optimizer is used during the training process, with the learning rate set to 0.001, the number of training epochs to 100, and the curvature parameter c to 0.1. By default, the loss weight α_d increases linearly from 0 to 1.0. In addition, the number of clusters is set to $K = 3$, the node representation dimension F is 13, the hyperbolic space projection dimension is 256, and the deep learning framework uses Python 3.8.20 and PyTorch 2.4.1.

4.4. Experiment on the impact of labeling ratio: identifying the optimal configuration

To systematically investigate the impact of the label-utilization rate, experiments were conducted with different proportions of the labeled data: 10%, 20%, 30%, 50%, and 70%. The remaining labeled samples are regarded as unlabeled. The quantitative indicators of model performance corresponding to each label utilization rate are presented in Table 2, while the curve of the performance with varying label utilization rates is intuitively illustrated in the performance curve graph of Figure 4. Analysis of the experimental results shows that under different volumes of labeled data, the model exhibits high stability in predicting the lower bound of the reference interval (i.e., head interval). This indicates that the prediction of the head interval is weakly affected by fluctuations in the volume of labeled data. Therefore, subsequent model performance evaluation focuses on the upper bound of the reference interval (i.e., tail interval), which is more sensitive to the volume of labeled data. The core evaluation indicators include upper-bound error rate e_{upper}, upper-bound bias rate BR_{UL}, and overall error rate e_{overall}, aiming to more accurately capture the impact of the labeled data volume on the model's key predictive capabilities.

From the overall trend in Figure 4, the performance of SS-HGAE does not scale monotonically with the increase of the label-utilization rate, peaking at 10% to 30% and degrading beyond 30%. As the proportion of labeled data increases from 10% to 70%, the predicted reference interval narrows. This nuanced phenomenon arises primarily because the labels are derived exclusively from one-class healthy samples. Therefore, a high label ratio signifies a severe one-class imbalance, which amplifies the distribution mismatch between the labeled data and the true unlabeled samples. Specifically, abundant healthy labels pull the model toward the normal core, artificially shrinking the estimated intervals. The optimal label utilization rate is highly context-dependent. At a 10% label rate, the model achieves the best comprehensive performance across six experimental groups. When the rate increases to 20%, the consistency between the predicted and true reference intervals peaks in four experimental groups. Notably, in the "13 to < 18 years (male)" subgroup, the prediction consistency for both ALT and AST reaches its optimum at a 30% label rate, yielding e_{overall} values as low as 10.28 and 8.70, respectively.

Synthesizing the experimental results across multiple scenarios, the label utilization rate for optimal model performance is typically between 10% and 30%. Specifically, the 10%–20% label utilization

Table 2. Model performance under different label utilization rates. The rank-1 results are in boldface.

Age/Sex Partition	Label Utilization Rate	Biomarkers									
		ALT					AST				
		e_{lower}	BR_{LL}	e_{upper}	BR_{UL}	e_{overall}	e_{lower}	BR_{LL}	e_{upper}	BR_{UL}	e_{overall}
28 days to < 1 year	10%	33.75	4.29	**−5.92**	**−6.67**	**19.83**	30.95	11.02	**−0.25**	**−0.34**	**15.60**
	20%	33.75	4.29	−7.61	−8.57	20.68	30.95	11.02	**−0.25**	**−0.34**	**15.60**
	30%	33.75	4.29	−7.04	−7.94	20.40	30.95	11.02	−0.75	−1.02	15.85
	50%	32.50	4.13	−11.69	−13.17	22.10	30.95	11.02	−2.88	−3.90	16.91
	70%	36.25	4.60	−28.45	−32.06	32.35	32.38	11.53	−12.50	−16.95	22.44
1 to < 2 years	10%	13.75	3.24	**−13.81**	**−17.06**	**13.78**	24.09	14.32	4.41	7.03	14.25
	20%	12.50	2.94	−21.67	−26.76	17.08	23.64	14.05	**−1.69**	**−2.70**	**12.67**
	30%	12.50	2.94	−22.38	−27.65	17.44	25.00	14.86	−1.69	−2.70	13.35
	50%	12.50	2.94	−25.00	−30.88	18.75	26.36	15.68	−3.73	−5.95	15.05
	70%	12.50	2.94	−25.95	−32.06	19.23	26.36	15.68	−4.75	−7.57	15.55
2 to < 13 years	10%	2.86	0.87	**−2.67**	**−3.48**	**2.76**	22.14	10.33	**0.00**	**0.00**	**11.07**
	20%	2.86	0.87	−5.33	−6.96	4.10	22.86	10.67	**0.00**	**0.00**	11.43
	30%	2.86	0.87	−6.00	−7.83	4.43	22.14	10.33	−0.23	−0.33	11.19
	50%	2.86	0.87	−20.33	−26.52	11.60	22.86	10.67	−1.82	−2.67	12.34
	70%	2.86	0.87	−21.67	−28.26	12.26	22.86	10.67	−1.59	−2.33	12.22
13 to < 18 years male	10%	−12.86	−2.50	−25.58	−30.56	19.22	9.17	4.40	−15.14	−22.40	12.15
	20%	−12.86	−2.50	−15.35	−18.33	14.10	18.33	8.80	−7.03	−10.40	12.68
	30%	−14.29	−2.78	**−6.28**	**−7.50**	**10.28**	14.17	6.80	**−3.24**	**−4.80**	**8.70**
	50%	−15.71	−3.06	−33.95	−40.56	24.83	13.33	6.40	−15.41	−22.80	14.37
	70%	−14.29	−2.78	−30.70	−36.67	22.49	17.50	8.40	−12.70	−18.80	15.10
13 to < 18 years female	10%	−1.67	−0.43	−12.07	−15.22	6.87	16.00	7.62	**−16.45**	**−24.29**	**16.23**
	20%	−1.67	−0.43	**−12.07**	**−15.22**	**6.87**	18.00	8.57	**−16.45**	**−24.29**	17.23
	30%	−8.33	−2.17	−13.10	−16.52	10.72	18.00	8.57	−16.45	−24.29	17.23
	50%	−5.00	−1.30	−12.76	−16.09	8.88	18.00	8.57	−16.45	−24.29	17.23
	70%	−5.00	−1.30	−10.69	−13.48	7.84	19.00	9.05	−16.13	−23.81	17.56

rate constitutes the core interval for balancing performance and labeling cost. Within this interval, the model not only achieves optimal or sub-optimal performance in 80% of the datasets but also reduces the labeling cost to 1/5–1/10 of the full labeling set, significantly cutting down the labor and time investment in clinical data labeling. This conclusion provides quantitative guidance for clinical data collection: when applying the model proposed in this study, only 10%–20% of the normal sample labeling volume needs to be collected to meet the requirements for efficient and accurate reference interval prediction. Take the "28 days to < 1 year" group (totaling 3243 samples) as an example in Table 1: the labeled samples among all samples are 1517/4201 = 0.36. Therefore, 20% of the labeled samples utilized accounts for less than $0.36 \times 20\% = 7.2\%$ of the total samples, which is below 8%.

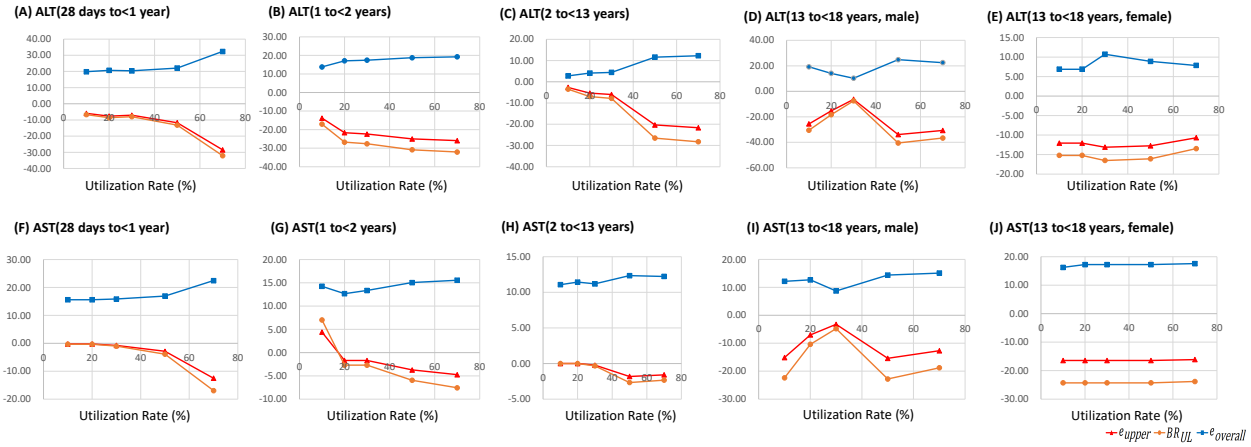

Figure 4. Utilization rate comparison of reference intervals for ALT and AST among different pediatric age-gender groups. The graph is divided into ten subgraphs based on indicator types and population characteristics. Subgraphs (A)–(E) correspond to the five age-gender groups of the ALT ("28 days to < 1 year", "1 to < 2 years", "2 to < 13 years", "13 to < 18 years [male]", "13 to < 18 years [female]"), while subgraphs (F)–(J) correspond to the same groups of the AST. The x-axis represents the utilization rate (%), while the y-axis intuitively presents the numerical differences and fluctuation characteristics of the three indicators (e_{upper}, BR_{UL}, e_{overall}) among different populations.

For all subsequent comparative experiments, we adopted a 20% label utilization rate. This rate not only yields optimal performance across most datasets but also provides a near-optimal basis for complex subgroups (e.g., "13 to < 18 years, female"). Training all baseline and proposed models on this unified benchmark ensures that any performance differences stem solely from architectural designs rather than variations in data configuration.

4.5. Comparative experiment: SS-HGAE vs. baseline methods

To systematically evaluate the effectiveness of the proposed SS-HGAE in pediatric biomarker reference interval prediction, we selected two categories of representative baseline methods for comparison. The first is traditional unsupervised clustering methods, including K-means [32] and the GMM [21]. These methods do not require labeled data and rely solely on intrinsic data distributions for interval estimation, serving as classical baselines for RI establishment. The second is mainstream semi-supervised learning methods, covering Semi-K-means [14], SSGMM [69], SSVAE [51], Semi-GAT [70], and Semi-GCN [71]. By leveraging a limited number of labeled samples to guide model learning, these methods provide a fair and direct comparison to our proposed semi-supervised framework.

Experiments were conducted under a unified experimental dataset and evaluation metrics. Quantitative performance results of RI prediction tasks for two key biomarkers, ALT and AST, are presented in Table 3. From the experimental data in Table 3, it can be observed that all comparative models demonstrate high consistency in predicting the head interval. This is because the biomarker

Table 3. Reference intervals of ALT and AST acquired by different indirect estimation techniques. Rank-1 results are in boldface and rank-2 results are underlined.

Age/Sex Partition	Methods	Biomarkers									
		ALT					AST				
		e_{lower}	BR_{LL}	e_{upper}	BR_{UL}	e_{overall}	e_{lower}	BR_{LL}	e_{upper}	BR_{UL}	e_{overall}
28 days to < 1 year	GMM	41.25	5.24	−30.56	−34.44	35.91	32.38	11.53	−12.63	−17.12	22.50
	K-means	26.25	3.33	29.44	33.17	27.84	28.10	10.00	24.00	32.54	26.05
	Semi-K-means	31.25	3.97	39.58	44.60	35.41	28.10	10.00	39.13	53.05	33.61
	SSGMM	31.25	3.97	41.27	46.51	36.26	28.10	10.00	40.13	54.41	34.11
	SSVAE	31.25	3.97	43.52	49.05	37.39	28.10	10.00	42.88	58.14	35.49
	Semi-GCN	31.25	3.97	41.27	46.51	36.26	28.10	10.00	40.13	54.41	34.11
	Semi-GAT	31.25	3.97	43.52	49.05	37.39	28.10	10.00	40.13	54.41	34.11
	SS-HGAE	33.75	4.29	**−7.61**	**−8.57**	**20.68**	30.95	11.02	**−0.25**	**−0.34**	**15.60**
1 to < 2 years	GMM	18.75	4.41	−31.90	−39.41	25.33	28.64	17.03	−7.29	−11.62	17.96
	K-means	11.25	2.65	33.33	41.18	22.29	20.00	11.89	24.24	38.65	22.12
	Semi-K-means	11.25	2.65	33.10	40.88	22.17	20.00	11.89	24.24	38.65	22.12
	SSGMM	11.25	2.65	25.24	31.18	18.24	23.18	13.78	18.31	29.19	20.74
	SSVAE	11.25	2.65	25.95	32.06	18.60	22.27	13.24	18.81	30.00	20.54
	Semi-GCN	11.25	2.65	25.95	32.06	18.60	23.18	13.78	18.81	30.00	21.00
	Semi-GAT	11.25	2.65	26.19	32.35	18.72	22.73	13.51	19.32	30.81	21.02
	SS-HGAE	12.50	2.94	**−21.67**	**−26.76**	**17.08**	23.64	14.05	**−1.69**	**−2.70**	**12.67**
2 to < 13 years	GMM	2.86	0.87	−29.67	−38.70	16.26	21.43	10.00	−5.00	−7.33	13.21
	K-means	1.43	0.43	67.33	87.83	34.38	17.14	8.00	22.95	33.67	20.05
	Semi-K-means	−1.43	−0.43	81.33	106.09	41.38	24.29	11.33	6.82	10.00	15.55
	SSGMM	2.86	0.87	68.00	88.70	35.43	20.00	9.33	17.05	25.00	18.52
	SSVAE	2.86	0.87	68.67	89.57	35.76	20.00	9.33	17.27	25.33	18.64
	Semi-GCN	2.86	0.87	68.67	89.57	35.76	20.00	9.33	17.27	25.33	18.64
	Semi-GAT	2.86	0.87	65.33	85.22	34.10	20.00	9.33	15.23	22.33	17.61
	SS-HGAE	2.86	0.87	**−5.33**	**−6.96**	**4.10**	22.86	10.67	**0.00**	**0.00**	**11.43**
13 to < 18 years male	GMM	−12.86	−2.50	−44.19	−52.78	28.52	10.83	5.20	−21.08	−31.20	15.96
	K-means	−8.57	−1.67	110.23	131.67	59.40	4.17	2.00	26.76	39.60	15.46
	Semi-K-means	−11.43	−2.22	110.23	131.67	60.83	6.67	3.20	34.59	51.20	20.63
	SSGMM	−10.00	−1.94	80.00	95.56	45.00	10.00	4.80	22.43	33.20	16.22
	SSVAE	−11.43	−2.22	66.28	79.17	38.85	10.00	4.80	19.46	28.80	14.73
	Semi-GCN	−10.00	−1.94	82.79	98.89	46.40	10.00	4.80	23.51	34.80	16.76
	Semi-GAT	−10.00	−1.94	82.79	98.89	46.40	10.00	4.80	23.51	34.80	16.76
	SS-HGAE	−12.86	−2.50	**−15.35**	**−18.33**	**14.10**	18.33	8.80	**−7.03**	**−10.40**	**12.68**
13 to < 18 years female	GMM	−5.00	−1.30	−18.28	−23.04	11.64	22.00	10.48	−17.10	−25.24	19.55
	K-means	−5.00	−1.30	56.55	71.30	30.78	20.00	9.52	19.68	29.05	19.84
	Semi-K-means	−5.00	−1.30	75.17	94.78	40.09	22.00	10.48	22.90	33.81	22.45
	SSGMM	−5.00	−1.30	66.90	84.35	35.95	21.00	10.00	22.90	33.81	21.95
	SSVAE	−5.00	−1.30	75.17	94.78	40.09	22.00	10.48	22.90	33.81	22.45
	Semi-GCN	−5.00	−1.30	75.17	94.78	40.09	21.00	10.00	24.19	35.71	22.60
	Semi-GAT	−5.00	−1.30	75.17	94.78	40.09	22.00	10.48	22.90	33.81	22.45
	SS-HGAE	−1.67	−0.43	**−12.07**	**−15.22**	**6.87**	18.00	8.57	**−16.45**	**−24.29**	**17.23**

concentrations corresponding to the head interval mostly fall within a physiologically stable range, where features are highly discriminative. All models can capture these patterns through basic statistical

analysis or feature learning, resulting in minimal differences in prediction results. Based on this observation, we focused the performance evaluation on the tail interval, which is more sensitive to the representational capability of the models. The tail interval corresponds to the high-concentration range of biomarkers and is directly associated with clinical abnormal value detection (e.g., an elevated upper bound of the ALT may indicate liver injury). Moreover, the data distribution in the tail interval is sparser and more affected by individual physiological differences than that in the head interval, making it more suitable for evaluating a model's ability to capture complex distributions. Therefore, a comprehensive evaluation was conducted using three core metrics: upper-bound error rate e_{upper}, upper-bound bias rate BR_{UL}, and overall error rate e_{overall}. These three metrics reflect two critical attributes of the model's performance: its accuracy in key intervals and the reliability of its predictions over the full-range interval.

Collectively, from the results of five metrics, the proposed SS-HGAE achieves the optimal performance in the reference interval prediction tasks for both ALT and AST across all datasets. In the ALT reference interval prediction task, compared with the second-best method corresponding to each subgroup, the SS-HGAE model reduces the overall error rate e_{overall} by 7.17, 1.16, 12.17, 14.42, and 4.77, respectively. Notably, the improvements of our proposal are particularly prominent in the two subgroups of "28 days to < 1 year" and "2 to < 13 years": the e_{upper} of the SS-HGAE drops to −7.61 and −5.33, respectively, while the BR_{UL} reaches −8.57 and −6.96. This result indicates that the SS-HGAE exhibits higher accuracy in interval prediction for biomarkers like ALT, which are significantly influenced by age and gender. In the AST reference interval prediction task, the performance advantage of the SS-HGAE is further highlighted. Compared with the second-best method in each subgroup, SS-HGAE reduces e_{overall} by 6.9, 5.3, 1.79, 3.28, and 2.32, respectively. Critically, in the three age groups of "28 days to < 1 year", "1 to < 2 years", and "2 to < 13 years", the upper-bound error rate between the predicted value of the AST reference interval by the SS-HGAE and the true reference interval value is less than 1.69, and the BR_{UL} is close to 2.7, being consistent with the true value. The AST values of these three age groups exhibit high variance and a wide physiological range (e.g., AST in infants is significantly affected by growth and development). So, the SS-HGAE still maintains high accuracy even on complex data distributions.

Within the scope of our current empirical evaluation on ALT and AST, the favorable performance of the SS-HGAE can be primarily attributed to two factors. First, unlike the Euclidean representations in traditional semi-supervised methods (e.g., Semi-GAT, SSVAE), the SS-HGAE employs hyperbolic embeddings to more accurately capture the similarity relationships between samples. The negative curvature property of hyperbolic space is compatible with the hierarchical distribution of the evaluated ALT and AST data, particularly when coupled with the supervised contrastive objective. Leveraging this geometry, the joint mechanism utilizes the exponential volume growth near boundaries to push apart distinct physiological states, yielding more reliable boundary estimations. Second, compared to unsupervised baselines that rely entirely on unlabeled data, the SS-HGAE uses a small fraction of labeled samples (< 8%) to efficiently locate the reference interval's core distribution. This mitigates the noise-induced interval deviations common in unsupervised methods. The SS-HGAE operates effectively under low label settings; however, excessive labels cause the model to over-collapse and degrade performance.

Table 4. Ablation experimental analysis on label information and hyperbolic space projection. The "√" mark indicates that the model includes the corresponding core component. The rank-1 results are in boldface.

Age/Sex Partition	Label	Hyperbolic Projection	Biomarkers									
			ALT					AST				
			e_{lower}	BR_{LL}	e_{upper}	BR_{UL}	e_{overall}	e_{lower}	BR_{LL}	e_{upper}	BR_{UL}	e_{overall}
28 days to < 1 year			36.25	4.60	−28.87	−32.54	32.56	31.90	11.36	−13.25	−17.97	22.58
	√		36.25	4.60	−21.55	−24.29	28.90	32.38	11.53	−12.63	−17.12	22.50
	√	√	33.75	4.29	**−7.61**	**−8.57**	**20.68**	30.95	11.02	**−0.25**	**−0.34**	**15.60**
1 to < 2 years			12.50	2.94	−29.76	−36.76	21.13	26.36	15.68	−10.00	−15.95	18.18
	√		12.50	2.94	−26.90	−33.24	19.70	29.09	17.30	−4.92	−7.84	17.00
	√	√	12.50	2.94	**−21.67**	**−26.76**	**17.08**	23.64	14.05	**−1.69**	**−2.70**	**12.67**
2 to < 13 years			1.43	0.43	−8.67	−11.30	5.05	20.71	9.67	−0.23	−0.33	**10.47**
	√		1.43	0.43	−22.33	−29.13	11.88	22.86	10.67	−4.77	−7.00	13.81
	√	√	2.86	0.87	**−5.33**	**−6.96**	**4.10**	22.86	10.67	**0.00**	**0.00**	11.43
13 to < 18 years male			−14.29	−2.78	−30.93	−36.94	22.61	14.17	6.80	−15.41	−22.80	14.79
	√		−15.71	−3.06	−30.47	−36.39	23.09	5.00	2.40	−13.51	−20.00	**9.26**
	√	√	−12.86	−2.50	**−15.35**	**−18.33**	**14.10**	18.33	8.80	**−7.03**	**−10.40**	12.68
13 to < 18 years female			−1.67	−0.43	−17.59	−22.17	9.63	17.00	8.10	−17.74	−26.19	17.37
	√		−5.00	−1.30	−17.59	−22.17	11.29	19.00	9.05	−16.45	−24.29	17.73
	√	√	−1.67	−0.43	**−12.07**	**−15.22**	**6.87**	18.00	8.57	**−16.45**	**−24.29**	**17.23**

Note: The hyperbolic projection module is designed for the supervised contrastive loss, which requires labeled data. Hence, the "no labels + hyperbolic projection" setting is not feasible in our framework.

4.6. Ablation study

To verify the necessity and gains of the core components of the SS-HGAE, we conducted an ablation study focusing on the independent contributions of the supervised information and hyperbolic space projection. The experiment strictly followed the design principle: two simplified models with missing functions were constructed to form a comparison with the original complete model, so as to quantitatively evaluate the mechanism of action of each component separately. This design can effectively eliminate interference between components, ensuring that performance differences only result from each component, and providing rigorous experimental evidence for the effectiveness of the components. The construction is as follows: The unsupervised baseline model (SS-HGAE w/o label) discards the annotations of the label set and the supervised losses, retaining only the learning framework based on unsupervised loss functions. This model is used to isolate and analyze the role of supervised information in capturing prior knowledge of data categories. The Euclidean-space baseline model (SS-HGAE w/o hyperbolic projection) discards the hyperbolic projection component, and all operations, including feature learning, distance calculation, and reference-interval prediction, are calculated in the traditional Euclidean space. This model is used to verify the representational advantages of hyperbolic space projection for the intrinsic structure of high-dimensional data. All models (the original SS-HGAE, SS-HGAE w/o label, and SS-HGAE w/o hyperbolic projection) adopt a unified experimental configuration, with a 20% subset of the labeled dataset used as the training

benchmark, core evaluation metrics including reference-interval prediction errors (including e_{upper}, e_{overall}, and BR_{UL}), and tests conducted on the full set of subgroup datasets (covering different age and gender groups) for the two biomarkers, ALT and AST.

Table 4 presents the quantitative performance results of the three types of models. By comparing the performance differences between the Euclidean space baseline model (SS-HGAE w/o hyperbolic projection) and the original complete model, the critical role of the hyperbolic space projection module can be clarified: when only relying on Euclidean space for loss constraint and feature learning, the model performance decays significantly, and this decay is particularly prominent in the tail interval, which is sensitive to distribution complexity. On a macro level, the complete model achieves the lowest overall error (e_{overall}) in 8 out of 10 experimental groups. More importantly, across all 10 groups without exception, it secures absolute optimal performance on both critical upper bound metrics (e_{upper} and BR_{UL}). Specifically, in the ALT reference interval prediction task, the upper-bound bias rate BR_{UL} of the Euclidean space baseline model decreased by 15.71, 6.47, 22.17, 18.06, and 6.96, respectively, compared with the original model, with an average reduction of 13.87. In the AST reference interval prediction task, the average reduction of BR_{UL} of the Euclidean space baseline model compared with the original model was 7.70. The above results fully confirm that the hyperbolic space projection module is not a simple spatial transformation operation. Instead, by adapting to the hierarchical distribution of biomarker data, it enhances the representation of complex structure, and particularly plays an irreplaceable role in improving the consistency and accuracy of tail interval prediction. By comparing the performance of the unsupervised baseline model (SS-HGAE w/o label) and the Euclidean space baseline model, the auxiliary role of supervised information can be further clarified. Although the unsupervised baseline occasionally outperforms the Euclidean baseline, the complete model with hyperbolic projection dominates in most cases. This demonstrates the complex hierarchical structure of the data and proves that supervised signals alone are insufficient, making the hyperbolic module indispensable. It should be noted that the role of supervised information lies primarily in ensuring performance stability, and that the hyperbolic space version still performs the best.

From the ablation study comparing the three model versions on the ALT and AST datasets, a core conclusion can be drawn: both supervised information and hyperbolic space projection contribute to the reference interval prediction, exhibiting functional complementarity. The hyperbolic space projection is responsible for accurately representing data structures, while the supervised module anchors the distribution boundary. Due to these two components, the SS-HGAE achieves optimal performance in this ablation study. This result verifies the rationality of the model architecture design: the synergistic combination of hyperbolic geometry and a strictly small amount of labeled data effectively enhances prediction performance, providing a feasible approach specifically for the evaluated pediatric ALT and AST biomarkers.

5. Discussion and limitations

This study has several limitations that warrant discussion. First, the role of the label ratio in this semi-supervised setting is nuanced and constrained. The labeled part helps to anchor the healthy cluster center, but it introduces a bias when the label ratio is high, as the labeled samples are only from healthy samples without pathological boundary constraints. Consequently, this method is strictly constrained to low-resource label settings (e.g., 10%–20%). Second, the current study serves only as a proof of

concept, the proposed method is currently applied to two functionally related liver enzymes (ALT and AST) across five age-sex groups. Due to the distinct groups of different pediatric biomarkers, expanding the evaluation requires a massive scale of independent modeling pipelines. Therefore, the framework's performance on other diverse clinical indicators remains to be further explored in future work.

6. Conclusions and future works

In this work, to alleviate the challenge of limited labeled clinical data, we propose a semi-supervised hyperbolic graph autoencoder (SS-HGAE) to establish pediatric reference intervals. By integrating supervised information guidance and hyperbolic space representation, the SS-HGAE enables accurate RI prediction with scarce labels on the ALT and AST biomarkers. Validation on pediatric clinical datasets for these two biomarkers implies the effectiveness of the model in this semi-supervised setting, with optimal performance at 10%–20% label utilization. This work offers a proof-of-concept for pediatric RI modeling.

Future work will focus on addressing these limitations by: (1) optimizing the feature fusion module to accommodate more biomarkers, and (2) exploring the details of the label ratio in one-class, labeled, semi-supervised settings.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are thankful for the financial support by the Ministry of Science and Technology of China under Grant 2022YFB2703300, the National Natural Science Foundation of China (U22B2048,62476274), and the Post-Doctoral Later-Stage Foundation Project of Shenzhen Polytechnic University (6024271013K).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. S. Poole, L. F. Schroeder, N. Shah, An unsupervised learning method to identify reference intervals from a clinical database, *J. Biomed. Inf.*, **59** (2016), 276–284. https://doi.org/10.1016/j.jbi.2015.12.010
2. M. Pusparum, G. Ertaylan, O. Thas, Individual reference intervals for personalised interpretation of clinical and metabolomics measurements, *J. Biomed. Inf.*, **131** (2022), 104111. https://doi.org/10.1016/j.jbi.2022.104111
3. A. Rao, A computer program for the derivation of non-parametric reference ranges from patients' results, *Comput. Biol. Med.*, **20** (1990), 331–336. https://doi.org/10.1016/0010-4825(90)90012-E

4. X. Ni, W. Song, X. Peng, Y. Shen, Y. Peng, Q. Li, et al., Pediatric reference intervals in china (prince): design and rationale for a large, multicenter collaborative cross-sectional study, *Sci. Bull.*, **63** (2018), 1626–1634. https://doi.org/10.1016/j.scib.2018.11.024
5. CLSI EP28-A3C, *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory*, 2010. Available from: https://clsi.org/shop/standards/ep28/.
6. J. Zierk, F. Arzideh, T. Rechenauer, R. Haeckel, W. Rascher, M. Metzler, et al., Age- and sex-specific dynamics in 22 hematologic and biochemical analytes from birth to adolescence, *Clin. Chem.*, **61** (2015), 964–973. https://doi.org/10.1373/clinchem.2015.239731
7. J. Zierk, F. Arzideh, R. Haeckel, H. Cario, M. C. Frühwald, H. J. Groß, et al., Pediatric reference intervals for alkaline phosphatase, *Clin. Chem. Lab. Med.*, **55** (2017), 102–110. https://doi.org/10.1515/cclm-2016-0318
8. Y. Weng, Y. Zhang, W. Wang, T. Dening, Semi-supervised information fusion for medical image analysis: Recent progress and future perspectives, *Inf. Fusion*, **106** (2024), 102263. https://doi.org/10.1016/j.inffus.2024.102263
9. Z. Xu, D. Lu, J. Luo, Y. Zheng, R. K. Tong, Separated collaborative learning for semi-supervised prostate segmentation with multi-site heterogeneous unlabeled mri data, *Med. Image Anal.*, **93** (2024), 103095. https://doi.org/10.1016/j.media.2024.103095
10. A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, I. Goodfellow, Realistic evaluation of deep semi-supervised learning algorithms, *Adv. Neural Inf. Process. Syst*, **31** (2018).
11. H. Lin, Y. Liu, D. Shi, X. Cheng, A contrastive model with local factor clustering for semi-supervised few-shot learning, *Mathematics*, **11** (2023), 3394. https://doi.org/10.3390/math11153394
12. B. Xiao, C. Lu, Semi-supervised medical image classification combined with unsupervised deep clustering, *Appl. Sci.*, **13** (2023), 5520. https://doi.org/10.3390/app13095520
13. A. Iscen, G. Tolias, Y. Avrithis, O. Chum, Label propagation for deep semi-supervised learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 5070–5079. https://doi.org/10.1109/CVPR.2019.00521
14. S. Basu, A. Banerjee, R. J. Mooney, Semi-supervised clustering by seeding, in *Proceedings of the Nineteenth International Conference on Machine Learning*, (2002), 27–34.
15. O. Ganea, G. Bécigneul, T. Hofmann, Hyperbolic neural networks, *Adv. Neural Inf. Process. Syst.*, **31** (2018).
16. M. S. Alzuhair, M. M. Ben Ismail, O. Bchir, Soft semi-supervised deep learning-based clustering, *Appl. Sci.*, **13** (2023), 9673. https://doi.org/10.3390/app13179673
17. V. Khrulkov, L. Mirvakhabova, E. Ustinova, I. Oseledets, V. Lempitsky, Hyperbolic image embeddings, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 6418–6428. https://doi.org/10.1109/CVPR42600.2020.00645
18. A. Ermolov, L. Mirvakhabova, V. Khrulkov, N. Sebe, I. Oseledets, Hyperbolic vision transformers: Combining improvements in metric learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 7409–7419. https://doi.org/10.1109/CVPR52688.2022.00726

19. T. Jiang, L. Chen, W. Chen, W. Meng, P. Qi, Reliamatch: Semi-supervised classification with reliable match, *Appl. Sci.*, **13** (2023), 8856. https://doi.org/10.3390/app13158856

20. S. Weber, B. Zöngür, N. Araslanov, D. Cremers, Flattening the parent bias: Hierarchical semantic segmentation in the poincaré ball, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2024), 28223–28232. https://doi.org/10.1109/CVPR52733.2024.02666

21. J. Zierk, F. Arzideh, L. A. Kapsner, H. U. Prokosch, M. Metzler, M. Rauh, Reference interval estimation from mixed distributions using truncation points and the kolmogorov-smirnov distance (kosmic), *Sci. Rep.*, **10** (2020), 1704. https://doi.org/10.1038/s41598-020-58749-2

22. R. G. Hoffmann, Statistics in the practice of medicine, *Jama*, **185** (1963), 864–873. https://doi.org/10.1001/jama.1963.03060110068020

23. C. Bhattacharya, A simple method of resolution of a distribution into Gaussian components, *Biometrics*, **1967** (1967), 115–135. https://doi.org/10.2307/2528285

24. F. Arzideh, W. Wosniok, R. Haeckel, Indirect reference intervals of plasma and serum thyrotropin (TSH) concentrations from intra-laboratory data bases from several German and Italian medical centres, *Clin. Chem. Lab. Med.*, **49** (2011), 659–664. https://doi.org/10.1515/CCLM.2011.114

25. W. Wosniok, R. Haeckel, A new indirect estimation of reference intervals: Truncated minimum chi-square (TMC) approach, *Clin. Chem. Lab. Med.*, **57** (2019), 1933–1947. https://doi.org/10.1515/cclm-2018-1341

26. T. Ammer, A. Schützenmeister, H. U. Prokosch, M. Rauh, C. M. Rank, J. Zierk, RefineR: A novel algorithm for reference interval estimation from real-world data, *Sci. Rep.*, **11** (2021), 16023. https://doi.org/10.1038/s41598-021-95301-2

27. T. Hepp, J. Zierk, M. Rauh, M. Metzler, S. Seitz, Mixture density networks for the indirect estimation of reference intervals, *BMC Bioinf.*, **23** (2022), 307. https://doi.org/10.1186/s12859-022-04846-0

28. R. Yan, K. Li, Y. Lv, Y. Peng, N. Van Halm-Lutterodt, W. Song, et al., Comparison of reference distributions acquired by direct and indirect sampling techniques: Exemplified with the Pediatric Reference Interval in China (PRINCE) study, *BMC Med. Res. Methodol.*, **22** (2022), 106. https://doi.org/10.1186/s12874-022-01596-8

29. R. Chen, Y. Tang, W. Zhang, W. Feng, Deep multi-view semi-supervised clustering with sample pairwise constraints, *Neurocomputing*, **500** (2022), 832–845. https://doi.org/10.1016/j.neucom.2022.05.091

30. R. Chen, Y. Tang, X. Cai, X. Yuan, W. Feng, W. Zhang, Graph structure aware contrastive multi-view clustering, *IEEE Trans. Big Data*, **10** (2023), 260–274. https://doi.org/10.1109/TBDATA.2023.3334674

31. F. Tian, B. Gao, Q. Cui, E. Chen, T. Y. Liu, Learning deep representations for graph clustering, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **28** (2014). https://doi.org/10.1609/aaai.v28i1.8916

32. M. Ahmed, R. Seraj, S. M. S. Islam, The k-means algorithm: A comprehensive survey and performance evaluation, *Electronics*, **9** (2020), 1295. https://doi.org/10.3390/electronics9081295

33. J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in *International Conference on Machine Learning*, (2016), 478–487.
34. D. Huang, D. H. Chen, X. Chen, C. D. Wang, J. H. Lai, Deepclue: Enhanced deep clustering via multi-layer ensembles in neural networks, *IEEE Trans. Emerg. Top. Comput. Intell.*, **8** (2024), 1582–1594. https://doi.org/10.1109/TETCI.2024.3353598
35. H. Ju, J. Guo, W. Ding, W. Pedrycz, X. Cheng, X. Yang, FDGC: Fuzzy deep clustering with dual-granularity contrastive learning, *Knowl. Based Syst.*, **2025** (2025), 114401. https://doi.org/10.1016/j.knosys.2025.114401
36. T. N. Kipf, M. Welling, Variational graph auto-encoders, preprint, arXiv:1611.07308. https://doi.org/10.48550/arXiv.1611.07308
37. J. Zheng, Y. Tang, X. Peng, J. Zhao, R. Chen, R. Yan, et al., Indirect estimation of pediatric reference interval via density graph deep embedded clustering, *Comput. Biol. Med.*, **169** (2024), 107852. https://doi.org/10.1016/j.compbiomed.2023.107852
38. R. Guan, W. Tu, S. Wang, J. Liu, D. Hu, C. Tang, et al., Structure-adaptive multi-view graph clustering for remote sensing data, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2025), 16933–16941. https://doi.org/10.1609/aaai.v39i16.33861
39. S. Wang, X. Liu, S. Liu, W. Tu, E. Zhu, Scalable and structural multi-view graph clustering with adaptive anchor fusion, *IEEE Trans. Image Process.*, **33** (2024), 4627–4639. https://doi.org/10.1109/TIP.2024.3444320
40. D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu, P. Cui, Structural deep clustering network, in *Proceedings of the Web Conference*, (2020), 1400–1410. https://doi.org/10.1145/3366423.3380214
41. W. Tu, S. Zhou, X. Liu, X. Guo, Z. Cai, E. Zhu, et al., Deep fusion clustering network, in *Proc AAAI Conf Artif Intell*, **35** (2021), 9978–9987. https://doi.org/10.1609/aaai.v35i11.17198
42. Y. Liu, W. Tu, S. Zhou, X. Liu, L. Song, X. Yang, et al., Deep graph clustering via dual correlation reduction, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (2022), 7603–7611. https://doi.org/10.1609/aaai.v36i7.20726
43. P. Zhao, X. Li, Y. Pan, I. W. Tsang, M. Wang, L. Liao, Sharpening deep graph clustering via diverse bellwethers, *Knowl. Based Syst.*, **317** (2025), 113322. https://doi.org/10.1016/j.knosys.2025.113322
44. Y. Ma, H. He, Z. Lei, K. Shi, X. Peng, Z. Niu, Deepedd: Nonparametric deep graph clustering through effective distance and density, *Neurocomputing*, **2025** (2025), 131359. https://doi.org/10.1016/j.neucom.2025.131359
45. I. Triguero, S. García, F. Herrera, Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study, *KIIS*, **42** (2015), 245–284. https://doi.org/10.1007/s10115-013-0706-y
46. Y. Wang, Y. Huang, Q. Wang, C. Zhao, Z. Zhang, J. Chen, Graph-based self-training for semi-supervised deep similarity learning, *Sensors*, **23** (2023), 3944. https://doi.org/10.3390/s23083944
47. A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, (1998), 92–100. https://doi.org/10.1145/279943.279962

48. T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, (2016), 785–794. https://doi.org/10.1145/2939672.2939785

49. D. Zhou, O. Bousquet, T. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, *Adv. Neural Inf. Process. Syst.*, **2003** (2003), 16.

50. O. Chapelle, A. Zien, Semi-supervised classification by low density separation, in *International Workshop on Artificial Intelligence and Statistics*, (2005), 57–64. https://doi.org/10.7551/mitpress/9780262033589.001.0001

51. D. P. Kingma, D. J. Rezende, S. Mohamed, M. Welling, Semi-supervised learning with deep generative models, *Adv. Neural Inf. Process. Syst.*, **2014** (2014), 27.

52. D. Guan, Y. Xing, J. Huang, A. Xiao, A. El Saddik, S. Lu, S2match: Self-paced sampling for data-limited semi-supervised learning, *Pattern Recognit.*, **159** (2025), 111121. https://doi.org/10.1016/j.patcog.2024.111121

53. W. Peng, T. Varanka, A. Mostafa, H. Shi , G. Zhao, Hyperbolic deep neural networks: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 10023–10044. https://doi.org/10.1109/TPAMI.2021.3136921

54. M. Nickel, D. Kiela, Poincaré embeddings for learning hierarchical representations, *Adv. Neural Inf. Process. Syst.*, **2017** (2017), 30.

55. M. Nickel, D. Kiela, Learning continuous hierarchies in the lorentz model of hyperbolic geometry, in *International Conference on Machine Learning*, (2018), 3779–3788.

56. C. Gulcehre, M. Denil, M. Malinowski, A. Razavi, R. Pascanu, K. M. Hermann, et al., Hyperbolic attention networks, preprint, arXiv:1805.09786. https://doi.org/10.48550/arXiv.1805.09786

57. R. Shimizu, Y. Mukuta, T. Harada, Hyperbolic neural networks++, preprint, arXiv:2006.08210. https://doi.org/10.48550/arXiv.2006.08210

58. Q. Ma, M. Yang, M. Ju, T. Zhao, N. Shah, R. Ying, Breaking information cocoons: A hyperbolic graph-llm framework for exploration and exploitation in recommender systems, preprint, arXiv:2411.13865. https://doi.org/10.48550/arXiv.2411.13865

59. Y. Liu, Z. He, K. Han, Hyperbolic category discovery, in *Proceedings of the Computer Vision and Pattern Recognition Conference*, (2025), 9891–9900. https://doi.org/10.1109/CVPR52734.2025.00924

60. National Health Commission of the People's Republic of China, *Reference Intervals of Clinical Biochemistry Tests Commonly used for Children (WS/T 780-2021)*, 2021. Available from: https://www.nhc.gov.cn/wjw/s9492/202105/3d5159ef7619452b9842ea9520189a11.shtml.

61. S. Sun, R. Huang, An adaptive k-nearest neighbor algorithm, in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, (2010), 91–94. https://doi.org/10.1109/FSKD.2010.5569740

62. S. Rastegar, H. Doughty, C. Snoek, Learn to categorize or categorize to learn? self-coding for generalized category discovery, *Adv. Neural Inf. Process. Syst.*, **36** (2023), 72794–72818. https://doi.org/10.52202/075280-3183

63. J. W. Cannon, W. J. Floyd, R. Kenyon, W. R. Parry, Hyperbolic geometry, *Flavors Geom.*, **31** (1997), 2.

64. A. Ungar, *A Gyrovector Space Approach to Hyperbolic Geometry*, Springer Nature, 2022.

65. L. Franco, P. Mandica, B. Munjal, F. Galasso, Hyperbolic self-paced learning for self-supervised skeleton-based action representations, preprint, arXiv:2303.06242. https://doi.org/10.48550/arXiv.2303.06242

66. S. Rastegar, M. Salehi, Y. M. Asano, H. Doughty, C. G. Snoek, Selex: Self-expertise in fine-grained generalized category discovery, in *European Conference on Computer Vision*, (2024), 440–458. https://doi.org/10.1007/978-3-031-72897-6_25

67. R. Z. Tan, C. Markus, S. Vasikaran, T. P. Loh, Comparison of 8 methods for univariate statistical exclusion of pathological subpopulations for indirect reference intervals and biological variation studies, *Clin. Biochem.*, **103** (2022), 16–24. https://doi.org/10.1016/j.clinbiochem.2022.02.006

68. G. J. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, 2008. https://doi.org/10.1002/9780470191613

69. W. Cao, R. Wang, M. Fan, X. Fu, H. Wang, Y. Wang, A new froth image classification method based on the MRMR-SSGMM hybrid model for recognition of reagent dosage condition in the coal flotation process, *Appl. Intell.*, **52** (2022), 732–752. https://doi.org/10.1007/s10489-021-02328-z

70. F. Yang, H. Zhang, S. Tao, Semi-supervised classification via full-graph attention neural networks, *Neurocomputing*, **476** (2022), 63–74. https://doi.org/10.1016/j.neucom.2021.12.077

71. B. Jiang, Z. Zhang, D. Lin, J. Tang, B. Luo, Semi-supervised learning with graph learning-convolutional networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 11313–11320. https://doi.org/10.1109/CVPR.2019.01157

72. Y. Ozarda, K. Ichihara, G. Jones, T. Streichert, R. Ahmadian, Comparison of reference intervals derived by direct and indirect methods based on compatible datasets obtained in Turkey, *Clin. Chim. Acta*, **520** (2021), 186–195. https://doi.org/10.1016/j.cca.2021.05.030

AIMS Press