



Predictive modeling of pediatric drug-induced liver injury: Dynamic classifier selection with clustering analysis

Zixin Shi, Linjun Huang  and Haolin Wang 

Abstract

Background: Pediatric populations are more vulnerable to drug-induced liver injury (DILI) due to distinct pharmacokinetic profiles and ongoing physiological maturation processes. However, early identification and assessment of DILI in pediatric patients present significant clinical challenges, primarily due to the inherent complexity of pediatric cases and substantial limitations in available clinical data.

Objective: This study introduces a framework that integrates clustering analysis with dynamic classifier selection (DCS) techniques to enhance pediatric DILI prediction. The proposed method addresses challenges such as patient heterogeneity and class imbalance, while optimizing predictive performance to support clinical decision-making.

Methods: We investigated a retrospective cohort of 12,555 pediatric inpatients across six hospitals in Chongqing, China. The dataset encompassed a wide range of biomedical parameters, including laboratory results and liver function profiles, along with clinical documentation spanning demographic characteristics, medical histories, and medication regimens. Patients were stratified into four distinct clinical subgroups based on silhouette coefficient. A diverse pool of base classifiers was generated with varied initialization strategies and hyperparameter optimizations tailored to each patient cluster. The classification process was further refined through the implementation of Dynamic Classifier Selection with Multiple Classifier Behavior (DCS-MCB) methodology, which adaptively customizes model selection based on the distinctive clinical profiles of each subgroup.

Results: The Clustering-enhanced DCS-MCB framework demonstrated superior performance compared to conventional machine learning models across evaluation metrics. The ensemble learning models consistently outperformed individual classifier models, with the presented study achieving the highest F1-score (0.926), MCC (0.917), G-mean (0.959), demonstrating the strength of this hybrid approach in addressing the complexities of pediatric DILI prediction.

Conclusion: The integration of clustering analysis with dynamic classifier selection has demonstrated efficacy in complex real-world clinical settings. This methodology provides a more robust, precise, and clinically adaptable framework for patient stratification and drug safety surveillance.

Keywords

Pediatric drug-induced liver injury, dynamic classifier selection, predictive modeling, patient stratification, drug safety

Received: 3 October 2024; accepted: 3 March 2025

Introduction

Adverse drug reactions (ADRs) in children are receiving increasing attention. ADRs are defined as “a noticeable harmful or unpleasant reaction resulting from an intervention related to the use of a medicinal product”.^{1,2} Due to the ongoing maturation of children’s physiological systems, such as in aspects of drug absorption, metabolism,

College of Medical Informatics, Chongqing Medical University, Chongqing, China

Corresponding author:

Haolin Wang, College of Medical Informatics, Chongqing Medical University, Chongqing, China.

Email: haolinwang@cqmu.edu.cn



Creative Commons NonCommercial-NoDerivs CC BY-NC-ND: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits any use, reproduction and distribution of the work as published without adaptation or alteration, provided the original work is attributed as specified on the SAGE and Open Access page (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

transport, elimination, as well as the use of off-label and unlicensed drugs, they are more susceptible to specific types of ADRs,^{3,4} with an incidence rate 2–4 times higher than that in adults. Adverse reactions typically signal the need for caution in future administration, requiring prevention, special treatment, adjustment of the dosage regimen, or discontinuation of the drug. According to a study based on a pediatric active monitoring system, approximately 15 children per 1000 experience adverse reactions.⁵ The majority of observed ADRs primarily affect the skin (e.g., rash, urticaria) and the gastrointestinal system (e.g., diarrhea, nausea, vomiting).^{6,7}

Drug-induced liver injury (DILI) is a rare but severe ADR, resulting from an adverse reaction to drugs or other exogenous agents,⁸ which can progress to acute liver failure (ALF). DILI can be categorized into two types: intrinsic and idiosyncratic DILI. Intrinsic DILI is dose-dependent, meaning the risk and severity of liver damage increase with higher drug doses, and it is more predictable as it is closely tied to the drug's pharmacological properties and mechanistic pathways. In contrast, idiosyncratic DILI is dose-independent, occurring unpredictably in only a small proportion of patients regardless of dosage, administration route, or treatment duration, and its complex causes involve genetic, immune-mediated, and metabolic factors unique to the individual. DILI is a common cause of pediatric liver disease.⁹ According to data from the U.S. Acute Liver Failure Study Group, DILI accounts for more than 50% of acute liver failure cases.¹⁰ DILI is a complex process driven by multifactorial etiologies and the combined effects of polypharmacy, with diverse underlying causes of adverse reactions among different pediatric patients. Providing clinical decision support for pediatric DILI presents a significant challenge. A pivotal concern lies in identifying the most appropriate treatment regimen to aid in the prognosis of DILI, considering the vast array of complex therapeutic alternatives. This can be effectively resolved through a comprehensive understanding of each patient's disease characteristics and medication profiles.

ADRs are preventable,¹¹ and the application of machine learning shows promise in assisting in the diagnosis of DILI. For example, Tracy L. Sandritter¹² employed an Electronic Health Records (EHRs)-based screening tool to identify potential pediatric DILI cases, identifying 12 patients over two years as possible or probable cases of DILI, and emphasized the need for future improvements to better identify and define DILI. In the context of predicting ADRs events, Ze Yu et al.¹³ developed predictive models using seven machine learning algorithms on a cohort of 1746 patients, with the best performance achieved by gradient boosting decision trees (GBDT), yielding a precision of 44%, recall of 25%, and F1 score of 38.88%.

However, conducting research on DILI data encounters substantial challenges, primarily stemming from the following factors: (1) The intricate interactions among different

medications, the prevalence of polypharmacy, and the impact of diverse underlying diseases such as sepsis and leukemia demand the application of sophisticated data analysis methods. (2) Data-related issues, like incompleteness, sparsity and imbalanced class distributions, pose significant difficulties in the analysis of real-world data. (3) There are also confounding elements involved, such as variability in clinical practices and the underreporting of adverse events.¹⁴ To address these challenges effectively, we proposed a framework that integrates clustering and DCS techniques to facilitate accurate and personalized clinical decision making.

Clustering algorithms play a crucial role in profiling patient subgroups that are associated with a wide array of suspected triggering factors related to different diseases and medications. These algorithms operate proactively by partitioning data into distinct, non-overlapping clusters based on feature similarity, ensuring that the patterns within each cluster are highly similar while remaining distinct from those in other clusters.¹⁵ The clustering problem has been extensively studied in diverse fields, including statistics and artificial intelligence. For example, Qianqian Yu¹⁶ proposed a novel clustering-based method for predicting potential associations between lncRNA and diseases. This method achieved an AUC exceeding 0.8, outperforming three other comparative methods. Similarly, Utkarsh Agrawal et al.¹⁷ introduced a new ensemble classification phase that followed the ensemble clustering stage, which improved the final classification outcomes for unclustered data. Peng Gaong Sun et al.¹⁸ used clustering algorithms, Markov clustering algorithms, and molecular complex detection to decompose the human PPI network into dense clusters. In the medical domain, identifying biological clusters such as patient subgroups with distinct characteristics, has been demonstrated to be both feasible and of great significance for optimizing treatment strategies. Therefore, developing frameworks that assign patients into groups can not only enhance the accuracy of classification but also enable researchers to gain deeper insights into the intricate relationships within the patient data.

Single classifiers often fall short when it comes to solving intricate classification problems. In contrast, multi-classifier systems incorporating DCS^{19,20} present a promising solution to common challenges like data incompleteness and imbalanced class distributions that are frequently encountered in EHRs data. These systems are capable of effectively handling such complexities by integrating a variety of feature selection and undersampling techniques directly on the original dataset. DCS enables the identification and selection of the most capable classifiers for each query instance. Notable examples include methods such as META-DES,²¹ multiple classifier behavior (MCB), and modified local accuracy (MLA). Typically, the process commences with an estimation of the classifiers' competitiveness, grounded in the local region of the feature space

in which the query samples are located. Subsequently, in accordance with specific selection criteria, the competence level of the base classifiers, such as accuracy or ranking, is evaluated by leveraging the samples within this local region. Finally, only the combination of classifiers that meet a certain competence level, or the single best-performing classifier, is selected. DCS has also demonstrated its efficacy in adeptly handling data characterized by imbalanced class distributions.²²

The integration of clustering techniques and DCS methods offers a hybrid strategy that capitalizes on the advantages of both to enhance predictive accuracy. Clustering categorizes patients into subgroups based on shared characteristics, allowing for a more tailored and precise application of classifiers. For instance, Wu²³ introduced a hybrid multi-clustering and bagged classifier generation (HMCBCG) method, achieving impressive results with an accuracy of 99.81%, an F1-score of 99.86%, and a G-mean of 99.78%. Such evidence underscores the potential of combining clustering and DCS techniques to improve classification performance, especially when dealing with complex and imbalanced datasets like those prevalent in medical research. Building upon this existing knowledge, our research adopts a hybrid approach specifically designed for predicting. Initially, patients are clustered into distinct groups based on clinical features such as underlying diseases and laboratory test results. Subsequently, DCS methods are employed to select the most competent classifiers for each subgroup, tailoring the classification process to the specific characteristics of each cluster. This approach not only enhances predictive performance but also increases the adaptability and robustness of the model. By tackling the heterogeneity within patient populations, this method facilitates personalized clinical decision support, advances predictive modeling for pediatric DILI, and demonstrates its potential to improve predictive accuracy and patient outcomes in real-world clinical settings.

Methods

Data collection

This study is a retrospective prognostic investigation utilizing data from pediatric inpatient records at six tertiary hospitals in Chongqing, China. According to the World Health Organization, childhood encompasses ages 0–10 years, while adolescence includes ages 10–19 years. Therefore, this study covers both children and adolescents within the study population. The initial cohort of this study consists of all pediatric patients hospitalized between January 1, 2013, and December 31, 2023, who were diagnosed with DILI during their hospitalization. First, we conducted screening using ICD-10 codes to identify potentially eligible cases. Subsequently, the research team invited clinicians to participate in the verification process to ensure

that the diagnosis met the criteria for DILI. The data we used were the first data records of patients upon admission, at which time the patients had not been diagnosed with DILI.

Exclusion criteria were as follows: (1) Absence of drug exposure: Patients who had not been exposed to drugs with a known potential to cause liver injury were excluded. (2) Incomplete or missing clinical data: Individuals with incomplete clinical data or a lack of essential medical information required for diagnosis were not included. (3) Unclear etiology: Patients in whom the cause of liver injury could not be confidently linked to either specific drug exposure or underlying diseases were excluded. This category included cases where liver injury was caused by common etiologies such as viral hepatitis or genetic liver disorders. This study included 1190 pediatric cases that were confirmed to have DILI. Additionally, a control cohort was formed through randomly sampling from the cases that were confirmed not to have DILI, excluding those with incomplete or missing clinical data. In the end, data from 12,555 patients were included in this study.

Statistical analysis

In the descriptive analyses, laboratory indices and other continuous variables were characterized using means and interquartile ranges (IQRs), while categorical data were presented as counts and percentages. A two-step variable selection approach was implemented: (1) Univariate screening: Continuous predictors were evaluated using the Mann-Whitney *U* test, appropriate for data with non-normal distributions, while binary variables, such as gender and medication usage, were assessed through the χ^2 test or Fisher's exact test, depending on the sample size. Variables demonstrating statistical significance (p -value < 0.05) were shortlisted for subsequent analysis, considering the dataset's high dimensionality. (2) Automated refinement: The variables selected from the univariate analysis underwent an automated selection procedure within the DCS framework. This step identified the most relevant predictors, balancing model complexity with predictive accuracy, as detailed in the Results section.

Specialized and standardized case report forms were designed for all cases, which captured a range of factors associated with pediatric DILI. These factors included are as follows: (1) Demographic information (e.g., gender, age); (2) Medication information (e.g., antibiotics, acetaminophen, non-steroidal anti-inflammatory drugs, drug counts); (3) Disease information (e.g., pneumonia, sepsis, leukemia); (4) Chief complaints (e.g., fever, cough, vomiting); (5) Hospital examination results (e.g., alanine transaminase (ALT), aspartate transaminase (AST), alkaline phosphatase (ALP), gamma-Glutamyl transferase (GGT), international normalized ratio (INR), total bilirubin (TBIL), direct bilirubin (DBIL), indirect bilirubin (IBIL));

Table 1. Demographic and clinical profile of patients.

Characteristic		label = 0 (n = 11,373) No.(%)/median (IQR)	label = 1 (n = 1190)``No.(%) /median (IQR)	P value
Demographic	Female	4391 (38.61)	440 (36.97)	0.267
	Male	6982 (61.39)	750 (63.03)	
	Age	5.57 (2.00–9.00)	6.01 (2.00–9.00)	0.001
Medication	Antibiotics	8513 (74.85)	560 (47.06)	<0.001
	Acetaminophen	219 (1.93)	14 (1.18)	0.027
	NSAIDs	2395 (21.06)	224 (18.82)	0.062
	Isoniazid	54 (0.47)	13 (1.09)	0.043
	Ribavirin	2531 (22.25)	221 (18.57)	0.002
	Cytarabine	93 (0.82)	163 (13.70)	<0.001
	Methotrexate	84 (0.74)	113 (9.50)	<0.001
	Drug counts	24.22 (16.00–29.00)	41.18 (31.00–49.00)	<0.001
Diagnosis	Pneumonia	4043 (35.55)	490 (41.18)	<0.001
	Septicaemia	824 (7.25)	567 (47.65)	<0.001
	Leukaemia	470 (4.13)	675 (56.72)	<0.001
	Radiotherapy	516 (4.54)	770 (64.71)	<0.001
	Coagulopathy	615 (5.41)	275 (23.11)	<0.001
	Ulcers	217 (1.91)	129 (10.84)	<0.001
	Illness counts	4.26 (2.00–6.00)	8.18 (6.00–10.00)	<0.001
Self-reported Symptoms	Fever	2416 (21.24)	327 (27.48)	<0.001
	Cough	2063 (18.14)	254 (21.34)	0.009
	Mental deficiency	1751 (15.40)	168 (14.12)	0.231
	Vomiting	1521 (13.37)	85 (7.14)	<0.001
	Soreness	2259 (19.86)	205 (17.23)	0.023
	Loss of appetite	1848 (16.25)	173 (14.54)	0.114
	Jaundice	125 (1.10)	41 (3.45)	<0.001
	Yellowing of the skin	177 (1.56)	38 (3.19)	0.002
Others	Hospitalization counts	1.37 (1.0–1.0)	3.86 (1.00–6.00)	<0.001
	Hospital Days Count	12.76 (6.00–14.00)	28.25 (13.00–37.00)	<0.001

(continued)

Table 1. Continued.

Characteristic		label = 0 (<i>n</i> = 11,373) No.(%)/median (IQR)	label = 1 (<i>n</i> = 1190) No.(%) /median (IQR)	<i>P</i> value
Laboratory Tests	ALT	29.98 (14.00–30.00)	169.43 (81.25–176.05)	<0.001
	AST	56.11 (24.40–63.58)	181.62 (128.90–187.80)	<0.001
	AST/ALT	1.78 (1.17–2.18)	1.34 (0.64–1.70)	<0.001
	ALP	188.01 (132.5–220.1)	190.28 (114.90–204.10)	0.667
	GGT	40.55 (10.4–45.6)	71.84 (15.22–78.75)	<0.001
	TBIL	21.59 (5.0–22.6)	24.63 (6.20–25.70)	0.107
	DBIL	3.26 (0–3.90)	11.81 (1.10–15.30)	<0.001
	IBIL	18.34 (13.41–19.70)	10.85 (4.20–10.88)	<0.001
	INR	1.09 (0.97–1.12)	1.12 (0.96–2.10)	0.035
	Serum Albumin	0.049 (0–1)	0.036 (0–1)	0.027

and (6) Other information (e.g., number of hospitalizations, length of hospital stay). Table 1 displays the demographic and clinical characteristics of the patients. Specifically, label = 0 represents patients who were not diagnosed with DILI, and label = 1 represents patients diagnosed with DILI.

The proposed framework

Figure 1 depicts the overall implementation pathway followed in this study. Step A represents the clustering phase, during which the pediatric DILI data are partitioned into distinct subgroups using the silhouette coefficient and k-means algorithm, as detailed in the Methods' clustering section. Subgrouping considers several factors including diseases (e.g., sepsis, leukemia), laboratory indices (e.g., ALT, TBIL), and drug treatment regimens (e.g., antibiotics, acetaminophen, NSAIDs). Step B represents the classifier pool generation stage. Various strategies are used such as different cluster initialization configurations, heterogeneous models with hyperparameter optimization, and data sampling. These strategies are implemented to address the data-related challenges including imbalanced class distribution, creating a pool of base classifiers that embrace multiple optimal solutions. The final stage is the dynamic selection phase. Classifying a new query sample usually has three key steps. First, define the competence region using methods like KNN, decision space, or the potential function model. Second, set selection criteria to assess base classifiers' performance, considering metrics such as data complexity, model ranking, and diversity. Finally,

determine the method for selecting base classifiers, which could be choosing a single classifier, dynamic classifier selection, or dynamic ensemble selection (DES). By assigning different test samples to their corresponding classifiers, we can effectively carry out the decision-making process.

Clustering. K-means is a classical unsupervised learning algorithm employed to partition a dataset into multiple clusters, grouping similar data points around a cluster center. The algorithmic workflow adopted in this study is as follows:

1. The K-means clustering algorithm was applied to perform clustering and stratification analysis on patients with similar characteristics, with the number of clusters (*K*) ranging from 2 to 10. For each patient's data, the distances to all cluster centers were computed, and the patient was assigned to the closest cluster. This approach ensured that each patient's data was allocated to the cluster with the minimum distance to its center. The formula used for this calculation is as follows:

$$\text{dis}(X_i, C_j) = \sqrt{\sum_{t=1}^m (X_{it} - C_{jt})^2}$$

In the formula above, X_i represents the i -th object where $1 \leq i \leq n$, C_j represents the j -th cluster center where $1 \leq j \leq k$, X_{it} denotes the t -th attribute of the i -th object where $1 \leq$

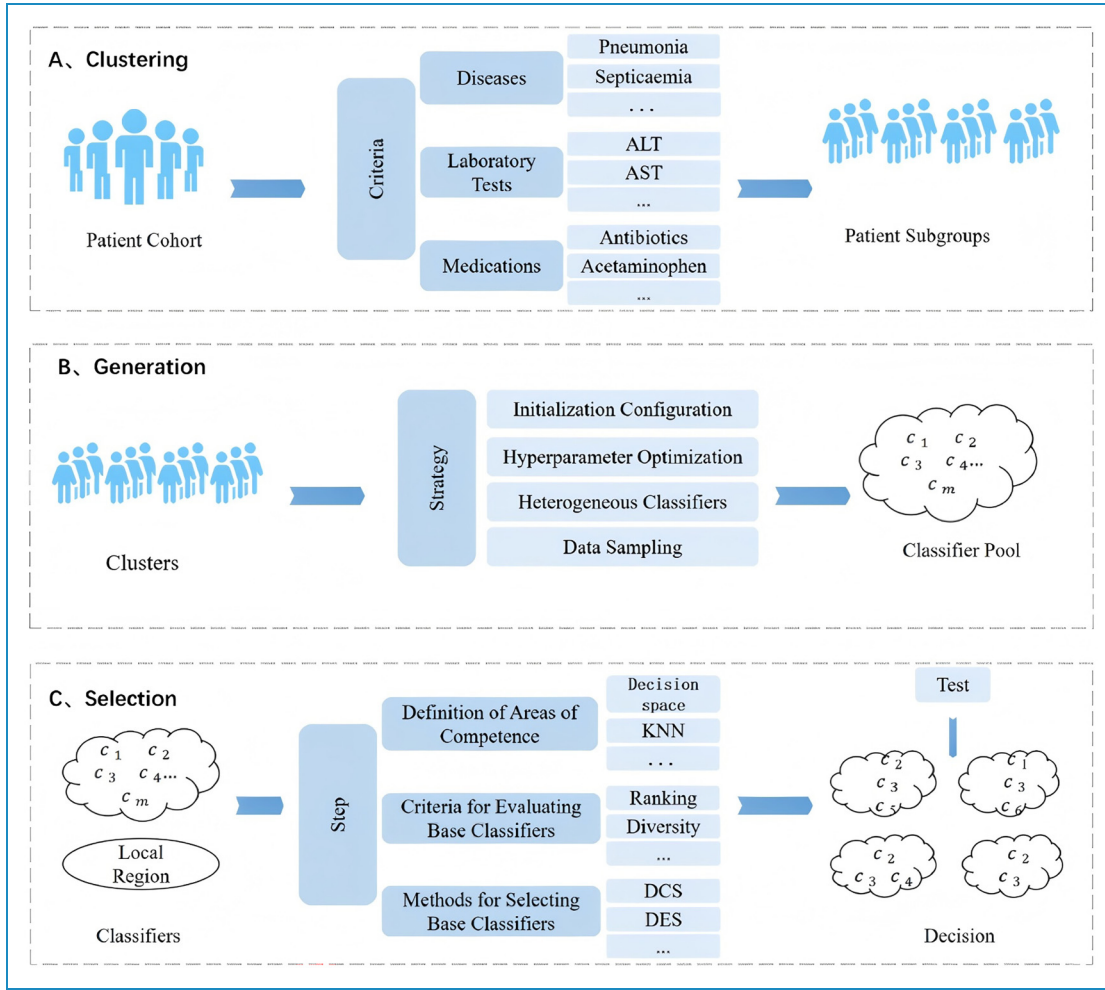


Figure 1. Experimental path diagram for this study.

$t \leq m$, and C_{jt} denotes the t -th attribute of the j -th cluster center.

(2) The optimal number of clusters K (ranging from 2 to 10) was determined according to the silhouette coefficient, a commonly used metric to evaluate the quality of clustering results. This coefficient measures both the compactness and separation of clusters. The silhouette coefficient $s(i)$ for each sample i is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance between sample i and all other samples within the same cluster (intra-cluster cohesion), $b(i)$ is the average distance between sample i and all samples in the nearest neighboring cluster (inter-cluster separation).

In this study, the silhouette coefficient, a metric that quantifies both intra-cluster cohesion and inter-cluster

separation, was used to evaluate the clustering quality. The cluster number corresponding to the highest silhouette coefficient was regarded as the optimal K , representing the most balanced clustering result. Once the optimal K was identified, the data were grouped into K clusters, thus laying the foundation for subsequent analyses.

Dynamic classifier selection. Under the assumption that data from different groups may exhibit distinct patterns and relationships, an appropriate classifier for instances from each group is selected in this phase. Class imbalance poses a significant challenge, potentially leading to biased classification results that prioritize the majority class. DCS has demonstrated its effectiveness in dealing with imbalanced data.

Classifier Pool Generation: The goal is to create a pool C containing m base classifiers that are both accurate and diverse. These base classifiers need to include a variety of models capable of achieving optimal results on the training data. The base classifier pool is generated by using various

initializations (clustering the data into diverse groups) and setting different parameters (each classifier is configured with distinct hyperparameters tailored to the data it is trained on). This approach ensures that the resulting classifier pool can generate informed predictions across a wide range of datasets, thereby improving overall classification performance.

DCS based on Multiple Classifier Behavior (DCS-MCB): This is a dynamic selection technique based on the behavior knowledge space (BSK) and classifier local accuracy.¹⁹ Given a new test sample x_j , its competent region θ_j is estimated. Subsequently, the BSK algorithm is used to compute the output profiles of the test sample and those of the corresponding region. The similarity $S(\tilde{x}_j, \tilde{x}_k)$ between the output profiles of the test sample \tilde{x}_j and the output profiles of its competent region $\tilde{x}_k \in \theta_j$ is computed using the following formula.

$$S(\tilde{x}_j, \tilde{x}_k) = \frac{1}{M} \sum_{i=1}^M T(x_j, x_k)$$

where

$$T(x_j, x_k) = \begin{cases} 1 & \text{if } c_i(x_j) = c_i(x_k), \\ 0 & \text{if } c_i(x_j) \neq c_i(x_k). \end{cases}$$

Samples with a similarity below the specified threshold are excluded from the competence region θ_j . As a result, the size of the competence region varies, depending on the similarity between the query sample and the samples within its competence region. After all similar samples have been selected, the competence of each base classifier is estimated by its classification accuracy within the defined competence region. If a classifier notably outperforms others in the pool, with the competence level surpassing a predefined threshold, it is used to classify the test sample. Otherwise, a majority voting rule will be applied, integrating all classifiers in the pool. The flowchart of the framework is presented in Figure 2.

In our study, separate base classifier pools are constructed for distinct data clusters. The DCS-MCB method is employed to ensure accurate classification of instances within each cluster. This approach allows for a thorough assessment of the base classifiers' performance across their full range of capabilities. To further validate the model's robustness and reliability, the data undergoes multiple rounds of cross-validation. This iterative validation not only reinforces the model's stability but also boosts its overall predictive performance, ensuring it can be effectively applied in various clinical settings.

In summary, we utilized a framework integrating clustering and dynamic selection techniques to enhance predictive accuracy in pediatric DILI. First, we identified the optimal number of clusters using the silhouette coefficient and applied the K-means algorithm to cluster the DILI data. Subsequently, we selected base classifiers with performance

surpassing a predefined threshold to create a classifier pool. In the final dynamic selection stage, we applied DCS-MCB to tailor classifier selection for each query sample, ensuring accurate and reliable predictions. The model's stability was validated through multiple rounds of cross-validation, rendering it appropriate for a wide range of clinical uses.

Results

This study presents a framework integrating clustering and dynamic classifier selection (DCS-MCB) to enhance the accuracy and robustness of prediction models for DILI in children. The results demonstrate that the proposed model outperforms traditional models, demonstrating its potential for use in clinical settings. It can assist in the early identification of high-risk pediatric DILI patients, enabling precise interventions, and improve the overall efficiency of health-care services.

Data processing

Figure 3 depicts the data processing workflow in this study. First, the data was divided into a training set and a test set. Subsequently, the training set was clustered into four clusters (c_0 to c_3) using silhouette coefficients and k-means clustering criteria. The clustered training sets were further split into training subsets (c0_train to c3_train) and validation subsets (c0_val to c3_val). According to the characteristic disparities among each cluster, four distinct base classifier pools were constructed (the specific criteria are detailed in the "Dynamic Classifier Selection" section of the Results). Each pool might contain base classifiers with diverse or similar characteristics. For instance, if the evaluation metrics of the KNN algorithm in cluster_0 all exceed 0.8, it is incorporated into pool_0. On the contrary, if its F1 score in cluster_1 is less than 0.8, it is excluded from pool_1. Eventually, four distinct base classifier pools were formed, and the DCS-MCB algorithm was employed to integrate these pools.

Notably, we left the test set unprocessed, maintaining its data independence. When the DCS-MCB algorithm predicts the test set, it first identifies the cluster (c_0 to c_3) that each data point belongs to. Then, it chooses the corresponding base-classifier pool (pool_0 to pool_3) for prediction. This data-processing approach ensures that the test set utilizes the most suitable base-classifier pool for prediction. It sets this method apart from traditional machine-learning models, leading to more accurate predictions.

Clustering analysis

As described in the Methods section, we first utilized the silhouette coefficient to determine the optimal value of K within the range of 2–10. Subsequently, we applied the k-means algorithm to cluster the data into K groups.

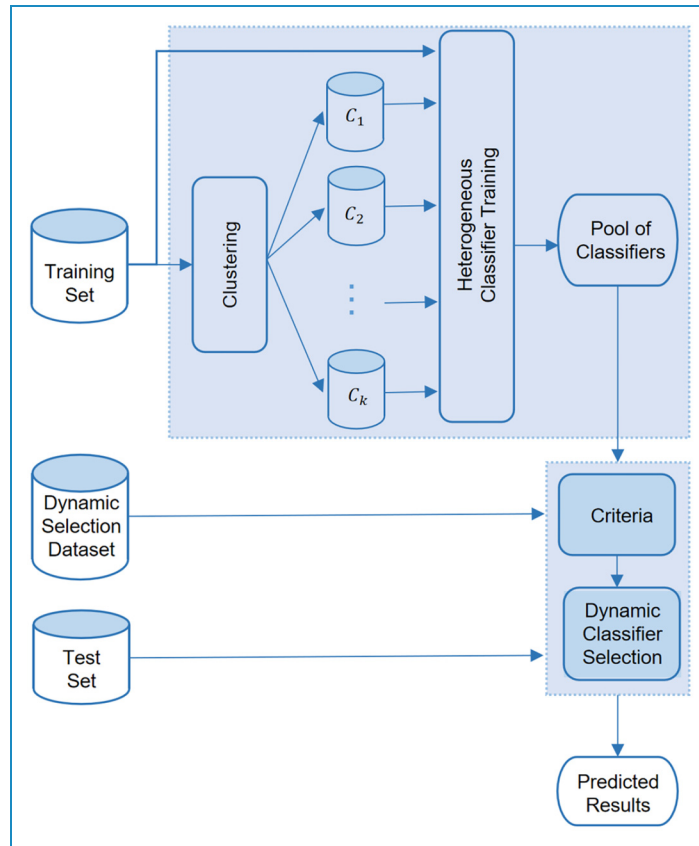


Figure 2. The integration of clustering and DCS framework flowchart.

Figure 4 presents the silhouette coefficient plot of this study, with the x-axis representing the number of clusters and the y-axis representing the silhouette score. A higher silhouette score implies a better clustering result. From this evaluation, we concluded that four clusters resulted in the optimal outcome within the potential range of clusters. Consequently, patients were divided into four distinct subgroups, discriminating those with complex underlying diseases from those with milder conditions. This stratification enables a more accurate analysis of the DILI incidence across different populations.

Subsequently, the k-means algorithm was applied to cluster the data into four groups. Table 2 presents the composition of each cluster, with the ratio of 0:1 for each category shown in parentheses. A detailed examination of the data clusters indicated that complex diseases were more prevalent in cluster_0 and cluster_2. For example, in the training set, 50.84% of patients in cluster_2 had a sepsis diagnosis, compared to just 1.43% in cluster_1. The differences between cluster_1 and cluster_3 were mainly driven by variations in laboratory indicators and medication use. For instance, the proportion of patients using ambroxol in cluster_0 was 21.59%, but in cluster_3, this figure increased significantly to 53.31%.

Figure 5 showcases a heatmap designed to visualize the normalized features of patients. The four distinct clusters,

namely cluster_0 through cluster_3, are clearly demarcated by prominent green dashed lines. Along the x-axis, the features are listed, while the y-axis represents the patients. The color-coded scale serves as an intuitive guide for the feature values, with darker hues indicating higher values. This heatmap offers a comprehensive view of the distribution patterns of features across diverse patient clusters. It reveals information about how features are distributed among the clusters. Each cluster exhibits unique feature prominence, accompanied by block-wise missing data. Notably, traditional statistical and machine learning techniques are found to be inadequate for effectively managing such missing data patterns. The clustering patterns unveiled in this heatmap underscore the inherent heterogeneity in feature expression among patient subgroups. This heterogeneity emphasizes the need for customized and tailored methods for predicting DILI.

Dynamic classifier selection

Evaluation metrics comprised conventional ones and those for imbalanced class distributions, such as accuracy (ACC), precision-recall area under the curve (PR-AUC), F1-score, recall, precision, Matthews correlation coefficient (MCC),

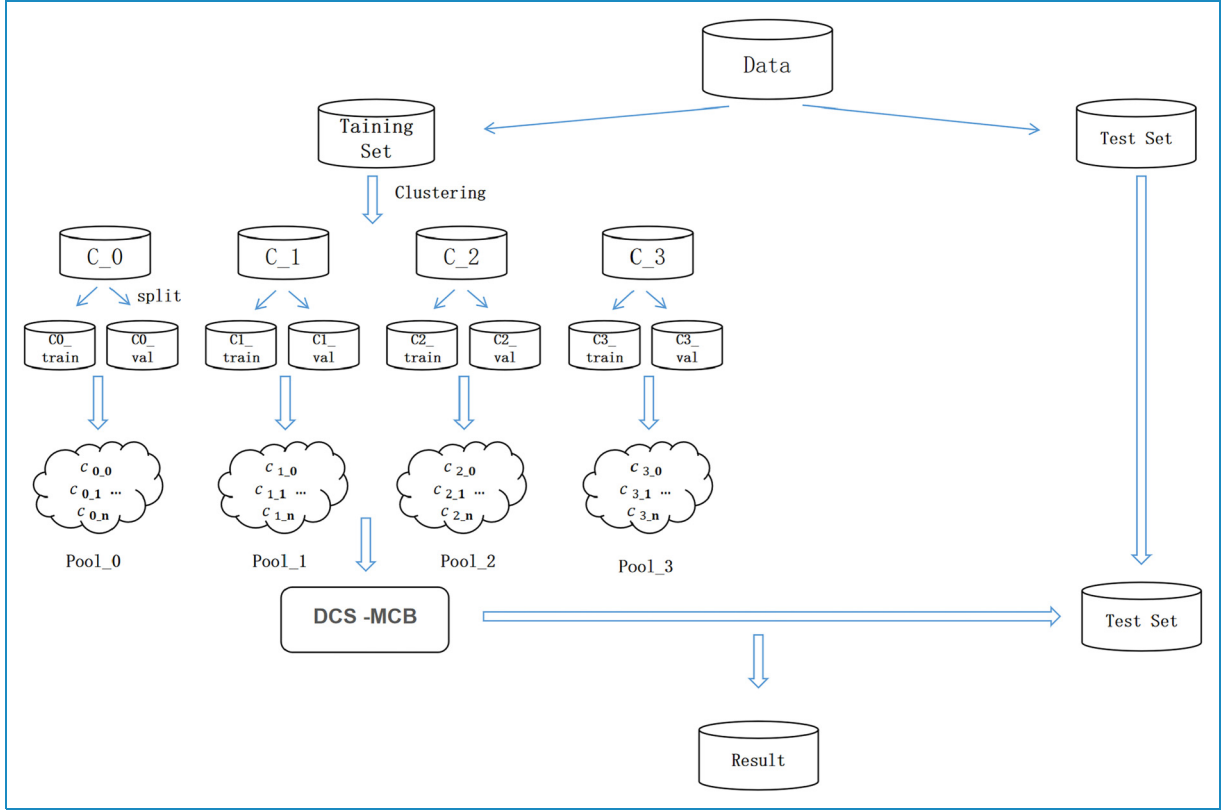


Figure 3. Overview of data processing flow in this study.

and geometric mean (G-mean). Formulas for these metrics are:

True Positives (TP): Correctly predicted positive cases

True Negatives (TN): Correctly predicted negative cases

False Positives (FP): Incorrectly predicted positive cases

False Negatives (FN): Incorrectly predicted negative cases

ACC (Accuracy):

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

It indicates the proportion of correct model predictions.

PR-AUC: It is the area under the Precision-Recall curve and reflects the trade-off between precision and recall of the model across various thresholds. It is obtained by graphing the Precision-Recall curve and computing the area under that curve.

Recall:

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

It represents the proportion of all positive class samples that are correctly predicted by the model.

Precision:

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

It shows the proportion of all samples predicted as positive that are indeed positive.

F1 score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

It is the weighted average of Precision and Recall, used to balance the Precision and Recall performance of the model.

Matthews correlation coefficient (MCC):

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

It is a comprehensive metric for assessing the classifier's performance, ranging from -1 to 1. A value of 1 implies perfect classification, 0 implies random classification, and -1 implies completely incorrect classification.

G-mean:

$$\text{G-mean} = \sqrt{\text{Recall} \times \text{Specificity}}$$

where specificity is given by the formula:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

It measures the balanced performance of the model for both positive and negative classes. The higher the

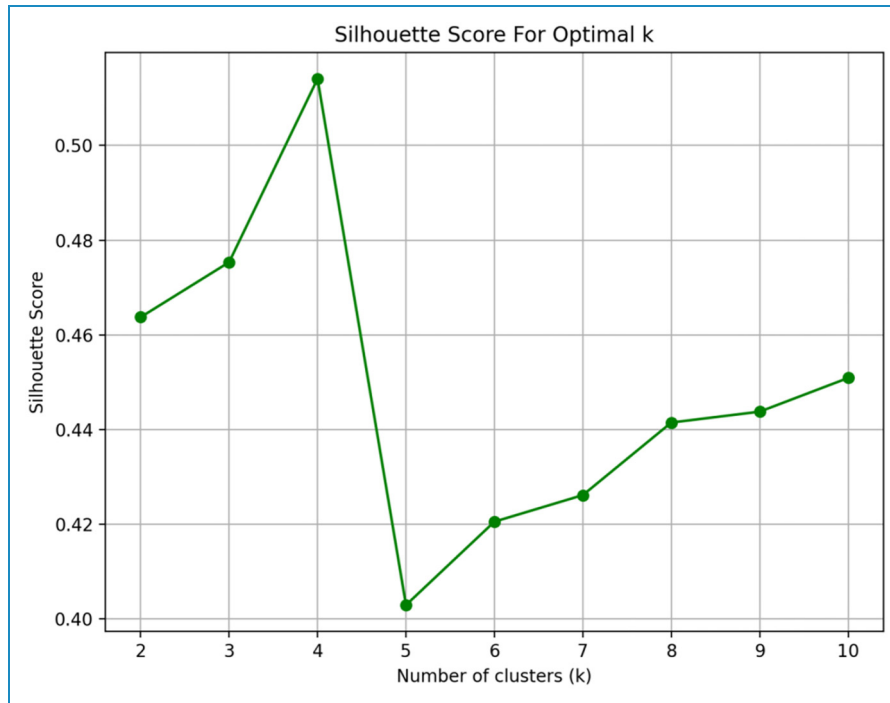


Figure 4. The silhouette coefficient plot of this study.

Table 2. Comparison of training and test data across four clusters.

	Cluster_0	Cluster_1	Cluster_2	Cluster_3
Train set	3641 (3330:3111)	3225 (3089:136)	1015 (556:459)	2163 (2117:46)
Test set	916 (847:69)	781 (752:29)	277 (155:122)	537 (519:18)

G-mean value, the better the model performs for both classes.

When selecting the base classifier pool, we chose a variety of individual and ensemble learning models and evaluated their performance using multiple metrics, such as ACC, Precision, Recall, F1 score, PR-AUC, MCC, and G-mean. The performance of candidate models was assessed against these metrics, and only models with all metrics above 0.5 were included in the final base classifier pool. The models finally included in the base classifier pool are: individual models (Decision Tree [DT], K-Nearest Neighbors [KNN], Logistic Regression [LR], Linear Discriminant Analysis [LDA]) and ensemble models (Random Forest [RF], Bagging, Adaptive Boosting [ADA], LightGBM [LGB]).

Table 3 shows the performance of DCS-MCB and base classifiers on original data. Without clustering, DCS-MCB achieved the highest values in ACC (0.952), F1 (0.750), Precision (0.745), and MCC (0.724). Meanwhile, Bagging achieved the highest Recall value of 0.823, and RF

outperformed DCS-MCB in PR-AUC and G-mean, with values of 0.815 and 0.883, respectively. Table 4 represents evaluation metrics after data clustering. Almost all base classifiers improved in all metrics. For instance, DT's F1 score rose from 0.647 before clustering to 0.858 after. Notably, unlike pre-clustering results, the proposed Cluster_DCS-MCB framework consistently outperformed other classifiers in every metric, achieving the highest ACC (0.986), PR-AUC (0.970), F1 (0.926), Recall (0.932), Precision (0.925), MCC (0.917), and G-mean (0.959). As expected, ensemble models generally outperformed individual ones, and our proposed framework always outperform other ensemble classifiers in all comparisons.

Figure 6 shows the normalized confusion matrices for the four clustering groups. In all four clusters, $TP > 0.9$, showing the DCS_MCB model predicts each category well. Also, misclassification rates (FP and FN) in all four clusters are < 0.1 . In clusters 0 and 1, $FP = 0$, and FN are 0.087 and 0.059 respectively. In clusters 2 and 3, FP (0.022 and 0.05) and FN (0.035 and 0.054) increase slightly

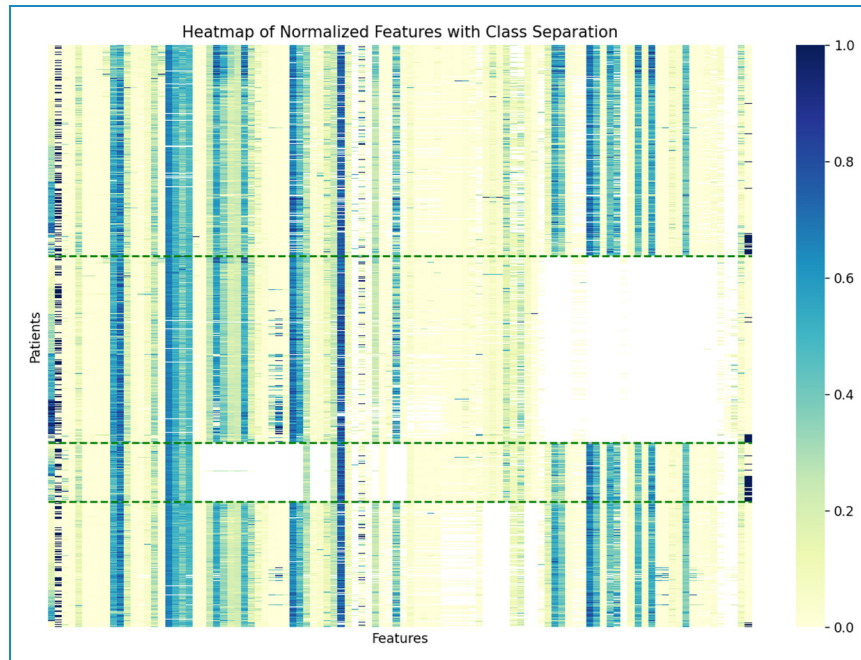


Figure 5. Heatmap analysis of patient features across four distinct clusters in pediatric DILI.

but stay low. These results suggest the model has excellent error control across all categories and reliable predictive performance.

These findings highlight the need to customize classification models for each data cluster's unique attributes. Varying cluster performance shows the importance of dynamic selection methods that adapt to these differences, improving the model's overall accuracy and robustness. This approach is vital in clinical applications, where accurate, context-specific predictions are key for effective patient management and treatment planning.

In summary, our study demonstrates that the integration of clustering and dynamic selection techniques substantially enhances the accuracy and robustness of pediatric DILI prediction. By stratifying patients into distinct subgroups based on complex underlying diseases and lab indicators, we achieved a more detailed analysis of drug effects across different populations. The proposed Cluster_DCS-MCB framework outperformed all evaluated metrics, highlighting the effectiveness of this hybrid approach. These results emphasize the importance of customizing predictive models to each patient subgroup's unique features, enabling more accurate and personalized clinical decision-making. The approach in this study offers a promising way to advance precision medicine and improve patient outcomes in complex clinical settings.

Interpretability

Experiments show DCS usually finds the best classifier for each patient in every cluster, accurately predicting DILI

risk. Whether using a single or multiple classifiers, interpretability methods like SHAP can explain the model's predictions. Figure 7 shows the distribution of feature importance for predicting pediatric DILI risk in one patient subgroup. Higher ALT levels, chemotherapy exposure, more diagnosed diseases, longer hospital stays, more medications, and more frequent hospitalizations were found to significantly increase the predicted DILI risk. The varying feature importance rankings among different clusters match clinical expectations, validating the model's robustness and reliability. Figure 8 depicts the distribution of significant DILI factors in two patients, with predicted probabilities of 0.83 and 0.16. For patient A, elevated GGT, AST/ALT ratio, and number of diagnosed diseases were predictors. For patient B, the AST/ALT ratio, D-DIMER levels, and HBV cAb indicated a lower DILI risk. By identifying key DILI-risk factors for each subgroup, clinicians can better understand individual patient profiles and make tailored medical decisions.

Discussion

The susceptibility factors for pediatric DILI mainly center on the interaction among gender, polypharmacy, and underlying diseases. Although research shows females are more sensitive to DILI,²⁴ in our study, the proportion of female DILI patients (440/1190) was lower than that of males. This difference might be due to environmental factors or data inclusion criteria. Pediatric DILI patients have higher levels of ALP, GGT, TBIL, DBIL, IBIL, and INR than non-DILI patients, mainly because of hepatocellular apoptosis or necrosis.²⁵

Table 3. Performance evaluation of DCS-MCB and base classifiers.

	ACC	PR-AUC	FI	Recall	Precision	Mcc	G-mean
DT	0.932 (0.926–0.938)	0.451 (0.413–0.489)	0.647 (0.614–0.668)	0.658 (0.618–0.676)	0.637 (0.610–0.665)	0.609 (0.586–0.632)	0.795 (0.769–0.822)
KNN	0.926 (0.920–0.932)	0.540 (0.527–0.554)	0.635 (0.614–0.656)	0.677 (0.653–0.702)	0.597 (0.578–0.616)	0.595 (0.572–0.619)	0.803 (0.789–0.818)
LR	0.935 (0.926–0.945)	0.660 (0.623–0.697)	0.684 (0.663–0.704)	0.744 (0.655–0.833)	0.642 (0.558–0.726)	0.653 (0.634–0.672)	0.842 (0.808–0.876)
LDA	0.928 (0.919–0.937)	0.678 (0.652–0.704)	0.684 (0.680–0.689)	0.762 (0.727–0.798)	0.623 (0.600–0.646)	0.649 (0.640–0.659)	0.854 (0.841–0.868)
RF	0.948 (0.941–0.954)	0.815 (0.805–0.825)	0.747 (0.730–0.764)	0.810 (0.781–0.840)	0.695 (0.643–0.747)	0.722 (0.705–0.739)	0.883 (0.871–0.896)
Bagging	0.943 (0.938–0.947)	0.806 (0.779–0.812)	0.729 (0.708–0.750)	0.823 (0.785–0.861)	0.669 (0.638–0.701)	0.702 (0.679–0.725)	0.877 (0.849–0.905)
ADA	0.949 (0.944–0.954)	0.810 (0.799–0.820)	0.748 (0.733–0.763)	0.797 (0.758–0.836)	0.707 (0.649–0.765)	0.723 (0.708–0.738)	0.877 (0.860–0.894)
LGB	0.949 (0.943–0.955)	0.812 (0.797–0.828)	0.743 (0.717–0.769)	0.778 (0.739–0.816)	0.712 (0.664–0.759)	0.716 (0.688–0.744)	0.867 (0.851–0.883)
DCS-MCB	0.952 (0.946–0.959)	0.800 (0.789–0.811)	0.750 (0.722–0.778)	0.756 (0.726–0.786)	0.745 (0.684–0.806)	0.724 (0.692–0.756)	0.858 (0.845–0.870)

Bold values represent the top results for each corresponding indicator.

Table 4. Performance evaluation of cluster_DCS-MCB and base classifiers post clustering.

	ACC	PR-AUC	FI	Recall	Precision	Mcc	G-mean
Cluster_DT	0.969 (0.962–0.976)	0.755 (0.637–0.873)	0.858 (0.805–0.911)	0.893 (0.870–0.916)	0.828 (0.797–0.859)	0.836 (0.815–0.858)	0.933 (0.919–0.947)
Cluster_KNN	0.861 (0.819–0.903)	0.668 (0.560–0.776)	0.625 (0.549–0.701)	0.865 (0.854–0.876)	0.505 (0.403–0.608)	0.543 (0.494–0.592)	0.791 (0.751–0.832)
Cluster_LR	0.934 (0.889–0.979)	0.833 (0.774–0.892)	0.799 (0.761–0.837)	0.863 (0.790–0.934)	0.750 (0.701–0.800)	0.749 (0.733–0.765)	0.880 (0.824–0.936)
Cluster_LDA	0.906 (0.877–0.935)	0.764 (0.693–0.834)	0.747 (0.667–0.827)	0.743 (0.684–0.802)	0.787 (0.759–0.815)	0.674 (0.659–0.689)	0.790 (0.960–0.839)
Cluster_RF	0.968 (0.957–0.979)	0.924 (0.881–0.968)	0.890 (0.842–0.938)	0.848 (0.795–0.901)	0.816 (0.774–0.859)	0.888 (0.844–0.932)	0.921 (0.888–0.954)
Cluster_Bagging	0.976 (0.965–0.988)	0.948 (0.911–0.986)	0.891 (0.851–0.931)	0.889 (0.822–0.956)	0.894 (0.868–0.919)	0.872 (0.845–0.899)	0.930 (0.905–0.956)
Cluster_ADA	0.979 (0.959–0.991)	0.967 (0.950–0.985)	0.920 (0.897–0.943)	0.890 (0.851–0.930)	0.951 (0.934–0.969)	0.903 (0.865–0.941)	0.935 (0.912–0.958)
Cluster_LGB	0.983 (0.972–0.994)	0.965 (0.959–0.971)	0.922 (0.894–0.950)	0.892 (0.869–0.916)	0.916 (0.908–0.924)	0.910 (0.874–0.0947)	0.936 (0.917–0.955)
Cluster_DCS-MCB	0.986 (0.981–0.991)	0.970 (0.958–0.982)	0.926 (0.915–0.938)	0.932 (0.897–0.937)	0.925 (0.910–0.941)	0.917 (0.898–0.937)	0.959 (0.934–0.984)

Bold values represent the top results for each corresponding indicator.

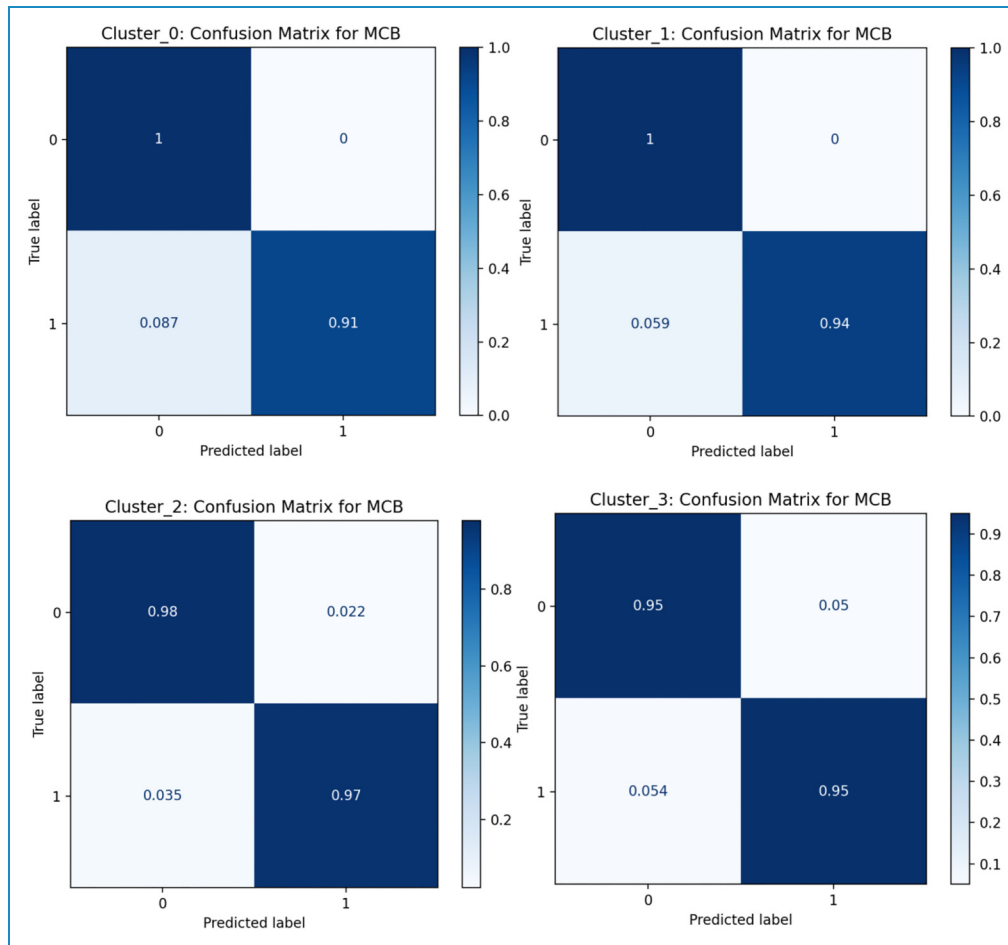


Figure 6. Normalized confusion matrix for four clustering groups.

In cases involving multiple drugs, drug combinations significantly contribute to hepatotoxicity.⁹ Drugs like acetaminophen, antibiotics, NSAIDs, antimicrobial agents, and antineoplastic agents are known DILI triggers,^{26–28} which matches our findings. Also, we observed that drug use patterns and their importance to outcome variables vary among different patient groups. Drug action mechanisms differ too. For example, ceftriaxone shouldn't be given to hyperbilirubinemic neonates as it can displace bilirubin, raising bilirubin levels and reducing unconjugated bilirubin.²⁹ Symptoms from different drug combinations vary between individuals. Healthcare providers must monitor not just traditional liver injury signs like nonspecific symptoms (fatigue, nausea, jaundice, vomiting) and biochemical dysfunctions (elevated ALT and ALP), but also be aware of common pediatric antibiotics and related drugs causing severe liver injury. Thus, careful consideration of drug dosages, frequencies, and timely assessment of potential symptoms and signs are vital in clinical practice.

Moreover, it is crucial to recognize that misclassification rates (false positives and false negatives) can give rise to ethical issues. In the diagnosis of pediatric DILI, for

instance, false positives might result in needless treatment or hospitalizations. This exposes patients to avoidable medical risks, unwanted drug side-effects, and psychological distress. Conversely, false negatives can cause delay in DILI diagnosis and treatment, resulting in more severe liver damage and potentially life-threatening complications. To this end, healthcare professionals must always be cognizant of the potential consequences of misclassifications and utilize the model as an auxiliary decision-making tool, not the sole determinant for decisions. Making treatment decisions requires comprehensively integrating the physician's professional acumen, rich clinical experience, and the model's recommendations to ensure the formulation of more personalized and reliable treatment strategies, thus optimizing patient care and outcomes.

The framework proposed in this study addresses a major shortcoming of traditional methods that treat all patients alike. By optimizing classifiers for the distinct features of each patient subgroup, it enhances both prediction accuracy and robustness. It also proves advantageous in risk stratification, providing practical value for the clinical management of pediatric DILI. However, this study has several

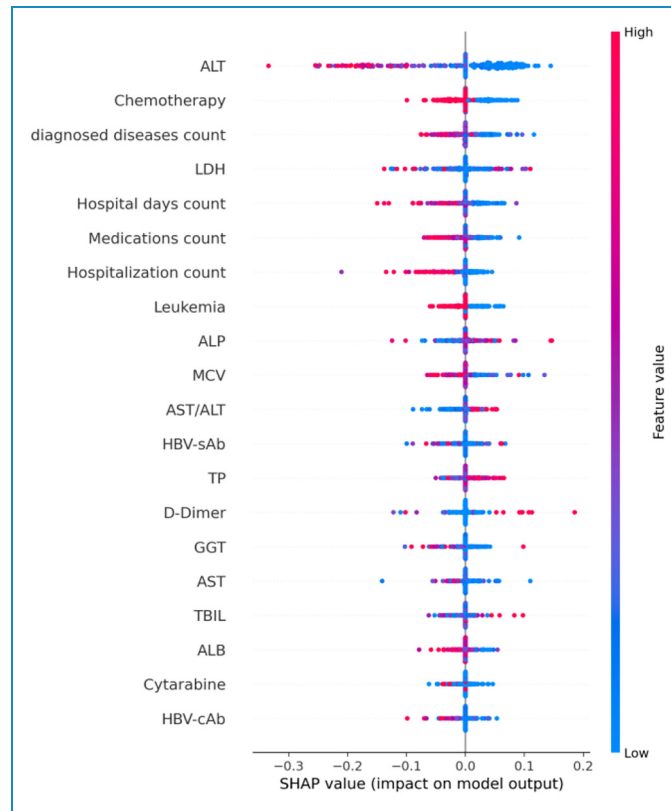


Figure 7. SHAP charts for predicting pediatric DILI risk in one patient subgroup.

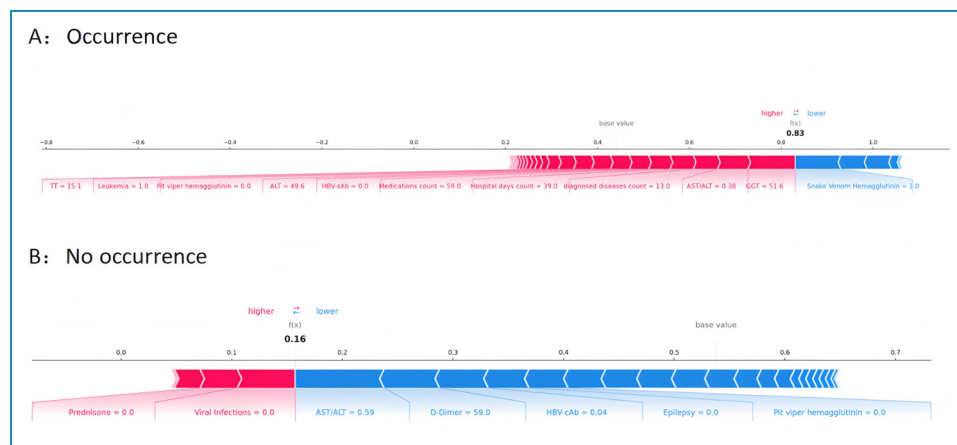


Figure 8. Distribution of significant factors in two pediatric DILI patients.

limitations. First, this study was confined to a specific region. As a result, the applicability of the currently trained model may have certain limitations. One of the strengths of our proposed framework is its ability to dynamically update the models within the classifier pool, thereby ensuring its continued effectiveness across different scenarios to a certain extent. Second, underreporting remains a persistent problem. This phenomenon may lead to the omission of

some pediatric cases, causing the dataset to be skewed towards more severe cases. Consequently, the dataset becomes less representative of the general pediatric population, which could potentially undermine the generalizability and accuracy of our findings. Third, medication patterns, including dosage and frequency, have not been comprehensively taken into account. Additionally, we failed to clearly differentiate between intrinsic and idiosyncratic DILI. Even

though we clustered patients according to underlying diseases, medication use, and laboratory parameters, the specific frequency and dosage of medications, which could be vital for assessing the risk of DILI, were not incorporated into the analysis. Finally, the study did not take into account the influences of factors such as genes, drug components and chemical structures,^{30,31} restricts a more in-depth understanding of DILI etiology and individual risk factors. Further research is expected to thoroughly explore the mechanisms, risk factors, and prognosis of pediatric DILI, and to establish corresponding diagnostic, preventive, and treatment strategies.

Conclusion

This study delves into a framework integrating clustering and dynamic classifier selection, with the aim of enhancing the accuracy and robustness of predictive models for pediatric DILI. The framework identifies the optimal number of clusters and categorizes patients into distinct subgroups. For each subgroup, a dedicated base classifier pool is constructed. Subsequently, the DCS-MCB method is utilized to adaptively pinpoint the most appropriate classifiers for every individual patient. The performance of the proposed framework is compared with common machine-learning models, evaluating its superiority, and assessing the impact of clustering enhancement on both the proposed framework and traditional baselines. Notably, the results demonstrate that the developed model outperforms existing approaches, exhibiting potential in the early identification of high-risk pediatric patients. By characterizing patient subgroups and generating cluster-specific classifiers that account for varying levels of clinical complexity, this framework provides a foundation for developing precise, clinically actionable strategies for patient stratification and drug safety monitoring.

Acknowledgments

The authors would like to thank all who supported this research.

Authors' contributions

Zixin Shi: Methodology, Investigation, Formal Analysis, Validation, Writing-Original Draft, Software. Linjun Huang: Methodology, Investigation, Formal analysis. Haolin Wang: Conceptualization, Methodology, Validation, Writing-Review & Editing, Supervision, Funding Acquisition.

Data availability statement

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval statement

The study was approved by the ethics committees of Chongqing Medical University (Reference Number: 2023096, Date Approved: Dec 20, 2023), and due to its retrospective nature, this study required no informed consent and represented minimal risk to participants.

Funding


The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China under Grant 72101040 and Graduate Research Innovation Project of Chongqing under Grant CYS23374.


National Natural Science Foundation of China, Graduate Research Innovation Project of Chongqing, (grant number 72101040, CYS23374).

Guarantor

Haolin Wang.

ORCID iDs

Linjun Huang  <https://orcid.org/0009-0008-1025-4510>

Haolin Wang  <https://orcid.org/0000-0002-1735-9525>

References

1. Coleman JJ and Pontefract SK. Adverse drug reactions. *CME Clin Pharmacol* 2016; 16(5): 481–485.
2. Alomar MJ. Factors affecting the development of adverse drug reactions (review article). *Saudi Pharm J* 2014; 22: 83–94.
3. Napoleone E. Children and ADRs (adverse drug reactions). *Ital J Pediatr* 2010; 36: 4.
4. Mitchell AA, Lacouture PG, Sheehan JE, et al. Adverse drug reactions in children leading to hospital admission. *Pediatrics* 1988; 82: 24–29.
5. Clavenna A and Bonati M. ADR Epidemiology in paediatrics. *M&B* 2009; 28: 503–504.
6. Gomes ER, Brockow K, Kuyucu S, et al. Drug hypersensitivity in children: report from the pediatric task force of the EAACI Drug Allergy Interest Group. *Allergy* 2016; 71: 149–161.
7. Elzagallaai AA and Rieder MJ. Model based evaluation of hypersensitivity adverse drug reactions to antimicrobial agents in children. *Front Pharmacol* 2021; 12: 638881.
8. Andrade RJ, Chalasani N, Björnsson ES, et al. Drug-induced liver injury. *Nat Rev Dis Primers* 2019; 5(1): 58.
9. Zhu Y, Li YG, Wang JB, et al. Causes, features, and outcomes of drug-induced liver injury in 69 children from China. *Gut Liver* 2015; 9: 525–533.
10. Fisher K, Vuppalaanchi R and Saxena R. Drug-induced liver injury. *Arch Pathol Lab Med* 2015; 139(7): 876–887.
11. Hakkarainen KM, Hedna K, Petzold M, et al. Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions—a meta-analysis. *PLoS One* 2012; 7: e33236.

12. Sandritter TL, Goldman JL, Habiger CJ, et al. An electronic medical records-based approach to identify idiosyncratic drug-induced liver injury in children. *Sci Rep* 2019; 9: 18090–8.
13. Yu Z, Ji H, Xiao J, et al. Predicting adverse drug events in Chinese pediatric inpatients with the associated risk factors: a machine learning study. *Front Pharmacol* 2021; 12: 659099–659099.
14. Hazell L and Shakir SAW. Under-reporting of adverse drug reactions: a systematic review. *Drug Saf* 2006; 29(5): 385–396.
15. Abualigah LMQ. *Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering*, vol. 816. Cham: Springer Nature, 2019.
16. Yuan Q, Guo X, Ren Y, et al. Cluster correlation based method for lncRNA-disease association prediction. *BMC Bioinformatics* 2020; 21: 180.
17. Agrawal U, Soria D, Wagner C, et al. Combining clustering and classification ensembles: a novel pipeline to identify breast cancer profiles. *Artif Intell Med* 2019; 97: 27–37.
18. Sun PG, Gao L and Han S. Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci* 2011; 7: 61–73.
19. Cruz RMO, Sabourin R and Cavalcanti GDC. Dynamic classifier selection: recent advances and perspectives. *Inf Fusion* 2018; 41: 195–216.
20. Cruz RMO, Hafemann LG and Sabourin R. DESlib: a dynamic ensemble selection library in Python. *J Mach Learn Res* 2020; 21(1): 283–287.
21. Cruz RM, Sabourin R, Cavalcanti GD, et al. META-DES: a dynamic ensemble selection framework using meta-learning. *Pattern Recognit* 2015; 48: 1925–1935.
22. Alceu RSLE and Britto S Jr. Dynamic selection of classifiers—a comprehensive review. *Pattern Recognit* 2014; 47(11): 3665–3680.
23. Wu J, Shen J, Xu M, et al. A novel combined dynamic ensemble selection model for imbalanced data to detect COVID-19 from complete blood count. *Comput Methods Programs Biomed* 2021; 211: 106444.
24. Rademaker M. Do women have more adverse drug reactions? *Am J Clin Dermatol* 2001; 2: 349–351.
25. Abboud G and Kaplowitz N. Drug-induced liver injury. *Clin Infect Dis* 2007; 30(4): 277–294.
26. Lai R, Li X, Zhang J, et al. Drug-induced liver injury in children: a nationwide cohort study from China. *JHEP Rep* 2024; 6: 101102.
27. Dipaola F, Molleston JP, Gu J, et al. Antimicrobials and anti-epileptics are the leading causes of Idiosyncratic Drug Induced Liver Injury in American Children. *J Pediatr Gastroenterol Nutr* 2019; 69: 152–159.
28. Niu H, Atallah E, Alvarez-Alvarez I, et al. Therapeutic management of idiosyncratic drug-induced liver injury and acetaminophen hepatotoxicity in the paediatric population: a systematic review. *Drug Saf* 2022; 45: 1329–1348.
29. Faa G, Ekstrom J, Castagnola M, et al. A developmental approach to drug-induced liver injury in newborns and children. *Curr Med Chem* 2012; 19: 4581–4594.
30. Kha Q-H, Le V-H, Hung TNK, et al. Development and validation of an explainable machine learning-based prediction model for drug–food interactions from chemical structures. *Sensors* 2023; 23: 3962. doi: 10.3390/s23083962
31. Vo TH, Nguyen NT and Le NQ. Improved prediction of drug–drug interactions using ensemble deep neural networks. *Med Drug Discov* 2023; 17: 100149.