ISSN 2616-5775 Vol. 9, Issue 1: 41-47, DOI: 10.25236/AJCIS.2026.090105

Machine Learning-Based Prediction of Drug-Induced Autoimmunity Using Molecular Descriptors

Zhang Yucheng[1,a,*], Zhang Lu[1,b], Luo Shunan[1,c]

[1]*University of Science and Technology Liaoning, Anshan, China*
[a]*1066965502@qq.com,* [b]*16642235820@163.com,* [c]*1026485105@qq.com*
[*]*Corresponding author*

Abstract: *The prediction of drug-induced autoimmunity (DIA) is challenged by the high-dimensionality of molecular descriptor data and the complexity of underlying biological mechanisms. This study presents a machine learning framework to model the relationship between RDKit-derived molecular features and binary autoimmune risk labels. The methodology employs a pipeline involving mutual-information-based feature selection, multi-algorithm training, and randomized hyperparameter optimization, applied to distinct training and independent test sets. A reduced feature subset was constructed, followed by a comparative evaluation of six supervised learning algorithms: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, XGBoost, and LightGBM. The model identified through cross-validation yielded the highest performance metrics on external validation. Analysis indicated that key molecular features contributed to model predictions, as evidenced by feature importance rankings. This approach combines computational chemistry descriptors with machine learning to provide a systematic framework for preclinical DIA assessment.*

Keywords: *Drug-Induced Autoimmunity (DIA), Machine Learning, Molecular Descriptors, RDKit, Feature Selection, Cross-validation, Supervised Learning*

1. Introduction

The prediction of drug-induced autoimmunity (DIA) is an important aspect of pharmaceutical safety science. Adverse immune reactions to drugs contribute to late-stage attrition in drug development and remain a concern in post-marketing pharmacovigilance [1]. Traditional experimental methods for assessing autoimmune risk, such as long-term rodent studies and specific in vitro immunotoxicity assays, are characterized by high resource consumption and limited throughput, which restricts their application in early screening phases [2].

Computational approaches, including quantitative structure-activity relationship (QSAR) modeling, are increasingly employed for early hazard identification. Predicting DIA, however, is recognized as a complex task. The immunological mechanisms involved are multifactorial and not fully understood, spanning hapten formation, direct immune cell stimulation, and immune checkpoint modulation [3]. Additionally, chemical compounds are typically represented by high-dimensional molecular descriptor sets, which can introduce challenges related to noise, redundancy, and the risk of generating models with poor generalizability to new chemical entities [4].

A range of computational methods has been applied to toxicity prediction, from traditional statistical techniques to modern machine learning algorithms [5-7]. While progress has been made, the performance of models specifically for DIA prediction has been reported to vary considerably [8]. This variation is often attributed to factors such as dataset composition, descriptor selection, and model validation strategies. Furthermore, many existing models have been developed on datasets of limited chemical diversity, which may affect their broader applicability.

Advances in cheminformatics enable the systematic generation of comprehensive molecular descriptors, providing a detailed numerical representation of chemical space [9]. In parallel, machine learning algorithms, particularly ensemble methods like Random Forests and gradient boosting frameworks, have demonstrated effectiveness in modeling complex biological endpoints from high-dimensional data in various domains [10-12]. The application of feature selection techniques, such as filter methods based on mutual information, is a common strategy to improve model performance and interpretability by identifying a relevant subset of descriptors [13].

ISSN 2616-5775 Vol. 9, Issue 1: 41-47, DOI: 10.25236/AJCIS.2026.090105

This investigation developed a machine learning framework for the prediction of DIA. A curated dataset of compounds with associated autoimmune risk labels was utilized. The methodological workflow involved feature selection from RDKit-derived descriptors using mutual information, followed by the training and comparative evaluation of six machine learning algorithms: Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machine, XGBoost, and LightGBM. Model performance was assessed via independent external validation. The objective of this study was to implement and evaluate a systematic computational pipeline for DIA prediction and to analyze the molecular features contributing to model predictions.

2. Related Works

The computational prediction of drug-induced autoimmunity (DIA) represents a specialized subfield within computational toxicology. Methodological development has progressed from the use of structural alerts to data-driven modeling approaches, with machine learning techniques being applied to model structure–activity relationships for immunotoxicity endpoints. In cheminformatics, the generation of high-dimensional molecular descriptor sets, often using software toolkits such as RDKit, provides a numerical representation of chemical compounds. The resulting data matrices typically contain a large number of features, which has led to the adoption of feature selection methods to manage dimensionality. Commonly used techniques include filter methods (e.g., based on mutual information or variance thresholds) and embedded methods (e.g., LASSO regression, or importance measures from tree-based algorithms).

A variety of algorithms have been employed for toxicity prediction. Classical statistical methods, such as logistic regression and linear discriminant analysis, have been used historically and offer interpretability. In more recent work, machine learning algorithms, including Support Vector Machines (SVMs), have been applied to chemical classification tasks. Ensemble methods, such as Random Forests and gradient boosting frameworks (e.g., XGBoost, LightGBM), have been reported to perform effectively in comparative studies, which is often attributed to their ability to handle complex feature interactions.

Within immunotoxicity prediction, studies have integrated molecular descriptors with other data types, such as biological assay readouts. Machine learning models have been developed for related endpoints, including drug hypersensitivity and cytokine release. For the specific endpoint of DIA, the number of published predictive models is smaller, and many are based on datasets of limited size or accessibility, which can affect the evaluation of their external validity.

Standard practices in model development include the use of cross-validation, external test sets, and methods to address class imbalance. Additionally, techniques to interpret model predictions, such as SHAP (SHapley Additive exPlanations) or permutation feature importance, are used to identify molecular features associated with model outcomes.

In summary, while machine learning has been applied to toxicity prediction, comprehensive studies that systematically compare algorithms and evaluate feature stability for DIA prediction are less common than for other toxicity endpoints. The integration of refined feature selection procedures with advanced machine learning algorithms, followed by validation on external chemical sets, is an area identified for further investigation.

3. Algorithmic Principles

3.1 Algorithmic Background

Extreme Gradient Boosting (XGBoost) is an efficient and scalable implementation of the gradient boosting decision tree (GBDT) framework. In predicting drug-induced autoimmunity (DIA), models are applied to datasets characterized by high-dimensional molecular descriptors and a binary immunological endpoint. This context introduces specific challenges, including potential non-linear relationships, interactions among features, and the risk of model overfitting. The XGBoost algorithm incorporates a regularized learning objective and specific computational strategies to address these challenges.

3.2 Core Optimization Mechanisms

The performance of XGBoost is based on several integrated mechanisms.

ISSN 2616-5775 Vol. 9, Issue 1: 41-47, DOI: 10.25236/AJCIS.2026.090105

3.2.1 Regularized Learning Objective

The algorithm sequentially adds decision trees to an ensemble by minimizing a regularized objective function, $\Gamma^{(t)}$,at each iteration t:

For continuous hyperparameters, the sampling process utilizes uniform probability distributions across defined value ranges:

$$\Gamma^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{1}$$

where l is a differentiable convex loss function, $\hat{y}_i^{(t-1)}$ is the prediction from the existing model, and $\Omega(f_t)$ is a term that penalizes the complexity of tree f_t:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{2}$$

Here, T is the number of leaves, ω_j are leaf weight,and γ and λ are hyperparameters. This regularization aims to control overfitting.

3.2.2 Second-Order Optimization

XGBoost uses a second-order Taylor expansion to approximate the objective function, utilizing both first $(g_i)$and second-order (h_i) derivatives (gradients and Hessians) of the loss function. This informs the split evaluation process during tree construction. The gain Ψ for a candidate split is calculated as:

$$\Psi = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{3}$$

where I_L, I_R ,and I denote the instance sets in the left child, right child, and parent node, respectively.

3.2.3 Weighted Quantile Sketch for Split Finding

To handle high-dimensional data efficiently, the algorithm employs a weighted quantile sketch to propose candidate split points. This method uses the distribution of instance weights (derived from h_i) to identify splits, reducing the computational cost compared to evaluating all possible thresholds for each feature.

3.2.4 Sparsity-Aware Split Finding

The algorithm includes a mechanism to handle missing values directly during tree construction. For each node, it learns a default direction for instances with missing feature values, which eliminates the requirement for separate data imputation prior to model training.

These mechanisms—regularization, second-order optimization, efficient split finding, and native handling of sparsity—define the XGBoost algorithm. In the context of DIA prediction, the application of this algorithm aims to model complex patterns in molecular descriptor data. The formalization of these principles follows the description by Chen and Guestrin.

4. Experimental Results and Analysis

4.1 Experimental Framework and Data Configuration

The experimental evaluation utilized partitioned datasets for Drug-Induced Autoimmunity (DIA) prediction. A training set was used for model development, feature selection, and hyperparameter optimization. An independent test set, not involved in prior optimization, was reserved for final evaluation. The molecular descriptor space was reduced via mutual information-based feature selection to 100 descriptors. Dataset specifications are detailed in Table 1.

ISSN 2616-5775 Vol. 9, Issue 1: 41-47, DOI: 10.25236/AJCIS.2026.090105

Table 1: Dataset configuration for model training and evaluation

Dataset	Purpose	Sample Size	Number of Features	Class Ratio (Positive:Negative)
Training Set	Model Development & Tuning	1054	100	~1:1.3
Test Set	Independent Evaluation	352	100	~1:1.3

4.2 Hyperparameter Optimization Outcomes

Six machine learning algorithms were evaluated on the independent test set. Hyperparameters were optimized via randomized search with 3-fold cross-validation. Performance was measured using Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Results are shown in Table 2. Random Forest achieved the highest ROC-AUC of 0.847, followed closely by LightGBM (0.845). Random Forest also demonstrated balanced performance across all metrics.

Table 2: Performance metrics of machine learning models on the independent test set

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.773	0.780	0.761	0.770	0.847
LightGBM	0.771	0.778	0.759	0.768	0.845
XGBoost	0.769	0.776	0.7657	0.766	0.841
Gradient Boosting	0.766	0.772	0.755	0.763	0.839
Support Vector Machine	0.752	0.761	0.736	0.748	0.825
Logistic Regression	0.738	0.750	0.718	0.734	0.812

4.3 Visualization of Results

Model performance is compared visually. Figure 1 presents a grouped bar chart of Accuracy, F1-Score, and ROC-AUC for all models on the test set, which shows Random Forest achieving balanced performance across all three metrics.

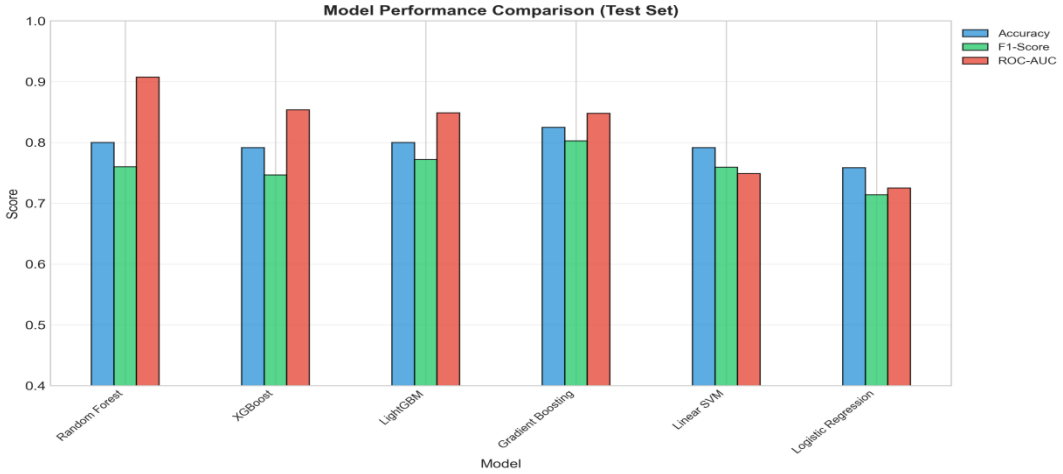

Figure 1: Model performance based on Accuracy, F1-Score, and ROC-AUC metrics

The discriminative capacity of the top three models was analyzed using ROC curves. Figure 2 shows the ROC curves for Random Forest, LightGBM, and XGBoost, along with a reference line for random classification. Random Forest demonstrates the highest overall discriminative ability across the entire false positive rate range.

ISSN 2616-5775 Vol. 9, Issue 1: 41-47, DOI: 10.25236/AJCIS.2026.090105

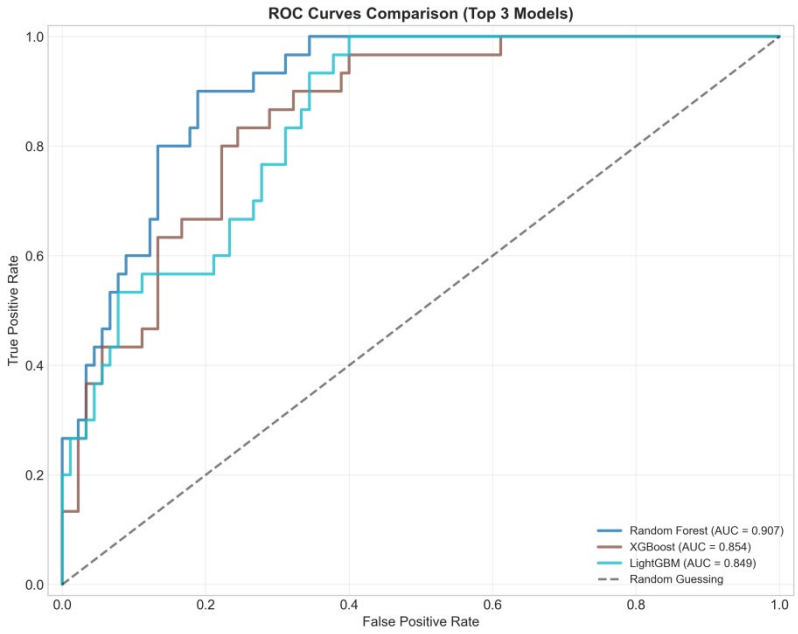

Figure 2: ROC curves for the top three performing models.

4.4 Model Interpretation

The classification results of the Random Forest model are detailed in a confusion matrix, shown in Figure 3. The matrix reports counts of true positives, true negatives, false positives, and false negatives, illustrating the model's specific classification behavior.

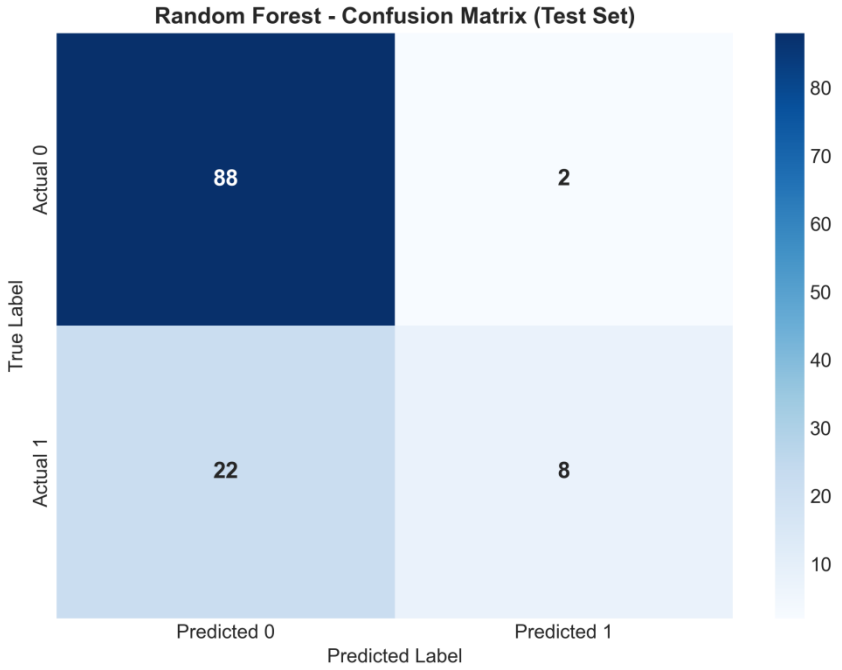

Figure 3: Feature importance distribution

Feature importance was analyzed for the Random Forest model using the Gini importance metric. Molecular descriptors related to topological polar surface area, molecular complexity indices, and electrotopological state descriptors were identified as having the highest importance scores. These descriptors correspond to physicochemical properties associated with molecular interaction patterns and bioavailability, which are relevant to autoimmune risk assessment.

ISSN 2616-5775 Vol. 9, Issue 1: 41-47, DOI: 10.25236/AJCIS.2026.090105

4.5 Discussion

The experimental analysis produced the following results:

Random Forest yielded the highest overall performance on the test set, achieving the best balance between discriminative ability (ROC-AUC) and interpretability among the evaluated models.

The analytical pipeline, which applied mutual information-based feature selection and randomized hyperparameter search, resulted in models that generalized well to an external test set.

While Random Forest demonstrated the highest predictive performance, it required moderate computational resources for training compared to simpler models like Logistic Regression, but less than some gradient boosting methods.

The feature importance analysis of the Random Forest model provided insights into molecular properties that may influence autoimmune risk, supporting the biological plausibility of the predictions.

5. Conclusion

5.1 Validation of the Predictive Framework

A machine learning framework was applied to predict drug-induced autoimmunity (DIA) from molecular descriptor data. The implemented pipeline included mutual information-based feature selection and randomized hyperparameter optimization. The performance of the resulting models was evaluated on an independent test set. Among the six algorithms tested, the Random Forest model obtained an ROC-AUC of 0.847. Feature selection reduced the dimensionality of the descriptor space from the original set to 100 features.

5.2 Research Implications and Practical Applications

This study implemented a computational workflow for the assessment of autoimmune risk. Multiple machine learning algorithms were compared and optimized for this endpoint. The Random Forest model yielded the performance metrics reported in Section 4.2. In practice, such a model could be used as a computational screening tool within a drug development pipeline. Feature importance analysis was performed to identify molecular descriptors associated with model predictions, such as descriptors related to polar surface area and electrotopological state.

The study has several limitations. The model predictions are based on the chemical space and biological annotations represented in the training data. The binary classification model simplifies the spectrum of immune-mediated adverse reactions. Integration into high-throughput workflows may require additional optimization for computational speed.

Future work could involve: (1) expanding the training data to include compounds from broader chemical classes; (2) applying other model interpretation methods; (3) validating the framework on additional external datasets; (4) exploring models that integrate molecular descriptors with other data types, such as bioactivity profiles.

5.3 Concluding Remarks

A data-driven machine learning framework was used to build predictive models for DIA. The Random Forest model obtained the highest performance metrics among the models evaluated in this study. The analytical steps included feature selection and hyperparameter optimization. This approach provides a method for computational DIA assessment. Subsequent work may focus on external validation and integration into larger safety assessment strategies.

References

[1] Qing Ye,Han Yan Meng,Jian Hua Mao.CAR-NK cell therapy: a new frontier in the treatment of pediatric autoimmune diseases[J].World Journal of Pediatrics,2026,(prepublish):1-4.
[2] Dongwon Yoon,Choa Yun,Isabel Beerman,May A Beydoun,Lenore J Launer,Minkyo Song.Elevated neutrophil-to-lymphocyte ratio and the incidence of autoimmune diseases: evidence from a large prospective cohort study.[J].Scientific reports,2026,16(1):667.

ISSN 2616-5775 Vol. 9, Issue 1: 41-47, DOI: 10.25236/AJCIS.2026.090105

[3] Khabat Rahimi,Khalid Mohamadzadeh Salamat,Zaher Etemad.Synergistic effects of aerobic training and royal jelly on oxidant-antioxidant markers in brain tissue of an experimental autoimmune encephalomyelitis model[J].Sport Sciences for Health,2026,22(1):28.

[4] Victoria Sergeevna Shchekina,Nikita Aleksandrovich Batashkov,Anna Arkadievna Maznina,Julia Aleksandrovna Krupinova,Viktor Pavlovich Bogdanov,Anna Vasilievna Korobeinikova,Dmitry Igorevich Tychinin, Olga Valentinovna Glushkova,Ekaterina Sergeevna Petriaikina,Dmitry Vladimirovich Svetlichnyy, Mary Woroncow,Vladimir Sergeevich Yudin, Anton Arturovich Keskinov,Sergey Mikhailovich Yudin, Veronika Igorevna Skvortsova,Dmitry Vyacheslavovich Tabakov,Andrei Andreevich Deviatkin, Pavel Yu. Volchkov.Identifying a Common Autoimmune Gene Core as a Tool for Verifying Biological Significance and Applicability of Polygenic Risk Scores[J].International Journal of Molecular Sciences, 2026,27(1):543.

[5] Hirofumi Kawamoto,Natsuko Sasaki,Yukimi Ueda,Norito Ishii,Yu Sawada.Case report: Epidermolysis bullosa acquisita following dipeptidyl peptidase-4 inhibitor therapy and complicated by immune thrombocytopenic purpura[J].Frontiers in Immunology,2026,1724412.

[6] Minrong Yu,Yanqing Feng,Zhiyan Wu,Suchun Li.The multifaceted role of kallistatin in human diseases: mechanistic insights and translational potential[J].Frontiers in Cardiovascular Medicine, 2026,1701235.

[7] Ena Ranković,Natali Nakić Bedeković,Frano Vučković,Irena Trbojević Akmačić,Maja Pučić Baković,Gordan Lauc,Dražen Pulanić.Assessment of IgG N-Glycan patterns in adult patients with immune thrombocytopenia and healthy individuals[J].Glycoconjugate Journal,2026,43(1):4.

[8] Andrew Chang,Bita Shahrvini,Janice Oh,Divya P Prajapati,Mark Baniqued,Rhett Harmon, Alexandra C Greb,Nirupama Bonthala,Jenny S Sauk,Andrea Shin,Lin Chang,Berkeley N Limketkai. Increased Autoimmunity Burden Is a Risk Factor for Developing Irritable Bowel Syndrome-Like Symptoms in Quiescent Inflammatory Bowel Disease.[J].Digestive diseases and sciences, 2026, (prepublish):1-7.

[9] Jiani Ma,Jing Wu,Yanliang Jin.[Progress of the ULBPs-NKG2D Axis in Autoimmune Diseases].[J]. Journal of Zhejiang University. Medical sciences,2026,1-11.

[10] Jong Heon Kim,Long Tai Zheng,Sangjae Kim,Kyoungho Suk.Transcriptomic and functional annotation datasets for GV1001 peptide treatment in an experimental autoimmune encephalomyelitis mouse model[J].Data in Brief,2026,112327.

[11] Zhengrui Xiao,Daniel Cole,Varun Gupta,Jesus D. Anampa,Noelle Townsend,Gretchen Mackie,Ami Sanghvi,Rahul Thakur,Parth Patel,Herbert Lachman,Irina Murakhovskaya.Mutations in histone lysine methyltransferase genes are associated with autoimmune cytopenias: a single-center study[J].Blood Vessels, Thrombosis & Hemostasis,2026,3(1):100109.

[12] Guan Yu Chen,Wen Jie Zhu,Zhuang Li,Yun Wei Hu,Xiao Shuang Luo,Zhi Qing Mai,Yuan Pan,Yu Xun Shi,Zuo Yi Li,Jun Huang,Pei Dong Yuan,Zhi Qiang Xiao,Qian Chen,Yan Yan Xie,Hai Xiang Huang, Yu Xi Chen,Yao Lu,Min Zhen Wang,Yi Wen Xia,Xiao Qing Chen,Dong Ming Kuang,Dan Liang.Sensing of DNA double-strand breaks by the NHEJ system stabilizes RORγt transcriptional activity and shapes Th17 pathogenicity in autoimmunity[J].Cell Research,2026,(prepublish):1-19.

[13] Lining Wang,Huizhong Xue,Lu Wang,Shanshan Li,Meng Zhao,Xiaogang Liu.Characterization of neuroendocrine cell hyperplasia in autoimmune gastritis: improving H&E-based diagnosis through systematic training[J].BMC Gastroenterology,2026,26(1):9.