Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Machine Learning for Predicting Progression From IgA Vasculitis to IgA Vasculitis Nephritis in Children: A Dual-Centre Retrospective Study

Qingkai Wang
Shanxian Central Hospital
Hao Qiu
Siyang Hospital
Liran Shen
Shanxian Central Hospital
Jinxing Dai
Siyang Hospital
Fachen Miao
Shanxian Central Hospital
Qianjin Shi
Siyang Hospital
Yunbiao Zhang
Shanxian Central Hospital
Kang Shen
Siyang Hospital
Weibing Qiu
qwb13511784020@126.com
Siyang Hospital

Research Article

Keywords: IgA vasculitis, IgA vasculitis nephritis, Machine learning, Risk strat-ification

Posted Date: April 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9261325/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Objective

IgA vasculitis nephritis (IgAVN) plays a decisive role in the long-term prognosis of pediatric IgA vasculitis (IgAV), yet the temporal lag of routine urinalysis frequently hinders early and precise risk stratification. This study aimed to develop a machine-learning predictive model using non-invasive peripheral blood parameters to facilitate early identification of IgAVN progression and to reveal underlying pathophysiological risk thresholds.

Methods

This retrospective study enrolled 509 pediatric IgAV patients from Siyang Hospital and Shanxian Central Hospital, among whom 213 developed IgAVN. Twelve core features were selected based on the intersection of LASSO regression and the Boruta algorithm. Seven machine learning algorithms were systematically evaluated to construct the optimal eXtreme Gradient Boosting (XGBoost) model. Furthermore, the Shapley Additive Explanations (SHAP) framework was incorporated to quantify feature importance and to decipher non-linear risk interactions.

Results

The XGBoost model demonstrated outstanding predictive performance in the independent validation set, achieving an area under the receiver operating characteristic curve (AUC) of 0.966, an accuracy of 0.907, an F1 score of 0.892, and a sensitivity of 0.921. SHAP analysis identified the Inflammatory Burden Index (IBI), C-reactive protein (CRP), and monocyte-to-lymphocyte ratio (MLR) as the primary driving factors. SHAP dependence plots revealed critical non-linear threshold effects: the risk of IgAVN escalated sharply and non-linearly in the presence of early subclinical albumin (ALB) depletion and decompensated inflammatory load. Decision curve analysis (DCA) demonstrated that the model achieved substantial clinical net benefit across a broad continuum of threshold probabilities.

Conclusion

The explainable XGBoost model, developed utilising routine non-invasive peripheral blood parameters, demonstrates promising potential as a supportive tool for the early risk stratification of IgAVN. By visualising complex, data-driven nonlinear risk inflexion points, this model may assist frontline clinicians in more effectively identifying high-risk pediatric patients in outpatient and emergency settings. Ultimately, these findings provide an objective reference to inform future clinical strategies to optimise the timing of interventions and potentially minimise unnecessary immunosuppressive exposure in low-risk patients.

Introduction

IgA vasculitis (IgAV), formerly known as Henoch-Schönlein purpura (HSP), is among the most common systemic small-vessel vasculitides in childhood, pathologically characterised by the predominant deposition of IgA1-dominant immune complexes [1, 2]. Although the classic presentations of cutaneous purpura, as well as articular and gastrointestinal involvement, typically follow a self-limiting course, the long-term prognosis and overall disease burden in pediatric patients are largely dictated by the onset and severity of IgA vasculitis nephritis (IgAVN), the most severe form of target-organ damage [3, 4].In clinical practice, the onset and progression of IgAVN exhibit substantial clinical heterogeneity and temporal unpredictability. While some patients present merely with transient microscopic hematuria or mild proteinuria, others may experience an insidious yet rapid progression to crescentic glomerulonephritis during the early stages of the disease, ultimately culminating in end-stage renal disease (ESRD) [5, 6]. This pronounced disparity in prognostic outcomes, coupled with the highly unpredictable course of disease, poses a significant challenge for clinicians in risk assessment at the time of initial IgAV diagnosis. Consequently, there is an urgent need to develop an early-warning tool that facilitates precise identification and risk-stratified management in the early phases of the disease.

Currently, renal biopsy remains the gold standard for the definitive diagnosis and pathological grading of IgAVN; however, its invasive nature precludes its widespread application as a routine screening modality for pediatric patients presenting at outpatient or emergency settings [7]. In clinical settings, the identification of early renal involvement relies primarily on routine urinalysis, including monitoring for proteinuria and microscopic hematuria [8]. However,

abnormalities in urinary parameters often exhibit a temporal lag; by the time definitive urinary abnormalities are clinically detected, glomerular microvascular injury and parenchymal lesions have typically already occurred [9]. These diagnostic limitations present substantial challenges for early intervention: high-risk patients with initially negative urinalysis results may miss the optimal therapeutic window to arrest inflammatory progression, whereas low-risk patients may be subjected to unnecessary exposure to glucocorticoids or immunosuppressive therapies.

Given that the pathogenesis of IgAV inherently involves systemic immune dysregulation and pervasive microvascular inflammation, an evident cascade of inflammatory activation and early metabolic stress typically emerges in the peripheral blood of pediatric patients before the onset of irreversible structural renal damage [10]. Routine peripheral blood biomarkers, including IBI and CRP (reflecting acute inflammatory burden), MLR (indicating immune network dysregulation), and ALB and UA (representing early endothelial function and nutritional-metabolic status), serve as noninvasive, readily accessible metrics. Theoretically, these indices can objectively delineate the pathophysiological trajectory of disease progression across multiple dimensions [11]. Previously, Özdemir et al. systematically evaluated the diagnostic efficacy of routine haematological parameters in predicting visceral involvement in pediatric IgAV using traditional logistic regression. Their findings demonstrated that elevated peripheral monocyte counts, the neutrophil-to-lymphocyte ratio (NLR), and CRP levels are independent risk factors for acute organ involvement [12]. However, these high-dimensional clinical features frequently exhibit complex multicollinearity and nonlinear interaction effects, which traditional statistical models predicated on linear assumptions struggle to adequately address[13, 14]. In recent years, machine learning (ML) algorithms—particularly tree-based ensemble methods such as Extreme Gradient Boosting (XGBoost)—have demonstrated substantial advantages in handling high-dimensional, collinear data and capturing higher-order nonlinear associations among variables [15, 16]. Concurrently, the Shapley Additive exPlanations (SHAP) framework mitigates the inherent "black-box" limitations of complex algorithms. It not only quantifies the global importance of individual features but also delineates nonlinear risk trajectories and potential threshold effects associated with deviations of continuous variables from physiological homeostasis [17, 18].

Building on this rationale, the present study aimed to develop and optimise a machine learning model based on routine peripheral blood inflammatory and metabolic indices to predict progression from IgAV to IgAVN in pediatric patients. Furthermore, we employed the SHAP framework to elucidate the pathophysiological contributions of core biomarkers to renal microvascular injury. Ultimately, this research endeavours to provide an efficient primary screening tool for outpatient and emergency settings, facilitating individualised precision risk stratification and establishing a robust evidence base to guide clinical decision-making.

Methods

Study Design and Population

In this retrospective study, we initially screened 684 pediatric patients diagnosed with IgAV who presented to the Departments of Pediatrics at Siyang Hospital and Shanxian Central Hospital between June 2022 and August 2025. After strictly applying the predefined exclusion criteria, 175 patients were excluded. Ultimately, 509 eligible patients—comprising 213 in the progression group and 296 in the non-progression group—were enrolled for subsequent feature extraction and model construction. This study was conducted in strict adherence to the principles of the Declaration of Helsinki and was approved by the respective Institutional Review Boards (Ethics Approval No. KS2026002). All included patients were aged ≤ 18 years at initial presentation, and the diagnosis of IgAV was established precisely according to the 2010 joint classification criteria endorsed by the European League Against Rheumatism (EULAR), the Paediatric Rheumatology International Trials Organisation (PRINTO), and the Paediatric Rheumatology European Society (PRES) [19].

The primary outcome was defined as progression to IgAVN within 6 months of disease onset. In accordance with the Kidney Disease: Improving Global Outcomes (KDIGO) guidelines [20], incident IgAVN was defined by the persistent presence of any of the following criteria during the follow-up period: (1) a random urinary albumin-to-creatinine ratio (ACR) > 30 mg/g, a urinary protein-to-creatinine ratio (UPCR) > 0.2 mg/mg, or a qualitative proteinuria of $\geq 1+$ on two consecutive routine urinalyses; (2) microscopic hematuria, defined as > 5 red blood cells per high-power field (HPF) on two consecutive examinations, or the presence of macroscopic hematuria; or (3) biopsy-proven IgAVN.

To mitigate the influence of potential confounders on baseline peripheral blood immunological and metabolic profiles, patients were excluded if they met any of the following criteria: (1) a prior definitive diagnosis of primary nephrotic syndrome, IgA nephropathy (IgAN), or other secondary renal diseases; (2) concomitant systemic autoimmune disorders, such as systemic lupus erythematosus (SLE) or antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis; (3) the presence of severe acute infections at admission (e.g., sepsis or severe pneumonia) that could significantly alter the systemic inflammatory burden;(4) exposure to systemic high-dose glucocorticoids or immunosuppressive therapies before baseline blood sample collection.

Data Selection and Preprocessing

We systematically reviewed the electronic medical records of the two participating centres to extract baseline data obtained within 24 hours of the initial diagnosis of IgAV, strictly before the administration of any systemic high-dose glucocorticoids or immunosuppressive therapies. Initial candidate features encompassed routinely available demographic data, prodromal infection-related parameters, biochemical indices, immunological profiles, coagulation function metrics, complete blood count (CBC) parameters, and acute-phase reactants in the outpatient or emergency settings. Specific variables included age, sex, alanine aminotransferase (ALT), alkaline phosphatase (ALP), gamma-glutamyl transferase (GGT), total protein (TP), ALB, urea, creatinine (CREA), UA, total carbon dioxide (TCO2), immunoglobulins (IgA, IgG, IgM), complement components (C3, C4), lactate dehydrogenase (LDH), antistreptolysin O (ASO), cystatin C (Cys-C), cholyglycine (CG), ischemia-modified albumin (IMA), prothrombin time (PT), activated partial thromboplastin time (APTT), fibrinogen (Fbg), thrombin time (TT), D-dimer (DD), white blood cell (WBC) count, platelet (PLT) count, CRP, total bile acid (TBA), adenosine deaminase (ADA), glutamate dehydrogenase (GLDH), homocysteine (HCY), non-esterified fatty acids (NEFA), and sialic acid (SA). To further quantify the systemic immune-inflammatory homeostasis in pediatric patients with IgAV, four higher-order composite indices were derived from primary blood cell counts and inflammatory markers: MLR, CRP-to-lymphocyte ratio (CLR), neutrophil-monocyte-lymphocyte derived ratio (NMLR), and IBI.

During the machine learning modelling phase, all data preprocessing steps were performed exclusively on the training set to prevent data leakage. For continuous variables with a missingness rate of less than 20%, imputation was performed using the k-nearest neighbours (KNN) algorithm to preserve the original feature distribution optimally. To mitigate the impact of dimensional and magnitude discrepancies among parameters on model training, all continuous features underwent Z-score standardisation (mean = 0, standard deviation = 1). At the same time, categorical variables were transformed via one-hot encoding. The implementation of this uniform standardisation protocol across the entire dataset neutralised potential batch effects and systematic baseline variations stemming from divergent laboratory assays between the two participating centres, ensuring the homogeneity and comparability of the pooled data. Following these procedures, a final high-dimensional feature matrix was constructed for subsequent feature selection, machine learning model training, performance evaluation, and SHAP interpretability analysis.

Machine Learning Model Construction

To identify the optimal algorithm for predicting the progression from IgAV to IgAVN, the 509 eligible patients were randomly partitioned into a training set (n = 356) and an independent validation set (n = 153) at a 7:3 ratio using a stratified sampling strategy. This stratification ensured that the proportional distribution of positive (progression to IgAVN) and negative outcomes remained consistent across both cohorts, thereby establishing a robust foundation for model training and reliable performance evaluation. To strictly prevent data leakage, all data preprocessing steps—including missing-value imputation, standardisation, categorical-variable encoding, and feature selection—were fitted exclusively on the training set, and the resulting transformation parameters were subsequently applied to the validation set. The optimal hyperparameter configurations for each evaluated model are detailed in Table 1.

Table 1
Optimal parameter combination for machine learning models

Model	Optimal hyperparameter combination
Decision tree (DT)	ccp_alpha = 0.0; max_depth = 10; max_features = None; min_samples_split = 20
Random forest (RF)	n_estimators = 50; max_features = 2
Extreme Gradient Boosting (XGBoost)	learning_rate = 0.1; max_depth = 3; n_estimators = 100; subsample = 0.6
Light Gradient Boosting Machine (LightGBM)	colsample_bytree = 0.6; learning_rate = 0.1; n_estimators = 100; num_leaves = 31; subsample = 0.6
Support vector machine (SVM)	C = 0.1; degree = 2; gamma = scale; kernel = linear
Artificial neural network (ANN)	activation = relu; hidden_layer_sizes = (50,)

During the model development phase, seven supervised classification algorithms, spanning traditional linear statistical models to complex tree-based ensembles and nonlinear networks, were trained in parallel and systematically evaluated on the training set. These algorithms included logistic regression (LR), decision tree (DT), random forest (RF), XGBoost, light gradient boosting machine (LightGBM), support vector machine (SVM), and artificial neural network (ANN). To mitigate the risk of overfitting and optimise predictive performance, hyperparameter tuning for the tree-based models, SVM, and ANN was conducted using 10-fold cross-validation and a grid search. The LR model was constructed using default parameters. Finally, the models with their optimised hyperparameters were deployed on the independent validation set to provide an unbiased assessment of their generalisation performance.

Evaluation of Model Performance and Clinical Utility

To comprehensively evaluate the predictive performance of the machine learning models in the independent validation set, we established a multidimensional evaluation framework. Initially, the overall discrimination of each model was quantified by plotting receiver operating characteristic (ROC) curves and calculating the area under the curve (AUC) with corresponding 95% confidence intervals (CIs). Subsequently, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), Youden’s J index, and Kappa value were calculated. These metrics, in conjunction with the F1 score, were utilised to systematically compare the classification performance of the various algorithms in identifying high-risk patients and ruling out low-risk cases.

Following confirmation of discriminative ability, the models' goodness of fit and clinical translational potential were further assessed. Calibration curves, coupled with Brier scores, were used to visualise and quantify the agreement (calibration) between predicted probabilities and the observed incidence of IgAVN. Furthermore, to address the critical issue of intervention thresholds in pediatric clinical practice, decision curve analysis (DCA) was utilised to calculate the clinical net benefit across a wide range of threshold probabilities. This analysis evaluated the practical utility of the models in balancing the risks of overtreatment (e.g., unnecessary immunosuppressive exposure) against those of under-intervention (e.g., missing the optimal therapeutic window to arrest disease progression). Ultimately, this multidimensional evaluation facilitated a rigorous comparison of the overall performance, robustness, and clinical feasibility of each model, providing a robust evidence base for selecting the optimal predictive tool.

Statistical Analysis and Model Interpretation

All baseline statistical analyses were conducted using R software (version 4.2.2). Normally distributed continuous variables were expressed as the mean ± standard deviation (SD) and compared using the independent-samples t-test. Non-normally distributed continuous variables were presented as medians and interquartile ranges (IQRs) and compared using the Wilcoxon rank-sum test. Categorical variables were summarised as frequencies and proportions (n, %) and compared using the Pearson chi-square test or Fisher’s exact test, as appropriate. Statistical significance was defined as a two-sided $P < 0.05$.

Machine learning model development and interpretability analyses were implemented in Python (version 3.10.4). Following a comprehensive evaluation of discrimination, classification performance, calibration, and clinical net benefit across all models in the validation set, the XGBoost model—yielding the optimal overall performance—was selected for SHAP analysis. Specifically, a SHAP bar plot was utilised to identify the core predictive features. A SHAP beeswarm plot was then employed to visualise the directional impact of individual feature values on the risk of IgAVN. Furthermore, SHAP dependence plots were constructed to quantify nonlinear associations and potential risk-threshold effects of the core continuous variables.

Results

Patient characteristics

A total of 509 pediatric patients were included in this study, comprising 296 in the non-IgAVN group and 213 in the IgAVN group. Patients in the IgAVN group were significantly older than those in the non-IgAVN group ($P = 0.021$). As detailed in Table 2, the IgAVN group exhibited significantly higher levels of ALT, urea, CREA, UA, ASO, Cys-C, Fbg, WBC, MLR, NMLR, CLR, IBI, PLT, and CRP, alongside significantly lower ALB levels, compared with the non-IgAVN group (all $P < 0.05$). Furthermore, both PT and APTT were significantly shorter in the IgAVN group ($P = 0.001$ for both). No statistically significant differences were observed between the two groups regarding sex distribution or the remaining baseline parameters (all $P > 0.05$).

Table 2
Baseline characteristics.

Variables	Total(N = 509)	No-IgAVN(N = 296)	IgAVN(N = 213)	P-Value
Sex (%)				0.216
Female	250 (49.12)	138 (46.62)	112 (52.58)	
Male	259 (50.88)	158 (53.38)	101 (47.42)	
Age (years)	8.00 [7.00, 11.00]	8.00 [7.00, 10.00]	9.00 [7.00, 11.00]	0.021
ALT (U/L)	10.00 [9.00, 11.00]	9.00 [8.00, 10.00]	10.00 [9.00, 11.00]	< 0.001
ALP (U/L)	182.00 [142.00, 224.00]	182.00 [147.00, 226.25]	179.00 [138.00, 217.00]	0.225
GGT (U/L)	16.00 [12.00, 19.00]	16.00 [12.00, 19.00]	15.00 [11.00, 19.00]	0.710
TP (g/L)	68.77 [65.48, 71.52]	68.83 [65.56, 71.37]	68.74 [65.37, 71.68]	0.971
ALB (g/L)	39.69 [38.44, 41.76]	40.27 [38.88, 42.29]	38.91 [36.72, 40.11]	< 0.001
UREA (mmol/L)	3.59 ± 0.24	3.52 ± 0.22	3.69 ± 0.23	< 0.001
CREA (μmol/L)	35.49 [33.70, 37.21]	35.01 [33.15, 36.47]	36.42 [34.58, 38.59]	< 0.001
UA (μmol/L)	238.00 [217.00, 257.00]	230.50 [208.00, 250.25]	249.00 [230.00, 275.00]	< 0.001
TCO2 (mmol/L)	22.27 [21.23, 23.84]	22.29 [21.08, 23.96]	22.20 [21.26, 23.83]	0.834
IgA (g/L)	1.72 [1.65, 1.79]	1.72 [1.64, 1.78]	1.73 [1.66, 1.80]	0.091
IgG (g/L)	10.87 [9.38, 12.23]	10.91 [9.43, 12.19]	10.80 [9.23, 12.40]	0.977
IgM (g/L)	1.31 [1.13, 1.59]	1.32 [1.09, 1.59]	1.28 [1.15, 1.58]	0.968
C3 (g/L)	1.37 [1.29, 1.46]	1.37 [1.29, 1.46]	1.38 [1.29, 1.46]	0.982
C4 (g/L)	0.33 [0.26, 0.38]	0.33 [0.26, 0.37]	0.33 [0.28, 0.38]	0.475
LDH (U/L)	181.00 [172.00, 190.00]	181.00 [173.00, 190.25]	181.00 [172.00, 190.00]	0.360
ASO (IU/mL)	30.55 [23.48, 38.40]	26.06 [20.70, 34.58]	33.32 [26.35, 41.45]	< 0.001
Cys. C (mg/L)	0.81 ± 0.05	0.80 ± 0.04	0.83 ± 0.04	< 0.001
CG (μg/mL)	1.02 [0.78, 1.19]	1.02 [0.79, 1.22]	1.02 [0.78, 1.17]	0.430
IMA (U/mL)	70.64 [70.38, 70.94]	70.65 [70.38, 70.98]	70.62 [70.38, 70.92]	0.708
PT (s)	11.47 [11.39, 11.56]	11.48 [11.41, 11.58]	11.45 [11.38, 11.54]	0.001
APTT (s)	30.15 [29.89, 30.51]	30.20 [29.92, 30.58]	30.10 [29.82, 30.41]	0.001
Fbg (g/L)	2.46 [2.36, 2.59]	2.42 [2.34, 2.49]	2.57 [2.43, 2.71]	< 0.001
TT (s)	14.89 [14.14, 16.08]	14.89 [14.16, 16.08]	14.87 [14.11, 16.07]	0.938
DD (mg/L)	0.25 ± 0.03	0.25 ± 0.03	0.25 ± 0.03	0.154
WBC (×10□/L)	6.02 [5.69, 6.44]	5.82 [5.55, 6.12]	6.37 [5.96, 6.85]	< 0.001
MLR	0.15 [0.12, 0.20]	0.13 [0.11, 0.17]	0.19 [0.15, 0.23]	< 0.001

ALT, alanine aminotransferase; ALP, alkaline phosphatase; GGT, gamma-glutamyl transferase; TP, total protein; ALB, albumin; UREA, urea; CREA, creatinine; UA, uric acid; TCO2, total carbon dioxide; IgA, immunoglobulin A; IgG, immunoglobulin G; IgM, immunoglobulin M; C3, complement 3; C4, complement 4; LDH, lactate dehydrogenase; ASO, antistreptolysin O; Cys. C, cystatin C; CG, cholylglycine; IMA, ischemia-modified albumin; PT, prothrombin time; APTT, activated partial thromboplastin time; Fbg, fibrinogen; TT, thrombin time; DD, D-dimer; WBC, white blood cell count; MLR, monocyte-to-lymphocyte ratio; NMLR, neutrophil-monocyte-to-lymphocyte ratio; CLR, C-reactive protein-to-lymphocyte ratio; IBI, inflammatory burden index; PLT, platelet count; CRP, C-reactive protein; TBA, total bile acid; ADA, adenosine deaminase; GLDH, glutamate dehydrogenase; HCY, homocysteine; NEFA, non-esterified fatty acid; SA, sialic acid.

Variables	Total(N = 509)	No-IgAVN(N = 296)	IgAVN(N = 213)	P-Value
NMLR	3.65 [2.58, 4.49]	3.34 [2.07, 3.96]	4.16 [3.51, 5.32]	< 0.001
CLR	1.78 [1.36, 2.67]	1.65 [1.18, 2.54]	2.26 [1.51, 2.80]	< 0.001
IBI	6.41 [4.94, 8.66]	5.30 [4.25, 7.44]	8.02 [5.93, 10.25]	< 0.001
PLT (×10⁹/L)	230.99 ± 11.91	228.29 ± 10.51	234.74 ± 12.72	< 0.001
CRP (mg/L)	3.14 [2.04, 4.82]	2.48 [1.82, 3.52]	4.49 [2.82, 6.02]	< 0.001
TBA (μmol/L)	4.81 [3.28, 6.20]	4.79 [3.27, 6.18]	4.84 [3.32, 6.24]	0.822
ADA (U/L)	11.30 [9.93, 13.23]	11.23 [9.86, 13.23]	11.37 [10.02, 13.23]	0.654
GLDH (U/L)	2.11 [1.68, 2.57]	2.18 [1.70, 2.57]	2.06 [1.66, 2.55]	0.160
HCY (μmol/L)	7.19 [6.50, 7.83]	7.16 [6.44, 7.79]	7.21 [6.56, 7.91]	0.495
NEFA (mmol/L)	0.50 [0.40, 0.65]	0.50 [0.40, 0.65]	0.50 [0.41, 0.64]	0.946
SA (mg/dL)	53.87 [50.45, 59.81]	53.95 [50.74, 59.94]	53.74 [49.94, 59.50]	0.393

ALT, alanine aminotransferase; ALP, alkaline phosphatase; GGT, gamma-glutamyl transferase; TP, total protein; ALB, albumin; UREA, urea; CREA, creatinine; UA, uric acid; TCO2, total carbon dioxide; IgA, immunoglobulin A; IgG, immunoglobulin G; IgM, immunoglobulin M; C3, complement 3; C4, complement 4; LDH, lactate dehydrogenase; ASO, antistreptolysin O; Cys. C, cystatin C; CG, cholylglycine; IMA, ischemia-modified albumin; PT, prothrombin time; APTT, activated partial thromboplastin time; Fbg, fibrinogen; TT, thrombin time; DD, D-dimer; WBC, white blood cell count; MLR, monocyte-to-lymphocyte ratio; NMLR, neutrophil-monocyte-to-lymphocyte ratio; CLR, C-reactive protein-to-lymphocyte ratio; IBI, inflammatory burden index; PLT, platelet count; CRP, C-reactive protein; TBA, total bile acid; ADA, adenosine deaminase; GLDH, glutamate dehydrogenase; HCY, homocysteine; NEFA, non-esterified fatty acid; SA, sialic acid.

Selection of Modelling Variables

To further identify key features associated with the incidence of IgAVN, the least absolute shrinkage and selection operator (LASSO) regression and the Boruta algorithm were independently applied for feature selection in the training set. LASSO regression, coupled with 10-fold cross-validation, retained 18 candidate variables at the optimal penalty parameter corresponding to the 1-standard-error (1-SE) criterion (Fig. 1A). The coefficient path plot further demonstrated that as the penalisation parameter increased, the regression coefficients for the respective variables shrank continuously, ultimately approaching zero (Fig. 1B). Concurrently, the Boruta algorithm identified 17 critical variables, all of which were classified as confirmed features (Fig. 1C-D). Subsequently, the intersection of the results from both feature selection methods yielded 12 consensus variables—namely, ALT, ALB, urea, UA, ASO, APTT, MLR, NMLR, CLR, IBI, PLT, and CRP. These shared features were ultimately incorporated into the development of downstream machine-learning predictive models (Fig. 1E).

Model development and performance assessment

Utilising the 12 selected feature variables, we developed seven predictive models—LR, DT, RF, XGBoost, LightGBM, SVM, and ANN—and systematically compared their performance in the independent validation set. The results demonstrated that all seven models exhibited robust predictive capabilities, with XGBoost and LightGBM demonstrating the optimal overall performance. Although the LightGBM model achieved the highest AUC of 0.967 (95% CI: 0.941–0.988), the XGBoost model (AUC: 0.966, 95% CI: 0.935–0.990) achieved the highest accuracy (0.907), with substantial sensitivity (0.921) and specificity (0.898). Furthermore, XGBoost achieved an F1 score of 0.892, a Kappa value of 0.811, and a Youden’s J index of 0.818, indicating the best comprehensive classification performance. The RF model also performed admirably, with an AUC of 0.952 (95% CI: 0.914–0.984) and a specificity of 0.909. In contrast, the overall predictive efficacies of the DT, SVM, and ANN models were relatively inferior (Table 3).

Table 3

Comparative analysis of the performance outcomes across machine learning models. XGBoost, Extreme Gradient Boosting, LightGBM, Light Gradient Boosting Machine, SVM, Support Vector Machine, ANN, Artificial Neural Network, PPV, Predictive Value, NPV, Negative Predictive Value, AUC, Area Under the Curve.

Model	AUC	Accuracy	Precision	Sensitivity	Specificity	F1 Score	Kappa	Youden's J	PPV	NPV
Logistic	0.934	0.868	0.803	0.905	0.841	0.851	0.732	0.746	0.803	0.925
Decision Tree	0.885	0.854	0.797	0.873	0.841	0.833	0.704	0.714	0.797	0.902
Random Forest	0.952	0.887	0.871	0.857	0.909	0.864	0.768	0.766	0.871	0.899
XGBoost	0.966	0.907	0.866	0.921	0.898	0.892	0.811	0.818	0.866	0.94
LightGBM	0.967	0.894	0.841	0.921	0.875	0.879	0.785	0.796	0.841	0.939
SVM	0.93	0.841	0.791	0.841	0.841	0.815	0.676	0.682	0.791	0.881
ANN	0.923	0.841	0.82	0.794	0.875	0.806	0.672	0.669	0.82	0.856

XGBoost, Extreme Gradient Boosting, LightGBM, Light Gradient Boosting Machine, SVM, Support Vector Machine, ANN, Artificial Neural Network, PPV, Predictive Value, NPV, Negative Predictive Value, AUC, Area Under the Curve.

Confusion matrices from the validation set further confirmed that all models exhibited a degree of discriminative ability between patients with and without IgAVN (Fig. 2A-G). Notably, the XGBoost model correctly identified 58 patients with IgAVN and 79 patients in the non-IgAVN group, misclassifying only 5 and 9 cases, respectively; this overall classification outcome was superior to those of the other models (Fig. 2D). The LightGBM model also performed well, accurately identifying 58 IgAVN cases and 77 non-IgAVN cases (Fig. 2E). The RF model exhibited a strong capacity to recognise non-IgAVN patients, correctly classifying 80 such cases, reflecting its higher specificity (Fig. 2C).
ROC curve analysis revealed that all seven models maintained good discriminative ability in the validation set, with XGBoost and LightGBM encompassing the largest areas under the curve (Fig. 3A). Calibration curves indicated a general concordance between the predicted probabilities and the actual observed incidences across the models, with XGBoost, LightGBM, and the LR model demonstrating superior calibration performance (Fig. 3B). DCA results showed that all models yielded a net clinical benefit across most threshold probabilities; however, the XGBoost model achieved a higher overall net benefit (Fig. 3C). Taking into careful consideration the discriminative ability, classification performance, calibration, and DCA results, XGBoost exhibited a more balanced and robust profile across accuracy, sensitivity, F1 score, Kappa value, Youden’s J index, and overall net benefit. Consequently, XGBoost was ultimately selected as the optimal model for subsequent interpretability analysis and potential clinical deployment.

Interpretability analysis in the model

To quantify the contribution of each predictive variable to the model output, SHAP analysis was applied to the optimal XGBoost model. The results demonstrated that IBI, CRP, and MLR exhibited the highest mean absolute SHAP values (all 0.11), followed sequentially by ASO (0.06), NMLR (0.06), ALB (0.05), CLR (0.04), UA (0.03), urea (0.02), ALT (0.02), APTT (0.01), and PLT (0.01) (Fig. 4A). The SHAP beeswarm plot revealed that elevated levels of IBI, CRP, MLR, ASO, NMLR, CLR, UA, urea, and ALT were generally associated with an increased risk of incident IgAVN. In contrast, higher levels of ALB and APTT were correlated with a reduced risk (Fig. 4B).

At the individual level, SHAP waterfall plots for specific patients showed that the directional contributions and effect magnitudes of identical features differed substantially between the non-IgAVN and IgAVN outcomes, with MLR, ASO, CRP, and NMLR emerging as the pivotal variables that elucidated the model's predictions (Fig. 4C-D). Furthermore, SHAP dependence plots indicated pervasive nonlinear relationships between individual variables and the model output (Fig. 5). At lower levels, parameters including ALT, urea, UA, ASO, MLR, NMLR, CLR, IBI, and CRP exerted relatively limited impacts on the model output. However, as these indices increased, their corresponding SHAP values progressively increased, transitioning from negative to positive or showing continuous increases within the moderate-to-high ranges. Conversely, an increase in ALB levels corresponded with an overall decline in its respective SHAP value. APTT exhibited a similar negative trajectory, whereas PLT showed nonlinear fluctuations.

Web application deployment

To enhance the clinical accessibility and practical utility of the optimal XGBoost model, we deployed it as an online web-based prediction system (Fig. 6). Titled "Kidney Injury Risk Prediction System," this platform integrates individualised risk prediction and personalised SHAP explanation functionalities for specific pediatric patients. Upon inputting the required clinical parameters—namely, ALT, ALB, urea, UA, ASO, APTT, MLR, NMLR, CLR, IBI, PLT, and CRP—users can click "Run Prediction" to obtain the corresponding predicted probability. Concurrently, the system generates a personalised SHAP explanation that provides a visual representation of the predicted outcome and its key driving factors. Furthermore, the webpage includes a dedicated results panel and usage instructions, explicitly cautioning that the tool is intended strictly for scientific research and adjunctive evaluation, rather than as a substitute for professional clinical judgment. The online prediction system is freely accessible at: https://predictinglymphnodemetastasisingastriccancer.shinyapps.io/shiny_kidney_app/. Ultimately, these results demonstrate the successful online deployment of the proposed XGBoost model, providing a convenient tool for individualised risk assessment of IgAVN.

Discussion

In pediatric outpatient and emergency settings, the early progression from IgAV to IgAVN often lacks definitive, specific clinical manifestations at initial presentation. The insidious nature of disease evolution, coupled with pronounced clinical heterogeneity, predisposes high-risk patients to miss the critical therapeutic window for early identification and timely intervention. Concurrently, it increases the risk of unnecessary exposure to immunosuppressive therapies among low-risk patients with a self-limiting disease course [21]. Given the invasive nature of renal biopsy and the inherent temporal lag of routine urinalysis in detecting early renal involvement, exploring noninvasive, objective, and clinically feasible early predictive strategies is of profound practical significance. Driven by this clinical imperative, the present study developed and validated multiple machine-learning predictive models using routine peripheral blood parameters readily available at initial presentation. Our findings revealed that the XGBoost model demonstrated exceptional discrimination and calibration within the independent validation set, underscoring its superior stability and generalisation potential. Furthermore, subsequent SHAP interpretability analysis elucidated that composite inflammatory burden indices—specifically IBI and MLR—alongside ALB and ASO, serve as pivotal predictors intricately associated with the progression from IgAV to IgAVN.

Historically, risk assessment for IgAVN has predominantly relied upon traditional clinical scoring systems or statistical models such as logistic regression [22, 23]. However, the onset and progression of pediatric IgAV involve complex immune-inflammatory responses and metabolic remodelling processes. Consequently, routine peripheral blood indices are rarely independent; instead, they frequently exhibit substantial multicollinearity and higher-order nonlinear interactions [24–26]. Employing a two-sample Mendelian randomisation approach, Xie et al. systematically analysed the causal associations among IgAV, immune cells, metabolites, and inflammatory cytokines, revealing that the pathogenesis of IgAV is not a solitary inflammatory event, but rather a complex biological cascade co-orchestrated by immune-inflammatory responses and metabolic networks [27]. Within a modelling framework predicated on linear assumptions, such intricate relationships are difficult to adequately capture, potentially impairing the model's capacity to extract key information and, in turn, compromising predictive accuracy. Our findings demonstrate that the XGBoost algorithm, leveraging a tree-based ensemble strategy, can effectively capture complex nonlinear association patterns among features without requiring the a priori manual elimination of collinear variables. Compared with the traditional logistic regression model, XGBoost exhibited superior overall predictive performance while achieving a more optimal balance between sensitivity and specificity. Furthermore, its lower Brier score indicates that the model not only possesses robust discriminative capacity but also generates individualised predicted probabilities that are highly concordant with the actual observed risk.

By leveraging the capacity of tree-based models to capture high-dimensional interactions, our SHAP analysis not only enhances model interpretability but also delineates the critical pathological sequence of IgAV: triggering by prodromal infections, amplification via immune dysregulation, and culmination in microvascular endothelial injury. The high predictive importance of ASO supports the classical hypothesis that prodromal infections (e.g., Streptococcus) induce aberrant immune responses and pathogenic IgA1 production, thereby facilitating immune complex deposition [28]. Furthermore, the prominence of composite inflammatory indices, such as IBI and MLR, suggests that systemic inflammatory activation is a primary driver of disease progression. Mechanistically, this aligns with Li et al., who

demonstrated that IgA1-containing immune complexes activate neutrophils via FcαRI, precipitating endothelial damage. This highlights the vital role of innate immune hyperactivation in mesangial injury and systemic microvascular destruction [29].

Notably, ALB demonstrated a strong inverse association with IgAVN risk in the SHAP analysis. This indicates that subtle albumin leakage or depletion—resulting from early systemic endothelial barrier impairment—may act as an early warning sign of disease progression before overt target-organ damage (e.g., massive proteinuria) manifests. This finding is consistent with Shen et al., who identified baseline albumin as an independent predictor of renal involvement in pediatric IgAV, implying that an early decline in ALB reflects a latent risk of IgAVN [30]. Finally, SHAP dependence plots revealed nonlinear associations and distinct threshold effects between these key predictors and IgAVN risk. The risk of disease progression escalates nonlinearly once the systemic inflammatory burden exceeds the body's compensatory limits or when an early reduction in ALB occurs. These threshold effects underscore the complex, phased nature of IgAVN pathogenesis, providing clinicians with actionable metrics to identify high-risk patients nearing physiological decompensation.

Crucially, the XGBoost model relies entirely on features derived from standardised laboratory tests (e.g., routine inflammatory markers, ALB, and composite immune indices) universally available in outpatient and emergency settings. This noninvasive, readily accessible biomarker panel mitigates the risks associated with early renal biopsy and overcomes the inherent temporal delay of routine urinalysis in detecting systemic microvascular inflammation [31, 32].

Furthermore, decision curve analysis (DCA) substantiated the model's clinical utility, demonstrating a superior net clinical benefit across a broad range of threshold probabilities compared to default "treat-all" or "treat-none" strategies. This underscores the model's efficacy as a robust clinical decision-support tool. By providing frontline clinicians with an objective basis for early risk stratification, the model optimises the clinical management of pediatric IgAV. Particularly in resource-constrained environments, it empowers clinicians to strike an optimal balance—promptly initiating targeted interventions for patients at high risk of renal progression while sparing low-risk patients from unnecessary immunosuppressive exposure.

Despite its robust predictive performance and clinical promise, these findings must be interpreted with caution. First, the retrospective design and limited sources of data introduce inherent selection bias. Furthermore, the current reliance on internal validation necessitates the use of independent external cohorts to rigorously verify the model's stability and generalizability across diverse geographic regions, healthcare tiers, and heterogeneous populations. In conclusion, by integrating early routine peripheral blood features with the XGBoost algorithm and SHAP framework, this study systematically delineates the nonlinear risk trajectories of IgAV progression to renal involvement. Ultimately, this model provides a noninvasive, objective, and highly actionable decision-support tool. It empowers frontline clinicians to promptly identify high-risk pediatric patients, optimise risk stratification, and maximise the efficiency of medical resource allocation.

Conclusion

In conclusion, this study developed an interpretable XGBoost model based on routine non-invasive peripheral blood parameters that enables accurate early identification of pediatric patients at high risk of progression from IgAV to IgAVN, while revealing nonlinear associations between key predictors and disease progression. This model provides a quantitative basis for early clinical risk stratification, with potential to optimise the timing of intervention and reduce unnecessary immunosuppressive exposure in low-risk children. Although further validation in large prospective cohorts is warranted, it shows promising potential as a visual decision-support tool for clinical translation.

Abbreviations

Abbreviation	Full term
ACR	albumin-to-creatinine ratio
ADA	adenosine deaminase
ALB	albumin
ALP	alkaline phosphatase
ALT	alanine aminotransferase
ANN	artificial neural network
CLR	C-reactive protein-to-lymphocyte ratio
HCY	homocysteine
HPF	high-power field
HSP	Henoch-Schönlein purpura
IBI	Inflammatory Burden Index
LDH	lactate dehydrogenase
LR	logistic regression
ML	machine learning
MLR	monocyte-to-lymphocyte ratio
NEFA	non-esterified fatty acids
NLR	neutrophil-to-lymphocyte ratio
NMLR	neutrophil-monocyte-lymphocyte derived ratio
PPV	positive predictive value
ROC	receiver operating characteristic
SA	sialic acid
SVM	support vector machine
TCO2	total carbon dioxide
TP	total protein
TT	thrombin time
UA	uric acid
UPCR	urinary protein-to-creatinine ratio
UREA	urea
XGBoost	eXtreme Gradient Boosting

Declarations

Ethics approval and consent to participate

All procedures involving human participants were conducted in accordance with the ethical standards of the institutional research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Ethical approval for this retrospective study was obtained from the Medical Ethics Committee of Siyang Hospital (approval number: KS2026002). The requirement for informed consent to participate was waived by the Medical Ethics Committee of Siyang Hospital due to the retrospective nature of the study.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are not publicly available due to institutional and patient privacy restrictions but are available from the corresponding author on reasonable request

Competing Interests

The authors declare they have no conflict of interest.

Funding

This study was jointly funded by Siyang Hospital and the Suqian Municipal Health Commission, as part of the 2024 Suqian Municipal Health Commission Medical Research Project (ZD202409).

Authors' contributions

Conceptualisation: Qingkai Wang, Hao Qiu, Weibing Qiu; Data curation: Liran Shen, Fachen Miao, Jinxing Dai; Formal analysis: Qingkai Wang, Hao Qiu, Jinxing Dai; Methodology: Kang Shen; Software: Qingkai Wang, Jinxing Dai; Supervision: Qingkai Wang, Yubiao Zhang, Liran Shen, Weibing Qiu; Writing - original draft: Jinxing Dai, Qingkai Wang; Writing - review & editing: Hao Qiu, Weibing Qiu. All authors have read and approved the manuscript.

Acknowledgements

None.

References

1. Castañeda S, Quiroga-Colina P, Floranes P, Uriarte-Ecenarro M, Valero-Martínez C, Vicente-Rabaneda EF, et al. IgA Vasculitis (Henoch-Schönlein Purpura): An Update on Treatment. J Clin Med. 2024;13(21). https://doi.org/10.3390/jcm13216621.
2. Sestan M, Jelusic M. Diagnostic and Management Strategies of IgA Vasculitis Nephritis/Henoch-Schönlein Purpura Nephritis in Pediatric Patients: Current Perspectives. Pediatr Health Med Ther. 2023;14:89–98. https://doi.org/10.2147/phmt.S379862.
3. Oni L, Platt C, Marlais M, McCann L, Barakat F, Hesseling M, et al. National recommendations for the management of children and young people with IgA vasculitis: a best available evidence, group agreement-based approach. Arch Dis Child. 2024;110(1):67–76. https://doi.org/10.1136/archdischild-2024-327364.
4. Jelusic M, Sestan M, Giani T, Cimaz R. New Insights and Challenges Associated With IgA Vasculitis and IgA Vasculitis With Nephritis-Is It Time to Change the Paradigm of the Most Common Systemic Vasculitis in Childhood? Front Pediatr. 2022;10:853724. https://doi.org/10.3389/fped.2022.853724.
5. Amatruda M, Carucci NS, Chimenz R, Conti G. Immunoglobulin A vasculitis nephritis: Current understanding of pathogenesis and treatment. World J Nephrol. 2023;12(4):82–92. https://doi.org/10.5527/wjn.v12.i4.82.
6. Gage A, Pepper RJ, Marro J, Salama AD, Oni L. IgA Vasculitis Across the Ages: Is It Time for a Precision Medicine Approach? ACR Open Rheumatol. 2025;7(9):e70083. https://doi.org/10.1002/acr2.70083.
7. Marro J, Chetwynd AJ, Wright RD, Dliso S, Oni L. Urinary Protein Array Analysis to Identify Key Inflammatory Markers in Children with IgA Vasculitis Nephritis. Child (Basel). 2022;9(5). https://doi.org/10.3390/children9050622.
8. Zhang Q, Lai LY, Cai YY, Wang MJ, Ma G, Qi LW, et al. Serum-Urine Matched Metabolomics for Predicting Progression of Henoch-Schonlein Purpura Nephritis. Front Med (Lausanne). 2021;8657073. https://doi.org/10.3389/fmed.2021.657073.
9. Oni L, Sampath S. Childhood IgA Vasculitis (Henoch Schonlein Purpura)-Advances and Knowledge Gaps. Front Pediatr. 2019;7:257. https://doi.org/10.3389/fped.2019.00257.
10. Xu S, Han S, Dai Y, Wang L, Zhang X, Ding Y. A Review of the Mechanism of Vascular Endothelial Injury in Immunoglobulin A Vasculitis. Front Physiol. 2022;13:833954. https://doi.org/10.3389/fphys.2022.833954.

11. Bi Y, Quan W, Hao W, Sun R, Li L, Jiang C, et al. A simple nomogram for assessing the risk of IgA vasculitis nephritis in IgA vasculitis Asian pediatric patients. Sci Rep. 2022;12(1):16809. https://doi.org/10.1038/s41598-022-20369-3.
12. Özdemir ZC, Çetin N, Kar YD, Öcal HO, Bilgin M, Bör Ö. Hemotologic Indices for Predicting Internal Organ Involvement in Henoch-Schönlein Purpura (IgA vasculitis). J Pediatr Hematol Oncol. 2020;42(1):e46–9. https://doi.org/10.1097/mph.0000000000001571.
13. Oh MY, Kim HS, Jung YM, Lee HC, Lee SB, Lee SM. Machine Learning-Based Explainable Automated Nonlinear Computation Scoring System for Health Score and an Application for Prediction of Perioperative Stroke: Retrospective Study. J Med Internet Res. 2025;27e58021. .https://doi.org/10.2196/58021.
14. Lee S, Kisiel MA, Lindberg P, Wheelock ÅM, Olofsson A, Eriksson J, et al. Using machine learning involving diagnoses and medications as a risk prediction tool for post-acute sequelae of COVID-19 (PASC) in primary care. BMC Med. 2025;23(1):251. https://doi.org/10.1186/s12916-025-04050-w.
15. Xie Y, Chen Y, Han Y, Zhai S, Xiao L, Yin D, et al. Identifying influencing factors associated with sleep quality in undergraduates based on partial least squares regression and XGBoost. Front Psychol. 2025;16:1732946. https://doi.org/10.3389/fpsyg.2025.1732946.
16. Yang S, Cao L, Zhou Y, Hu C. A Retrospective Cohort Study: Predicting 90-Day Mortality for ICU Trauma Patients with a Machine Learning Algorithm Using XGBoost Using MIMIC-III Database. J Multidiscip Healthc. 2023;16:2625–40. https://doi.org/10.2147/jmdh.S416943.
17. Luo S, Lai J, Mo L, Shen X, Fang R. Prediction of hospital mortality in sepsis-associated acute kidney injury using a machine-learning approach: a multicenter study using SHAP interpretability analysis. Clin Kidney J. 2026;19(1):sfaf372. .https://doi.org/10.1093/ckj/sfaf372.
18. Zhang C, Niu B, Wang R, Zhang L. From traditional metabolic markers to ensemble learning: comparative application of machine learning models for predicting NAFLD risk in adolescents. Front Endocrinol (Lausanne). 2025;16:1681686. https://doi.org/10.3389/fendo.2025.1681686.
19. Ozen S, Pistorio A, Iusan SM, Bakkaloglu A, Herlin T, Brik R, et al. EULAR/PRINTO/PRES criteria for Henoch-Schönlein purpura, childhood polyarteritis nodosa, childhood Wegener granulomatosis and childhood Takayasu arteritis: Ankara 2008. Part II: Final classification criteria. Ann Rheum Dis. 2010;69(5):798–806. https://doi.org/10.1136/ard.2009.116657.
20. KDIGO 2021 Clinical Practice Guideline for the Management of Glomerular Diseases. Kidney Int. 2021;100(4s):S1. s276.https://doi.org/10.1016/j.kint.2021.05.021.
21. Ercan Emreol H, Yildirim-Toruner C, Jelusic M, Twilt M, Ozen S. New avenues in childhood vasculitis. Pediatr Rheumatol Online J. 2025;23(1):97. https://doi.org/10.1186/s12969-025-01149-5.
22. Guo L, Zhu A, Li W, Zeng F, Wang F. Clinical prediction model for progression from henoch-schönlein purpura to nephritis in pediatric patients. Am J Transl Res. 2024;16(12):7385–95. https://doi.org/10.62347/xdor8531.
23. Podraza Z, Poplicha K, Ufniarski T, Ucieklak J, Łysiak N, Mizerska-Wasiak M. Laboratory Findings and Clinical Features in IgA Vasculitis: Identifying Predictors of Kidney Involvement and Disease Relapse in Pediatric Patients. J Clin Med. 2025;14(9). .https://doi.org/10.3390/jcm14093055.
24. Chen G, Yang Z. Risk prediction for gastrointestinal bleeding in pediatric Henoch-Schönlein purpura using an interpretable transformer model. Front Physiol. 2025;16:1630807. https://doi.org/10.3389/fphys.2025.1630807.
25. Pan M, Li M, Li N, Mao J. Predicting renal damage in children with IgA vasculitis by machine learning. Pediatr Nephrol. 2024;39(10):2997–3004. https://doi.org/10.1007/s00467-024-06432-3.
26. Cao T, Zhu Y, Zhu Y. Construction of Prediction Model of Renal Damage in Children with Henoch-Schönlein Purpura Based on Machine Learning. *Comput Math Methods Med* 2022, 2022:6991218.https://doi.org/10.1155/2022/6991218
27. Xie M, Zhou N, Liang Q, Lin Z, Yao Y. Impact of Immune Cells on IgA Vasculitis via Metabolites and Inflammatory Cytokines. J Clin Immunol. 2025;45(1):157. https://doi.org/10.1007/s10875-025-01946-3.
28. Held M, Kozmar A, Sestan M, Turudic D, Kifer N, Srsen S, et al. Insight into the Interplay of Gd-IgA1, HMGB1, RAGE and PCDH1 in IgA Vasculitis (IgAV). Int J Mol Sci. 2024;25(8). .https://doi.org/10.3390/ijms25084383.
29. Li Q, Wu H, Yuan X, Shi S, Liu L, Lv J, et al. FcαRI-mediated neutrophil activation contributed to the pathogenesis of adult IgA vasculitis with nephritis. Nephrol Dial Transpl. 2025. https://doi.org/10.1093/ndt/gfaf260.

30. Shen L, Miao L, Xu L. Risk factors associated with renal injury in patients initially diagnosed with IgA vasculitis. Front Pediatr. 2025;13:1584768. https://doi.org/10.3389/fped.2025.1584768.
31. Williams CEC, Ging H, Skoutelis N, Marro J, Roberts L, Chetwynd AJ, et al. Biomarkers to predict kidney outcomes in children with IgA vasculitis. Minerva Pediatr (Torino). 2025;77(3):256–71. https://doi.org/10.23736/s2724-5276.24.07715-2.
32. Jin Y, He X, Lin W, Peng Z, Li W, Xiang W, et al. Serum cytokine profiles in children with IgA vasculitis with nephritis. Biomol Biomed. 2025;25(6):1425–43. https://doi.org/10.17305/bb.2024.11081.

Figures

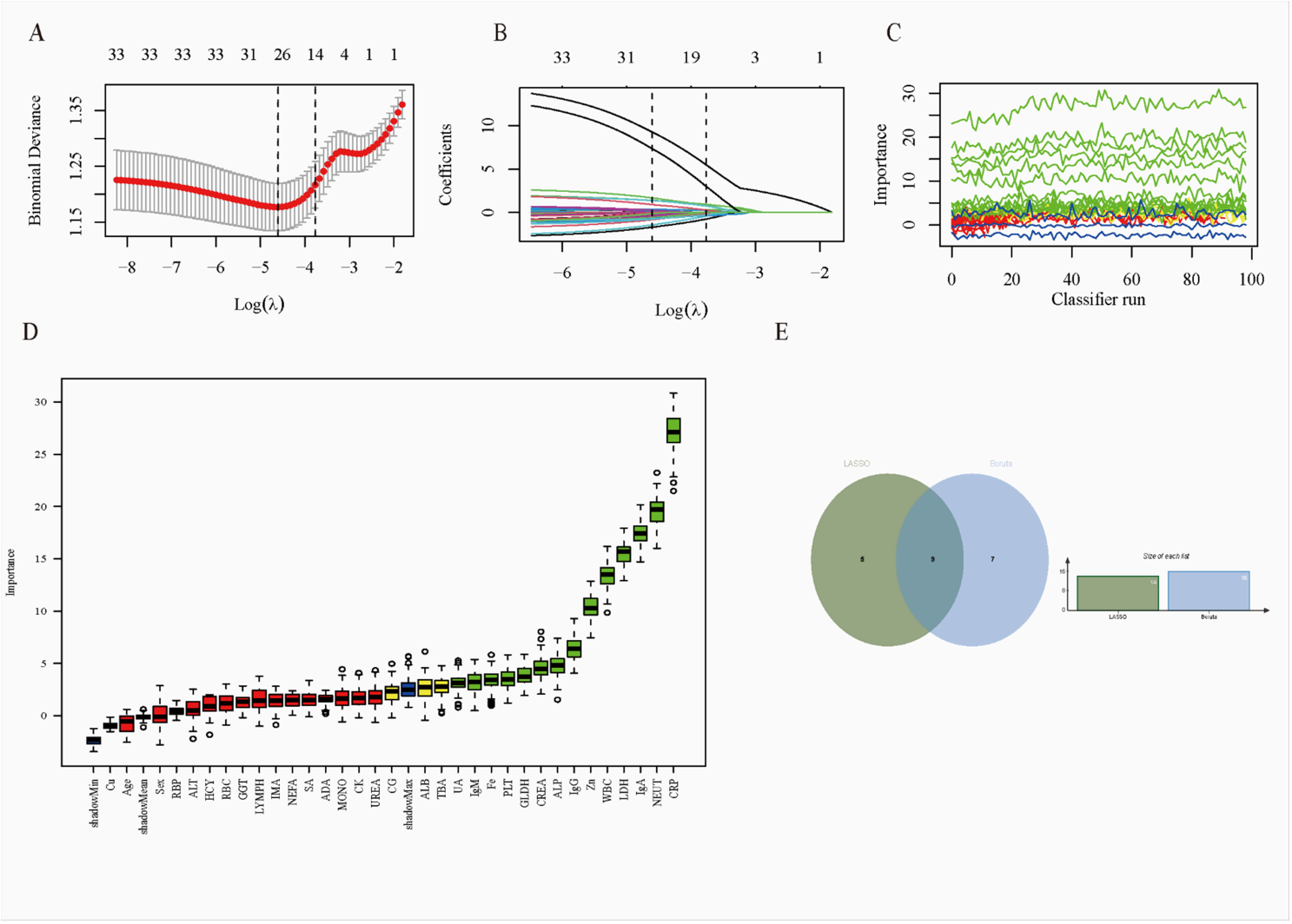


Figure 1

Feature selection process for model construction.
(A): Optimal penalty parameter λ identified by 10-fold cross-validation in the LASSO regression; (B): LASSO coefficient profiles plotted against log(λ); (C): Variable importance across 100 classifier runs in the Boruta algorithm; (D): Boxplots of feature importance generated by the Boruta algorithm; (E): Overlap analysis of variables selected by LASSO regression and the Boruta algorithm.

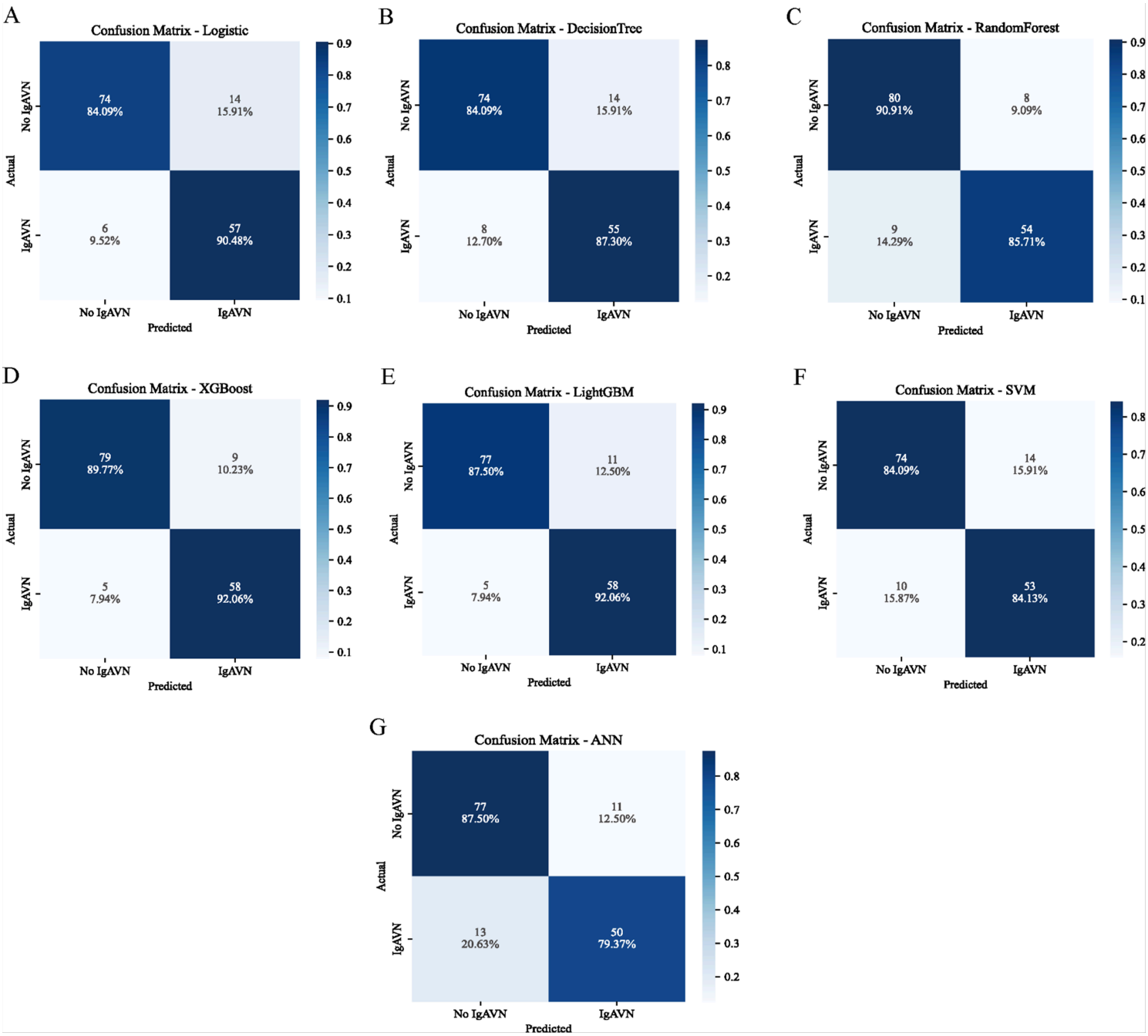


Figure 2

Confusion matrices of the seven models in the validation set.

(A): Logistic; (B): Decision Tree; (C): Random Forest; (D): XGBoost; (E): LightGBM; (F): SVM; (G): ANN. XGBoost, Extreme Gradient Boosting, LightGBM, Light Gradient Boosting Machine, SVM, Support Vector Machine, ANN, Artificial Neural Network.

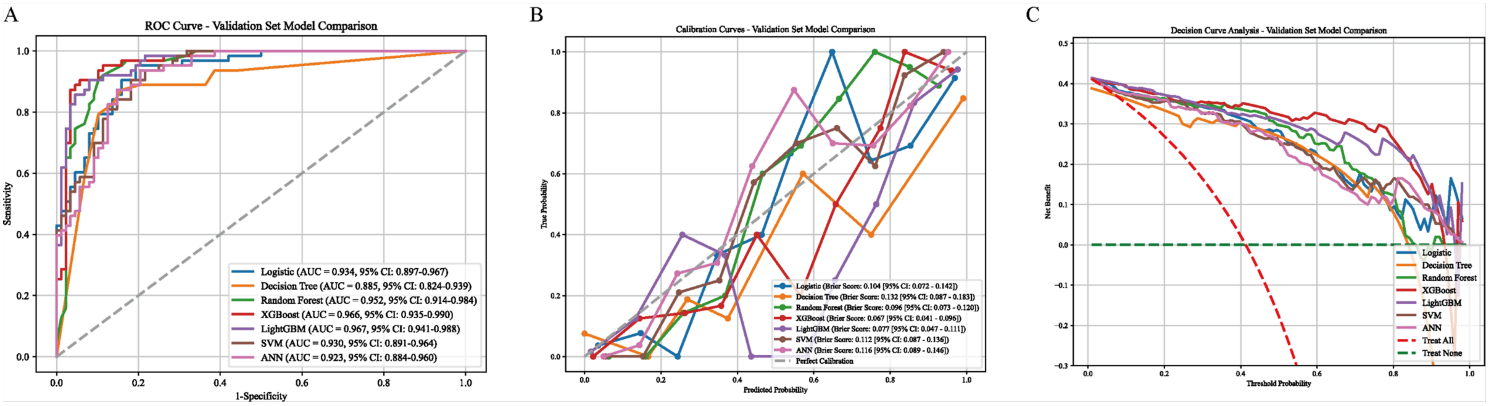


Figure 3

ROC curves, calibration curves, and decision curve analysis of the seven models in the validation set.
A: ROC curves of the seven models in the validation set; B: Calibration curves of the seven models in the validation set; C: Decision curve analysis of the seven models in the validation set.

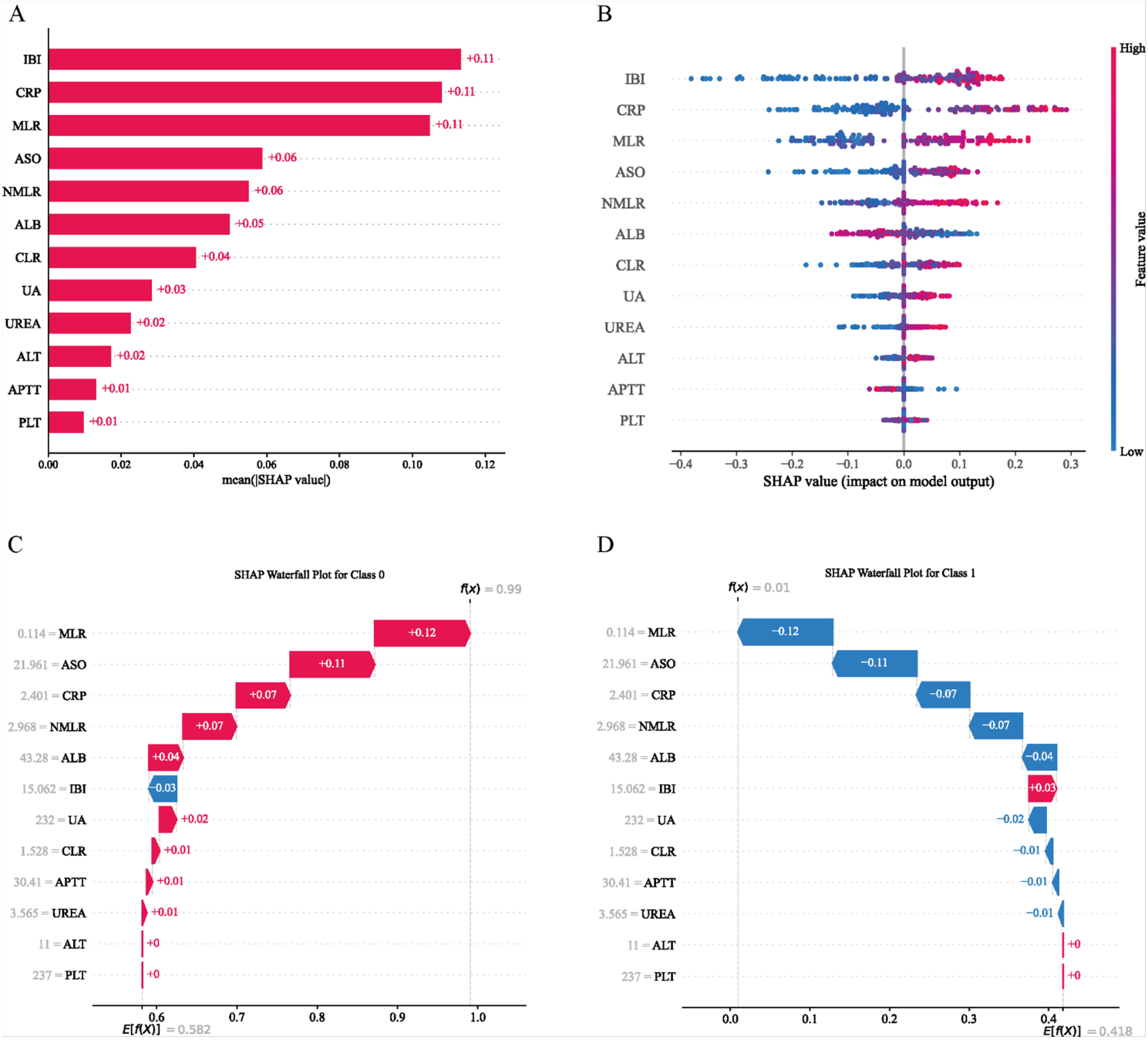


Figure 4

SHAP interpretation of the XGBoost model.
(A): Ranking of feature importance for the XGBoost model output; (B): SHAP beeswarm plot showing the relationships between feature values and model output; (C-D): SHAP waterfall plots of the same sample under different outcomes, C: no IgAVN (class0); D: IgAVN (class1).

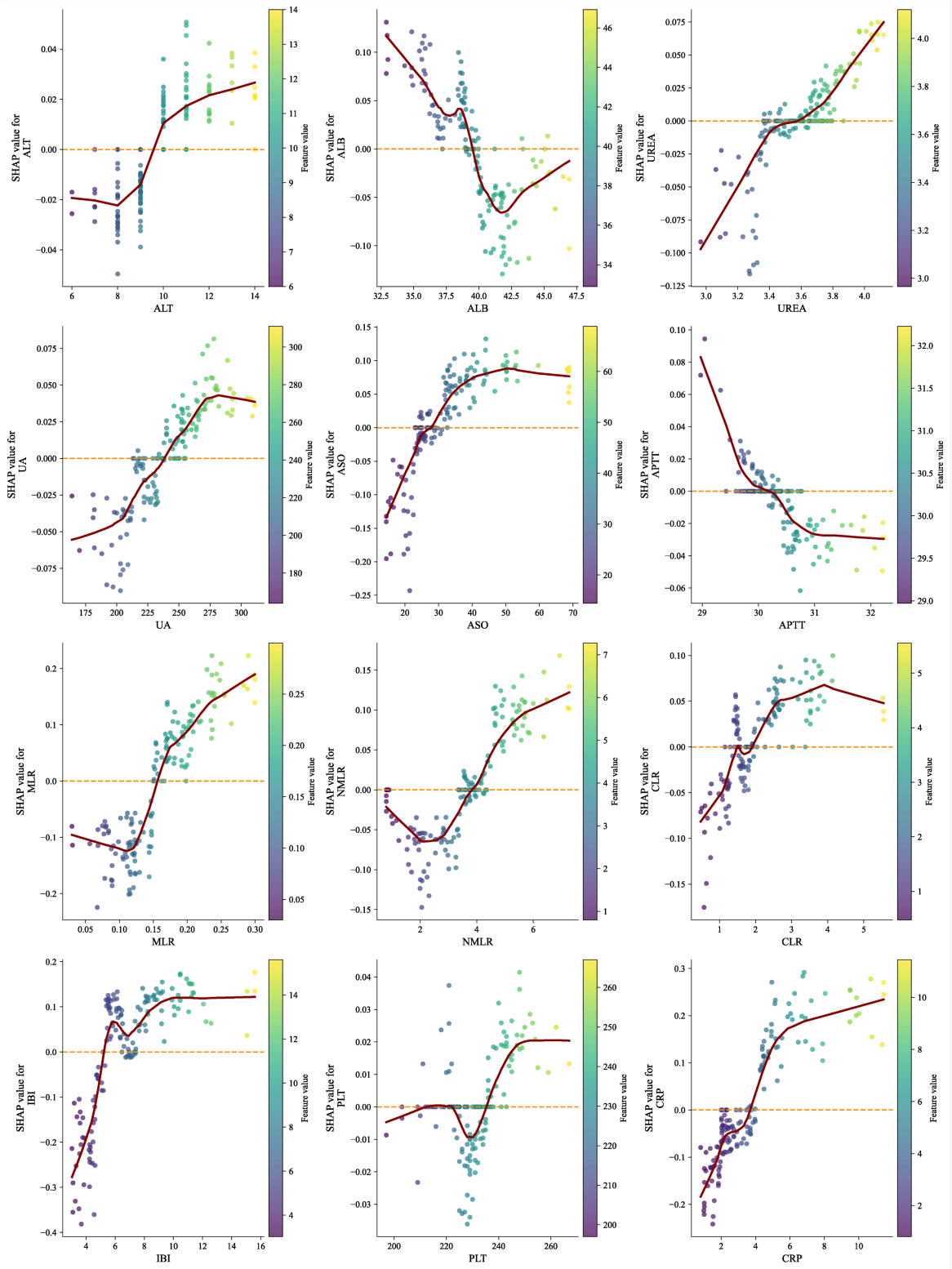


Figure 5

SHAP dependence plots showing the nonlinear relationships between predictors and model output in the XGBoost model.

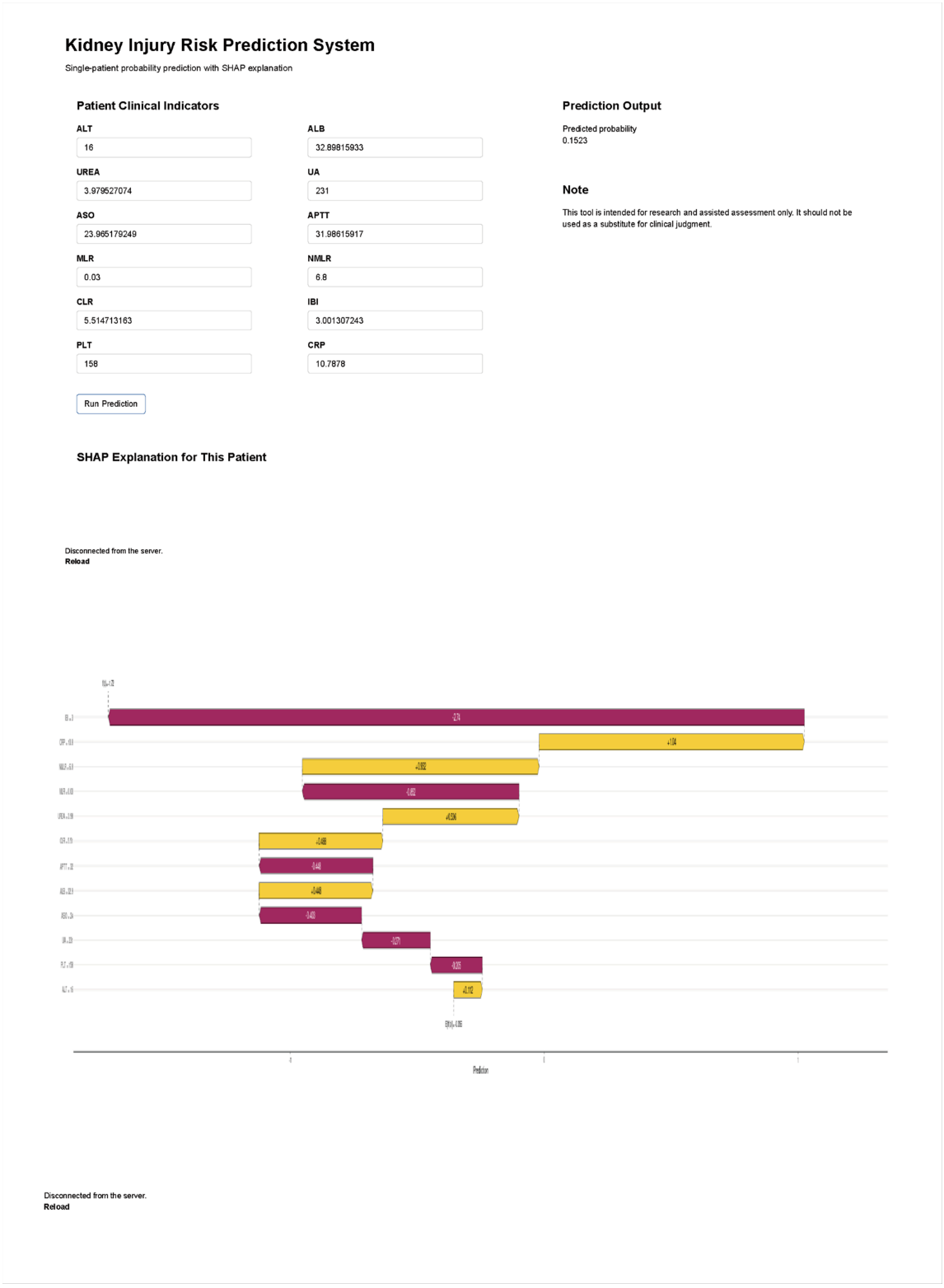


Figure 6

Interface of the online web-based prediction system.

The web-based system integrates IgAVN risk prediction with individualised SHAP interpretation, allowing users to obtain the predicted probability and corresponding explanations after entering relevant clinical variables.