Interpretable Machine Learning Model for Pulmonary hypertension Risk Prediction: a retrospective cohort study

Zhenshun Cheng Sr, Hongxia Jiang, Han Gao, Dexin Wang, Qingli Zeng, Xiaojun Hao

Submitted to: JMIR Medical Informatics
on: March 18, 2025

Table of Contents

Interpretable Machine Learning Model for Pulmonary hypertension Risk Prediction: a retrospective cohort study

Zhenshun Cheng Sr[1] PhD; Hongxia Jiang[2*]; Han Gao[2*]; Dexin Wang[3*]; Qingli Zeng[2]; Xiaojun Hao[2]

[1]Hubei Engineering Center for Infectious Disease Prevention, Control and Treatment, Chinese Academy of Medical Sciences Zhongnan Hospital of Wuhan University Wuhan CN
[2] Zhongnan Hospital of Wuhan University Wuhan CN
[3] Qichun County People's Hospital Qichun CN
[*]these authors contributed equally

Corresponding Author:
Zhenshun Cheng Sr PhD
Hubei Engineering Center for Infectious Disease Prevention, Control and Treatment,
Chinese Academy of Medical Sciences
Zhongnan Hospital of Wuhan University
Department of Respiratory and Critical Care Medicine, Zhongnan Hospital of Wuhan University, Wuhan, China.
Wuhan
CN

Abstract

Background: Pulmonary hypertension (PH) is a progressive disorder characterized by elevated pulmonary artery pressure and increased pulmonary vascular resistance, ultimately leading to right heart failure. Early detection is critical for improving patient outcomes.

Objective: To establish a novel machine learning-based diagnostic model for PH.

Methods: A diagnostic model for the early detection of pulmonary hypertension (PH) was developed through a two-step approach. First, Recursive Feature Elimination (RFE) was employed to select the most relevant echocardiographic variables, which were subsequently integrated into a composite ultrasound index using machine learning techniques such as XGBoost. In the second step, this ultrasound index was integrated with clinical variables identified through LASSO regression. Together, these elements were combined to construct a logistic regression model for diagnosis. The model's performance was rigorously evaluated using ROC curves, calibration plots, and decision curve analysis (DCA) to ensure its clinical relevance and accuracy.

Results: Machine learning identified key echocardiographic and clinical predictors, with the XGBoost model achieving high AUC, sensitivity, and specificity. LASSO regression identified critical clinical variables, including prothrombin time activity and serum cystatin C. The diagnostic model demonstrated high predictive accuracy, with an AUC of 0.997. Calibration and decision curve analyses indicated close alignment between predicted and observed outcomes validating the model's clinical value, especially at higher risk thresholds.

Conclusions: This model enhances early pulmonary hypertension (PH) diagnosis through a non-invasive approach and demonstrates strong predictive accuracy. It facilitates early intervention and personalized treatment, with potential applications in broader cardiovascular disease management. Clinical Trial: The study was approved by the Research Ethics Commission of Wuhan Zhongnan Hospital and the requirement for informed consent was waived by the Ethics Commission (Approval No.2023185)

(JMIR Preprints 18/03/2025:74117)
DOI: https://doi.org/10.2196/preprints.74117

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users. Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Interpretable Machine Learning Model for Pulmonary hypertension Risk Prediction: a retrospective cohort study

Hongxia Jiang [1#], Han Gao [1#], Dexin Wang [4#], Qingli Zeng [1], Xiaojun Hao [1], Zhenshun Cheng [123*]

1. Department of Respiratory and Critical Care Medicine, Zhongnan Hospital of Wuhan University, Wuhan, China.

2. Wuhan Research Center for Infectious Diseases and Cancer, Chinese Academy of Medical Sciences, Wuhan, China

3. Hubei Engineering Center for Infectious Disease Prevention, Control and Treatment, Wuhan, China

4. Department of Respiratory and Critical Care Medicine，Qichun County People's Hospital，Hubei，China

# These authors have contributed equally to this work and share first authorship.

*Corresponding author.

Correspondence: Zhenshun Cheng, E-mail: zhenshun_cheng@126.com

Abstract

Background: Pulmonary hypertension (PH) is a progressive disorder characterized by elevated pulmonary artery pressure and increased pulmonary vascular resistance, ultimately leading to right heart failure. Early detection is critical for improving patient outcomes.

Objective: To establish a novel machine learning-based diagnostic model for PH.

Methods: A diagnostic model for the early detection of pulmonary hypertension (PH) was developed through a two-step approach. First, Recursive Feature Elimination (RFE) was employed to select the most relevant echocardiographic variables, which were subsequently integrated into a composite ultrasound index using machine learning techniques such as XGBoost. In the second step, this ultrasound index was integrated with clinical variables identified through LASSO regression. Together, these elements were combined to construct a logistic regression model for diagnosis. The model's performance was rigorously evaluated using ROC curves, calibration plots, and decision curve analysis (DCA) to ensure its clinical relevance and accuracy.

Results: Machine learning identified key echocardiographic and clinical predictors, with the XGBoost model achieving high AUC, sensitivity, and specificity. LASSO regression identified critical clinical variables, including prothrombin time activity and serum cystatin C. The diagnostic model demonstrated high predictive accuracy, with an AUC of 0.997. Calibration and decision curve analyses indicated close alignment between predicted and observed outcomes validating the model's clinical value, especially at higher risk thresholds.

Conclusion: This model enhances early pulmonary hypertension (PH) diagnosis through a non-invasive approach and demonstrates strong predictive accuracy. It facilitates early intervention and personalized treatment, with potential applications in broader cardiovascular disease management.

Keywords: Pulmonary hypertension, machine learning, risk prediction, echocardiography, clinical data

Introduction

Pulmonary hypertension (PH) is a progressive and complex vascular disorder characterized by elevated pulmonary artery pressure and increased pulmonary vascular resistance, ultimately leading to right heart failure(1). Given its multifactorial etiology, which includes conditions such as idiopathic pulmonary arterial hypertension (IPAH), left heart disease, lung disease, chronic thromboembolic pulmonary hypertension (CTEPH), and other less common causes, early identification and risk assessment are critical for improving patient outcomes(2-5). PH is associated with a poor prognosis, as it significantly impairs right heart function and overall patient survival, making early detection essential(6).

Echocardiography, as a non-invasive imaging modality, has become integral to the evaluation of PH. It enables clinicians to assess right ventricular size and function, and estimate pulmonary artery pressures through tricuspid regurgitation velocity (TRV) and pulmonary artery systolic pressure (PASP). Several studies have highlighted the importance of echocardiographic parameters including right ventricular size, function (e.g., TAPSE), and TRV in predicting PH severity and outcomes(7-9). Moreover, comprehensive models such as the REVEAL-ECHO score, which integrate multiple echocardiographic variables, have significantly improved the ability to predict PH risk and have been validated across multiple cohorts(7, 10).

However, relying solely on echocardiographic data for PH diagnosis and risk prediction has limitations. Recent research has increasingly emphasized the need to integrate echocardiographic findings with clinical data, including age, sex, BMI, and comorbidities, to enhance predictive accuracy(11, 12). For instance, studies involving hyperthyroid patients demonstrated that models incorporating both clinical variables and echocardiographic parameters significantly improved PH risk assessment(13). Similarly, Nomogram models combining these data have proven effective in evaluating PH risk in postoperative patients and those with comorbidities such as obstructive sleep apnea(14, 15).

In recent years, the application of machine learning techniques to PH risk prediction has gained traction. Methods like recursive feature elimination (RFE) and advanced algorithms such as random forests and XGBoost have been shown to optimize feature selection, improving model robustness and prediction accuracy(15). These advancements, particularly when integrating echocardiographic and clinical data, have yielded models that outperform traditional methods in PH risk prediction(5).

In summary, echocardiography remains a cornerstone in PH assessment, but integrating clinical data and leveraging advanced statistical methods and machine learning techniques holds great promise for enhancing early detection and risk stratification. These combined approaches offer valuable insights into personalized treatment strategies, facilitating timely interventions and potentially improving clinical outcomes for patients with PH.

Materials and Methods

Study Population and Data Collection

This study retrospectively enrolled data from 294 patients with PH and 1231 control subjects who underwent echocardiographic examinations at Zhongnan Hospital of Wuhan University from January 2022 to April 2024. The inclusion criteria for the PH group were: (1) a discharge diagnosis of pulmonary hypertension; (2) confirmation of PH through right heart catheterization (RHC); and (3) age >14 years. Patients were excluded if: (1) the mean pulmonary arterial pressure (mPAP), as measured by RHC was <20 mmHg, or (2) the mPAP data from RHC were missing. The inclusion criteria for the control group were: (1) echocardiography results indicating no abnormalities in cardiac morphology, valve function, or ventricular wall motion; and (2) age >14 years. Control subjects were excluded if more than 5% of their data were missing. This study was approved by the Ethics Committee of Zhongnan Hospital of Wuhan University (Approval No. 2023185).

Data Collection and Variables

Data collected included demographic information (age and gender), echocardiographic parameters, and laboratory test results. Echocardiographic parameters assessed cardiac structure and function,

including measurements such as ascending aortic diameter, left atrial and left ventricular size, interventricular septal thickness, and pulmonary artery diameter. Additionally, the degree of regurgitation, flow velocity, and pressure gradients for the mitral, tricuspid, and aortic valves were recorded. All echocardiographic measurements were double-checked to ensure consistency and accuracy. Laboratory data included complete blood count, coagulation profiles, liver and kidney function, electrolytes, lipid profile, thyroid function, immune function, and arterial blood gas analysis.

Data Preprocessing

In the data preprocessing step, we used R packages (Version 1.4.3) to handle missing values through median imputation, replacing missing entries with the median of each variable to maintain data completeness. Continuous variables were centered and standardized, ensuring a mean of 0 and a standard deviation of 1 to meet the requirements of machine learning models. Categorical variables were converted to factors to ensure the accuracy of subsequent statistical analyses.

Feature Selection

We used stratified sampling to randomly split the data into training and testing sets in a 3:1 ratio. For selecting features from echocardiographic variables, we applied RFE using a Random Forest-based approach. RFE iteratively removed features with the least contribution to model performance, ultimately identifying the optimal subset of features. To ensure robustness and generalizability, we employed 10-fold cross-validation during feature selection.

Using the selected echocardiographic features, we trained multiple machine learning models, including logistic regression, LASSO regression, elastic net, decision tree, random forest, XGBoost, support vector machine (SVM), k-nearest neighbors (KNN), naive Bayes, and gradient boosting machine (GBM). Model training and evaluation were primarily conducted using the "caret" package(16). We used 10-fold cross-validation and grid search to optimize the hyperparameters. The primary performance metric was the Area Under the Receiver Operating Characteristic Curve (AUC-

ROC), which was computed and plotted using the "pROC" package. The model with the highest AUC-ROC was selected as the optimal model for the echocardiographic features. For the optimal model, we further performed internal and external validation using bootstrap methods to evaluate its robustness. To enhance model interpretability, Shapley Additive Explanations (SHAP) were used to analyze the contribution of each variable to the model's predictions(17). Finally, we developed an ultrasound index based on the optimal model in the training and test set.

For feature selection from routine clinical variables, we conducted LASSO regression analysis using the glmnet package. 10-fold cross-validation was used to select the optimal regularization parameter. Specifically, we calculated a range of lambda values and chose lambda.1se, which is the lambda value with the best performance in cross-validation, adjusted by one standard error.

Logistic Regression and Nomogram Construction

After identifying the optimal echocardiographic features and selected clinical variables, we combined them to develop a comprehensive logistic regression model. The model was fitted using the "lrm" function from the "rms" package. To prevent overfitting, we incorporated L2 regularization (ridge regression) as needed. Analysis of model coefficients provided insights into the contribution of each feature to predicting PH.

To facilitate the clinical application of the model, we constructed a nomogram using the "nomogram" function from the "rmsv package(18). This nomogram visualizes the logistic regression model in an easy-to-use format, allowing clinicians to estimate an individual's probability of developing PH by summing the scores for each predictive variable. The use of a nomogram ensures the interpretability and practicality of the model in a clinical setting.

Model Calibration and Decision Curve Analysis

To assess the performance of the nomogram and logistic regression model, we used the Receiver Operating Characteristic (ROC) curve. To evaluate the consistency between the predicted probabilities and observed outcomes, calibration curves were plotted using the "calibrate" function

from the "rms" package. Additionally, decision curve analysis (DCA) was conducted using the "rmda" package to quantify the net benefit of the model across different threshold probabilities, providing insight into its clinical utility(19, 20). Finally, we developed an online tool to facilitate the prediction of the risk of PH using R packages shiny (version 1.9.1).

Statistical Analysis

The SHAP analysis was conducted using SHAP（Version 0.46.0） in Python (version 3.10.8), while all other statistical analyses were performed using R（version 4.4.1， USA）. For continuous variables that followed a normal distribution, group comparisons were made using t-tests. For continuous variables that did not follow a normal distribution, non-parametric tests, such as the Mann-Whitney U test or the Kruskal-Wallis test, were applied. Categorical variables were presented as percentages and compared between groups using the Chi-square test. A $P < 0.05$ was considered statistically significant.

Results

Differences in Echocardiographic Parameters and Laboratory Findings between Pulmonary Hypertension Patients and Controls

This study ultimately included a cohort of 710 control subjects and 181 patients with PH. The findings revealed significant differences between the PH and control groups in several echocardiographic parameters and laboratory test results. Compared to controls, PH patients exhibited significantly larger dimensions of the ascending aorta, left atrium, left ventricle, right atrium, and right ventricle, whereas their left ventricular ejection fraction (EF%) and fractional shortening (FS%) were significantly lower (all $P < 0.05$, Table 1). Additionally, PH patients exhibited elevated red blood cell counts, reduced platelet counts, an increased neutrophil percentage, and a decreased lymphocyte percentage (all $P < 0.05$, Table 2). In terms of coagulation function, PH patients demonstrated prolonged prothrombin time (PT) and an increased international normalized ratio (INR), as well as elevated levels of total bilirubin, direct bilirubin, and indirect bilirubin (all $P <$

0.05, Table 3 and 4). Regarding electrolyte levels, PH patients exhibited reduced potassium, sodium, and phosphorus levels, along with elevated chlorine and calcium (all $P < 0.05$, Table 5). Overall, these results indicate that PH patients exhibit distinct pathological changes in cardiac structure, hematology, and liver function.

Feature Selection, Model Performance, and SHAP Analysis for Predicting Pulmonary Hypertension

In this study, we employed RFE with a Random Forest model to select the most influential echocardiographic features for predicting PH. Using 10-fold cross-validation, we identified 16 optimal features that significantly contributed to model performance, including the right atrium (RA) and pulmonary artery (PA), which exhibited the highest importance scores (Figure 1A). Figure 1B illustrates the relationship between cross-validation accuracy and the number of variables during the RFE process. As depicted, the model accuracy stabilizes when the number of variables reaches 10 to 16, suggesting that the model achieves optimal predictive performance at this point. Figure 1C illustrates that the XGBoost model outperformed others, excelling in AUC, sensitivity, and specificity, demonstrating high robustness in distinguishing between false positives and negatives. Figures 1D highlight SHAP analysis, revealing that the RA, left ventricular outflow tract velocity, and pulmonary artery diameter had the most significant impact on model predictions, thus making them critical for PH classification.

The XGBoost model was evaluated using a 3:1 developing-to-validation split. The confusion matrix in Figure A shows a high number of correct predictions for the normal class, though there are more false positives (129) for the PH class. In contrast, Figure F displays a slightly lower number of false positives, suggesting improved performance in the second model. The ROC curves (Figures 2B and 2G) reveal AUCs of 0.999 and 0.997, indicating excellent discrimination between classes. Precision-recall curves (Figures 2C and 2H) show very high performance, with precision remaining close to 1 until recall values start dropping, which highlights robust performance in identifying true positives. Specificity and sensitivity metrics (Figures 2D and 2I) were both near 1, indicating high sensitivity

without a significant drop in specificity—an important balance for minimizing both false positives and false negatives. The Accuracy vs. Threshold curves (Figures E and J) reveal that optimal accuracy is achieved at a threshold between 0.7 and 0.8, indicating stable performance across a range of thresholds. Overall, the XGBoost model demonstrates excellent performance, with high AUC values, strong precision-recall trade-offs, and a well-balanced sensitivity and specificity, making it a reliable tool for PH diagnosis.

Model Development and Evaluation for Pulmonary Hypertension Diagnosis

LASSO regression identified key clinical features—including prothrombin time activity, serum cystatin C, and ultrasound indices—as the most influential variables (Figure 3A). The binomial deviance plot helped select the optimal λ to balance complexity and performance, whereas the coefficient shrinkage plot highlighted the most impactful features as λ increased (Figure 3B).

Based on these features, we developed a logistic regression model to diagnose pulmonary hypertension (PH). A nomogram was constructed to quantify each variable's contribution to the diagnosis. The nomogram visually illustrates the scoring for each feature, and the total score is utilized to estimate the risk probability of PH through a linear predictor (Figure 3C). The decision curve analysis further underscores the model's clinical utility, with the Ultrasound nomogram consistently yielding a higher net benefit, especially at higher risk thresholds (Figure 3D). The model's performance evaluation shows strong predictive accuracy. The ROC curves demonstrate excellent discriminatory power, with an AUC of 0.999 for the training set and 0.984 for the testing set (Figure 4 A and B). The calibration curves suggest that the model's predictions are well-aligned with actual outcomes, reflecting reliable calibration (Figure 4 C and D). Overall, the model exhibits robustness, stability, and valuable clinical applicability for predicting pulmonary hypertension.

Ultrasound and Clinical Variable-Based Pulmonary Hypertension Risk Prediction Model

A pulmonary hypertension risk calculator is introduced, which utilizes echocardiographic parameters including the right atrium, pulmonary artery diameter, and left atrium to predict an ultrasound index.

The accompanying SHAP analysis on the right illustrates the contribution of each feature to the model's output, enabling clinicians to rapidly assess a patient's risk by entering echocardiographic data (Figure 5A). Additionally, a nomogram-based pulmonary hypertension risk prediction tool is presented. This tool calculates total points based on clinical variables, such as prothrombin time and serum cystatin C levels. It then displays the total points along with the corresponding risk of pulmonary hypertension, thereby streamlining the assessment process for clinicians (Figure 5B).

Discussion

Our study introduces an integrated diagnostic model combining ultrasound-derived parameters and clinical data to enhance early risk prediction of PH. Using RFE, LASSO regression, and various machine learning algorithms, we optimized the model and rigorously assessed its predictive performance. Given that PH is a rapidly progressive disease with high mortality, early detection and intervention are essential for improving outcomes (21, 22). Current diagnostic methods for pulmonary artery pressure rely heavily on two approaches: right heart catheterization and echocardiography which estimates pressure based on tricuspid regurgitation velocity[23-25]. However, catheterization is invasive, and echocardiography can be prone to significant inaccuracies. To overcome these limitations, this research proposes a novel non-invasive diagnostic pathway that integrates ultrasound and clinical data, thereby providing an effective tool for the early detection of PH. Furthermore, the model establishes a foundation for personalized treatment strategies and continuous monitoring, advancing the precision medicine approach to PH management. In our study, we used RFE to identify an optimal subset of ultrasound features and validated their importance in predicting PH. The results demonstrated that variables such as right atrium (RA), left atrium (LA), and pulmonary artery diameter (PAD) significantly contributed to model performance. The physiological relevance of these ultrasound parameters aligns closely with the pathophysiology of PH. Enlargement of the right and left atria reflects increased pulmonary circulation pressure and heightened right ventricular load, both of which are common in PH patients(26). Additionally, an

enlarged pulmonary artery diameter correlates directly with elevated pulmonary vascular resistance, further exacerbating right ventricular pressure overload(27).

A key strength of our study lies in the rigorous validation of all cases in the PH group were verified by right heart catheterization with a mean pulmonary artery pressure >20 mmHg, combined with echocardiographic data. Additionally, all individuals in the control group had complete echocardiographic data, with all cardiac function parameters within normal ranges. In other words, our case group was diagnosed with PH by the gold standard, whereas our control group was verified by echocardiography to have no structural or functional abnormalities in the heart, thus eliminating the possibility of false positives or false negatives. Although these stringent inclusion criteria resulted in a more limited sample size, they substantially enhanced the precision and clinical relevance of our data and model.

Compared to previous studies, our research systematically identified key ultrasound parameters using RFE, with the process validated through 10-fold cross-validation to ensure the robustness of feature selection(28-30). Unlike traditional approaches that focus on individual ultrasound parameters, our innovation lies in integrating multidimensional ultrasound features with clinical data to construct an efficient and interpretable diagnostic model, significantly enhancing early prediction capabilities for PH(31, 32). This advancement provides clinicians with a more precise ultrasound-based detection strategy to facilitate the early identification of high-risk patients, enabling timely intervention.

In this study, we further modeled the ultrasound features using multiple machine learning algorithms, such as XGBoost, Random Forest, and Logistic Regression. Among these models, XGBoost demonstrated the best predictive performance, achieving an AUC of 0.997, highlighting its strength in handling high-dimensional data and modeling complex nonlinear relationships. As a gradient-boosting algorithm, XGBoost incrementally reduces the model error to enhance predictive accuracy, rendering it particularly effective for clinical data with intricate interactions(33, 34). To enhance the interpretability of the XGBoost model, we constructed an analysis using SHAP. SHAP values

quantified the contribution of each feature to the model's predictions, emphasizing the importance of variables such as the right atrium and pulmonary artery diameter in predicting PH. Compared to traditional "black-box" models, incorporating SHAP analysis significantly enhanced the model's interpretability, providing more clinical insights. SHAP not only increases transparency but also enables clinicians to better understand the predictive mechanisms of the model, thus supporting more informed and reasoned clinical decision-making.

Building on the ultrasound features, we further constructed feature selection for conventional clinical variables using LASSO regression, identifying key variables such as prothrombin time activity and serum cystatin C. Prothrombin time activity is closely associated with coagulopathy in patients with PH, as previous studies demonstrating a significant association between PH and increased thrombotic risk(35). Serum cystatin C, a marker of renal function, is frequently elevated in cases of increased cardiovascular risk(36). This study represents the first to integrate these serum biomarkers with ultrasound features, offering a novel and comprehensive multidimensional predictive model for PH diagnosis.

To facilitate the clinical application of our model, we developed a nomogram based on logistic regression. The nomogram serves as a straightforward, visual tool for clinicians, allowing them to calculate individual PH risk by summing scores of each predictor variable(37). Compared to conventional diagnostic methods, the nomogram provides a personalized risk assessment approach, aiding clinicians in more precise disease screening and early intervention(38). We further evaluated the clinical utility of the model using calibration curves and DCA. The calibration curves demonstrated that the predicted risk is closely aligned with observed outcomes, supporting the model's applicability and robustness across diverse populations. DCA indicated that our model, incorporating both ultrasound and clinical variables, offers significant clinical benefit over the baseline model at different risk thresholds. These results indicate that our model not only exhibits theoretical predictive accuracy but also demonstrates substantial practical value in real-world clinical

settings.

Although several machine learning models have been developed for PH, they exhibit certain limitations. Many models rely on a single data source, which restricts predictive accuracy and reliability. For instance, Athénaïs Boucly's study identified cytokines as prognostic biomarkers in pulmonary arterial hypertension but failed to account for potential confounding factors, such as clinical variables or imaging data, which may influence the results(39). Similarly, Yukina Hirata's research demonstrated improved accuracy of a machine learning model compared to traditional methods in the derivation cohort; however, its performance in the validation cohort was only comparable to guideline-based echocardiographic assessments, underscoring the need for further optimization and validation(40). Additionally, previous studies frequently employed univariable and multivariable models to evaluate the relationship between clinical and echocardiographic parameters in precapillary pulmonary hypertension (PH)(41). However, they lacked advanced feature selection methods like LASSO and Recursive Feature Elimination (RFE), which could have enhanced model performance by identifying key predictive variables, reducing overfitting, and providing deeper insights into prognostic factors. In contrast, our research proposes an innovative approach by integrating ultrasound parameters with clinical variables, thereby enhancing the model's predictive performance through multidimensional integration(42, 43). Unlike prior work, we optimized feature selection using both RFE and LASSO regression and ensured model robustness through comparison and tuning across multiple machine-learning models(44). We also significantly enhanced model interpretability using SHAP (Shapley Additive Explanations), which makes this high-performing "black-box" model more transparent and trustworthy for clinical use, ultimately enhancing its applicability and credibility in a healthcare context(17, 45, 46).

Despite the strong performance of our model in predicting PH, it exhibits certain limitations. First, the study sample size was relatively small, which limits the generalizability of the findings. Future work should involve validation using larger, multicenter datasets to ensure the model's external

validity. Second, although we employed cross-validation and SHAP analysis to assess model robustness and interpretability, further validation is needed to determine its effectiveness in real-world clinical settings, particularly with dynamic data. Additionally, as new ultrasound technologies and biomarkers emerge, future research should incorporate additional multidimensional biological data to enhance the predictive accuracy and applicability of the model.

Conclusion

Our study introduces an innovative multidimensional diagnostic model for PH, integrating ultrasound parameters and clinical data, and applying machine learning methods to optimize feature selection and model construction. Our approach significantly improves the early diagnostic capability for PH. The model not only exhibits high predictive performance and interpretability but also demonstrates substantial potential for clinical application. Future research could build on this model to explore its utility in diagnosing other cardiovascular diseases, thereby broadening its clinical impact.

Declarations

Ethics approval and consent to participate

The study was approved by the Research Ethics Commission of Wuhan Zhongnan Hospital and the requirement for informed consent was waived by the Ethics Commission (Approval No.2023185)

Consent for publication

All authors agreed with the submission of the manuscript for publication and agreed to be accountable for all aspects of the manuscript.

Competing interests

All authors declare that they have no conflicts of interest.

Data availability

All the original data involved in this investigation are available on reasonable request to the first author or corresponding author.

Founding

This study was funded by the following programs: The National Natural Science Foundation of China (No.82300078), the Fundamental Research Funds for the Central Universities (No.2042023kf0059), Wuhan University Clinical Medicine + Youth Supporting Program (No.413000557), The Climbing Project for Medical Talent of Zhongnan Hospital of Wuhan University (Grant No. PDJH202406), Science and Technology Innovation Cultivation Funding of Zhongnan Hospital of Wuhan University (No. CXPY2023060).

Authors' Contributions

Hongxia Jiang and Han Gao designed the research, analyzed data, and drafted the manuscript. Hao Li, Qingli Zeng, and Xiaojun Hao prepared and collected data. Zhenshun Cheng conceived and designed the research edited and revised the manuscript and supervised all aspects of the project. Hongxia Jiang, Han Gao, Hao Li, Qingli Zeng, Xiaojun Hao, and Zhenshun Cheng approved the final version of the manuscript.

Abbreviations

AUC-ROC: Area Under the Receiver Operating Characteristic Curve, BMI: Body Mass Index, CTEPH: Chronic Thromboembolic Pulmonary Hypertension, DCA: Decision Curve Analysis, EF:Ejection Fraction, FS: Fractional Shortening, GBM: Gradient Boosting Machine, IPAH: Idiopathic Pulmonary Arterial Hypertension, INR: International Normalized Ratio, KNN: K-Nearest Neighbors, LASSO: Least Absolute Shrinkage and Selection Operator, LA: Left Atrium, mPAP: mean Pulmonary Arterial Pressure, PH: Pulmonary hypertension, PT: Prolonged Prothrombin Time, PAD: Pulmonary Artery Diameter, PA: Pulmonary Artery, PASP: Pulmonary Artery Systolic Pressure, RFE: Recursive Feature Elimination, RA: Right Atrium, SHAP: Shapley Additive Explanations, REVEAL-ECHO: Registry to Evaluate Early and Long-term Pulmonary Arterial Hypertension Disease Management - Echocardiography, RHC: Right Heart Catheterization,

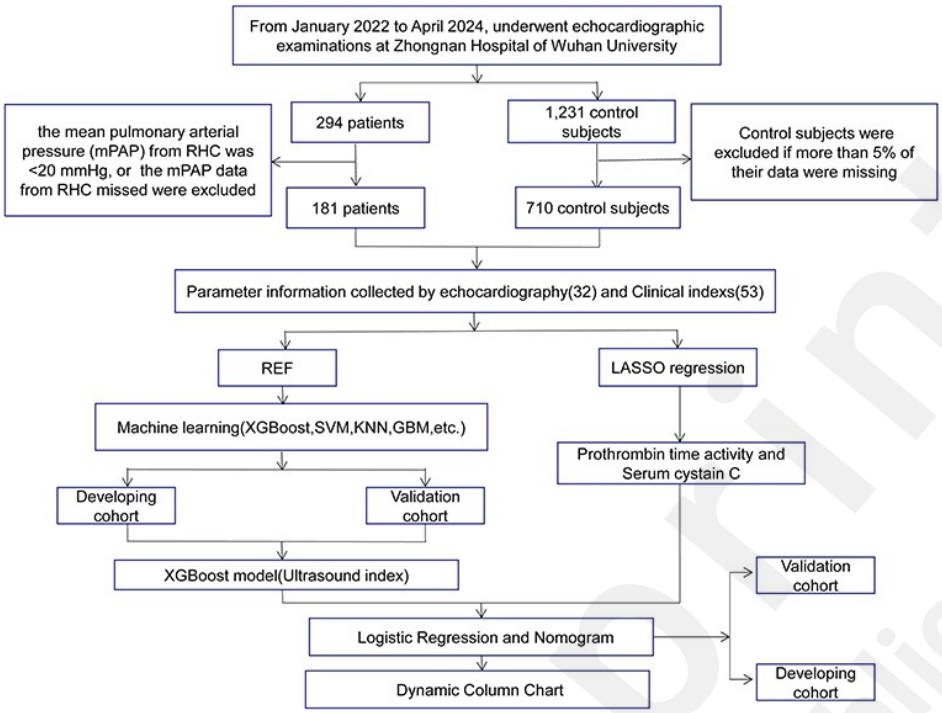
SVM:Support Vector Machine, TRV: Tricuspid Regurgitation Velocity, TAPSE: Tricuspid Annular Plane Systolic Excursion, XGBoost: Extreme Gradient Boosting.

Reference

1. Humbert M, Sitbon O, Guignabert C, Savale L, Boucly A, Gallant-Dewavrin M, et al. Treatment of pulmonary arterial hypertension: recent progress and a look to the future. Lancet Respir Med. 2023;11(9):804-19.
2. Johnson S, Sommer N, Cox-Flaherty K, Weissmann N, Ventetuolo CE, Maron BA. Pulmonary Hypertension: A Contemporary Review. Am J Respir Crit Care Med. 2023;208(5):528-48.
3. Weatherald J, Hemnes AR, Maron BA, Mielniczuk LM, Gerges C, Price LC, et al. Phenotypes in pulmonary hypertension. Eur Respir J. 2024;64(3).
4. Maron BA, Brittain EL, Hess E, Waldo SW, Barón AE, Huang S, et al. Pulmonary vascular resistance and clinical outcomes in patients with pulmonary hypertension: a retrospective cohort study. Lancet Respir Med. 2020;8(9):873-84.
5. Kovacs G, Avian A, Bachmaier G, Troester N, Tornyos A, Douschan P, et al. Severe Pulmonary Hypertension in COPD: Impact on Survival and Diagnostic Approach. Chest. 2022;162(1):202-12.
6. Balistrieri A, Makino A, Yuan JX. Pathophysiology and pathogenic mechanisms of pulmonary hypertension: role of membrane receptors, ion channels, and Ca(2+) signaling. Physiol Rev. 2023;103(3):1827-97.
7. El-Kersh K, Zhao C, Elliott G, Farber HW, Gomberg-Maitland M, Selej M, et al. Derivation of a Risk Score (REVEAL-ECHO) Based on Echocardiographic Parameters of Patients With Pulmonary Arterial Hypertension. Chest. 2023;163(5):1232-44.
8. Vonk Noordegraaf A, Chin KM, Haddad F, Hassoun PM, Hemnes AR, Hopkins SR, et al. Pathophysiology of the right ventricle and of the pulmonary circulation in pulmonary hypertension: an update. Eur Respir J. 2019;53(1).
9. Ley L, Grimminger F, Richter M, Tello K, Ghofrani A, Bandorski D. The Early Detection of Pulmonary Hypertension. Dtsch Arztebl Int. 2023;120(48):823-30.
10. Celestin BE, Bagherzadeh SP, Ichimura K, Santana EJ, Sanchez PA, Tobore T, et al. Identifying consistent echocardiographic thresholds for risk stratification in pulmonary arterial hypertension. Pulm Circ. 2024;14(2):e12361.
11. Shah AM, Cikes M, Prasad N, Li G, Getchevski S, Claggett B, et al. Echocardiographic Features of Patients With Heart Failure and Preserved Left Ventricular Ejection Fraction. J Am Coll Cardiol. 2019;74(23):2858-73.
12. Álvarez Troncoso J, Soto Abánades C, Robles-Marhuenda Á, Alcolea Batres S, Fernández Velilla Peña M, Jiménez Valero S, et al. Prevalence, risk factors and echocardiographic predictors of pulmonary hypertension in systemic lupus erythematosus: towards a screening protocol. RMD Open. 2024;10(1).
13. Yan T, Ma Q, Li X, Shen Q, Liu X, Zhang X. Establishment of a prediction model of pulmonary artery hypertension in patients with hyperthyroidism. Ann Noninvasive Electrocardiol. 2024;29(5):e13133.
14. Zhou Z, Gu Y, Tian L, Zheng H, Li S. Development and Validation of a Nomogram of Persistent Pulmonary Hypertension in Adult Pretricuspid Shunts After Correction. J Am Heart Assoc. 2024;13(9):e032412.
15. Hu M, Duan A, Huang Z, Zhao Z, Zhao Q, Yan L, et al. Development and Validation of a Nomogram for Predicting Obstructive Sleep Apnea in Patients with Pulmonary Arterial Hypertension. Nat Sci Sleep. 2022;14:1375-86.
16. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;28(5):1 - 26.
17. <1705.07874v2.pdf>.
18. <Normogram-rms.pdf>.
19. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making. 2006;26(6):565-74.
20. <Decision Curve Analysis-rmda.pdf>.
21. Fu W, He W, Li Y, Chen Y, Liang J, Lei H, et al. Efficacy and safety of novel-targeted drugs in the treatment of pulmonary arterial hypertension: a Bayesian network meta-analysis. Drug Deliv. 2021;28(1):1007-19.
22. Maron BA, Abman SH, Elliott CG, Frantz RP, Hopper RK, Horn EM, et al. Pulmonary Arterial Hypertension: Diagnosis, Treatment, and Novel Advances. Am J Respir Crit Care Med. 2021;203(12):1472-87.
23. Poch D, Mandel J. Pulmonary Hypertension. Ann Intern Med. 2021;174(4):Itc49-itc64.
24. Mandras SA, Mehta HS, Vaidya A. Pulmonary Hypertension: A Brief Guide for Clinicians. Mayo Clin Proc. 2020;95(9):1978-88.
25. Rosenkranz S, Gibbs JS, Wachter R, De Marco T, Vonk-Noordegraaf A, Vachiéry JL. Left ventricular heart failure and pulmonary hypertension. Eur Heart J. 2016;37(12):942-54.
26. Galderisi M, Cosyns B, Edvardsen T, Cardim N, Delgado V, Di Salvo G, et al. Standardization of adult transthoracic echocardiography reporting in agreement with recent chamber quantification, diastolic function, and

heart valve disease recommendations: an expert consensus document of the European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging. 2017;18(12):1301-10.
27. Jones R, Zapol WM, Reid L. Pulmonary artery remodeling and pulmonary hypertension after exposure to hyperoxia for 7 days. A morphometric and hemodynamic study. Am J Pathol. 1984;117(2):273-85.
28. Hameed A, Condliffe R, Swift AJ, Alabed S, Kiely DG, Charalampopoulos A. Assessment of Right Ventricular Function-a State of the Art. Curr Heart Fail Rep. 2023;20(3):194-207.
29. Arvanitaki A, Diller GP, Gatzoulis MA, McCabe C, Price LC, Wort SJ. Noninvasive diagnostic modalities and prediction models for detecting pulmonary hypertension associated with interstitial lung disease: a narrative review. Eur Respir Rev. 2024;33(174).
30. Ostasevicius V, Jurenas V, Venslauskas M, Kizauskiene L, Zigmantaite V, Stankevicius E, et al. Low-frequency ultrasound for pulmonary hypertension therapy. Respir Res. 2024;25(1):70.
31. Cassady SJ, Ramani GV. Right Heart Failure in Pulmonary Hypertension. Cardiol Clin. 2020;38(2):243-55.
32. D'Alto M, Liccardo B, Di Maio M, Del Giudice C, Romeo E, Argiento P, et al. Lung Ultrasound, Echocardiography, and Fluid Challenge for the Differential Diagnosis of Pulmonary Hypertension. J Am Soc Echocardiogr. 2023;36(11):1181-9.
33. Silva GFS, Fagundes TP, Teixeira BC, Chiavegatto Filho ADP. Machine Learning for Hypertension Prediction: a Systematic Review. Curr Hypertens Rep. 2022;24(11):523-33.
34. Hou N, Li M, He L, Xie B, Wang L, Zhang R, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. J Transl Med. 2020;18(1):462.
35. Lutsey PL, Evensen LH, Thenappan T, Prins KW, Walker RF, Farley JF, et al. Incidence and Risk Factors of Pulmonary Hypertension After Venous Thromboembolism: An Analysis of a Large Health Care Database. J Am Heart Assoc. 2022;11(14):e024358.
36. Shlipak MG, Sarnak MJ, Katz R, Fried LF, Seliger SL, Newman AB, et al. Cystatin C and the risk of death and cardiovascular events among elderly persons. N Engl J Med. 2005;352(20):2049-60.
37. Shi S, Zhu X, Cheang I, Liao S, Yin T, Lu X, et al. Development and validation of a diagnostic nomogram in pulmonary hypertension due to left heart disease. Heart Lung. 2024;65:11-8.
38. Deng X, Hou H, Wang X, Li Q, Li X, Yang Z, et al. Development and validation of a nomogram to better predict hypertension based on a 10-year retrospective cohort study in China. Elife. 2021;10.
39. Moss BJ, Ryter SW, Rosas IO. Pathogenic Mechanisms Underlying Idiopathic Pulmonary Fibrosis. Annu Rev Pathol. 2022;17:515-46.
40. Hirata Y, Tsuji T, Kotoku J, Sata M, Kusunose K. Echocardiographic artificial intelligence for pulmonary hypertension classification. Heart. 2024;110(8):586-93.
41. Siddiqui I, Rajagopal S, Brucker A, Chiswell K, Christopher B, Alenezi F, et al. Clinical and Echocardiographic Predictors of Outcomes in Patients With Pulmonary Hypertension. Am J Cardiol. 2018;122(5):872-8.
42. Hirsch K, Nolley S, Ralph DD, Zheng Y, Altemeier WA, Rhodes CJ, et al. Circulating markers of inflammation and angiogenesis and clinical outcomes across subtypes of pulmonary arterial hypertension. J Heart Lung Transplant. 2023;42(2):173-82.
43. Tello K, Wan J, Dalmer A, Vanderpool R, Ghofrani HA, Naeije R, et al. Validation of the Tricuspid Annular Plane Systolic Excursion/Systolic Pulmonary Artery Pressure Ratio for the Assessment of Right Ventricular-Arterial Coupling in Severe Pulmonary Hypertension. Circ Cardiovasc Imaging. 2019;12(9):e009047.
44. Huang Y, Ma SF, Oldham JM, Adegunsoye A, Zhu D, Murray S, et al. Machine Learning of Plasma Proteomics Classifies Diagnosis of Interstitial Lung Disease. Am J Respir Crit Care Med. 2024;210(4):444-54.
45. Hu J, Xu J, Li M, Jiang Z, Mao J, Feng L, et al. Identification and validation of an explainable prediction model of acute kidney injury with prognostic implications in critically ill children: a prospective multicenter cohort study. EClinicalMedicine. 2024;68:102409.
46. Raihan MJ, Khan MA, Kee SH, Nahid AA. Detection of the chronic kidney disease using XGBoost classifier and explaining the influence of the attributes on the model using SHAP. Sci Rep. 2023;13(1):6263.

Figures and Figure legends



Workflow of the analysis

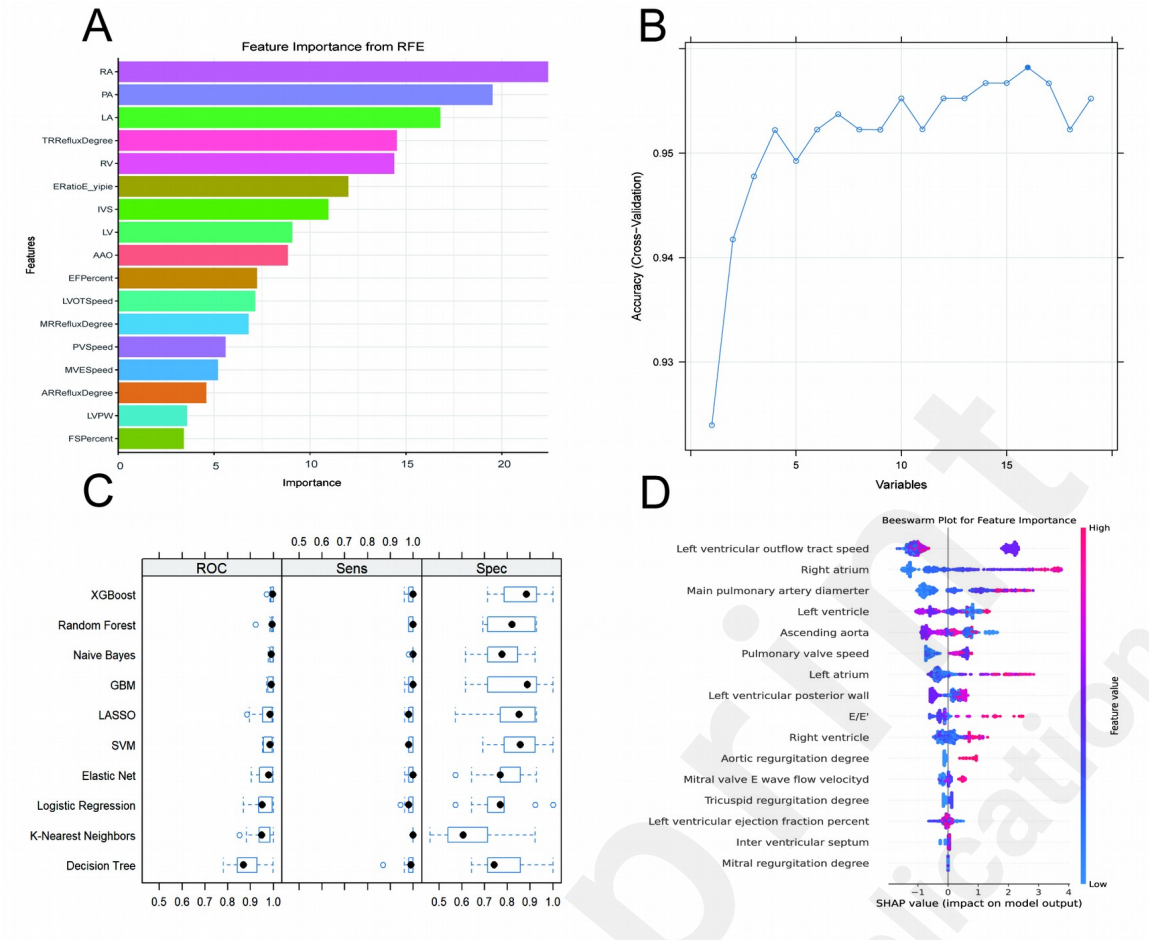


Figure 1: (A) eDeviance Plot from Lasso Regression. (B) Coefficient Path for Lasso Regression. (C) Nomogram

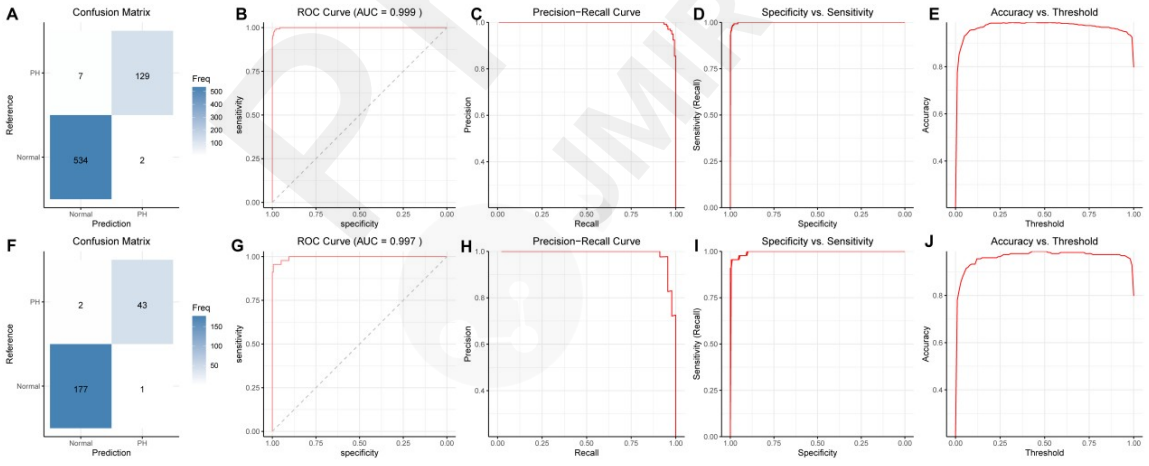


Figure 2: (A) Confusion Matrix for Developing cohort. (B) ROC Curve for Developing cohort (AUC = 0.999). (C) Precision-Recall Curve for Developing cohort. (D) Specificity vs Sensitivity for Developing cohort. (E) Accuracy vs Threshold for Developing cohort. (F) Confusion Matrix for Validation cohort. (G) ROC Curve for Validation cohort (AUC = 0.997). (H) Precision-Recall Curve

for Validation cohort. (I) Specificity vs Sensitivity for Validation cohort. (J) Accuracy vs Threshold for Validation cohort.

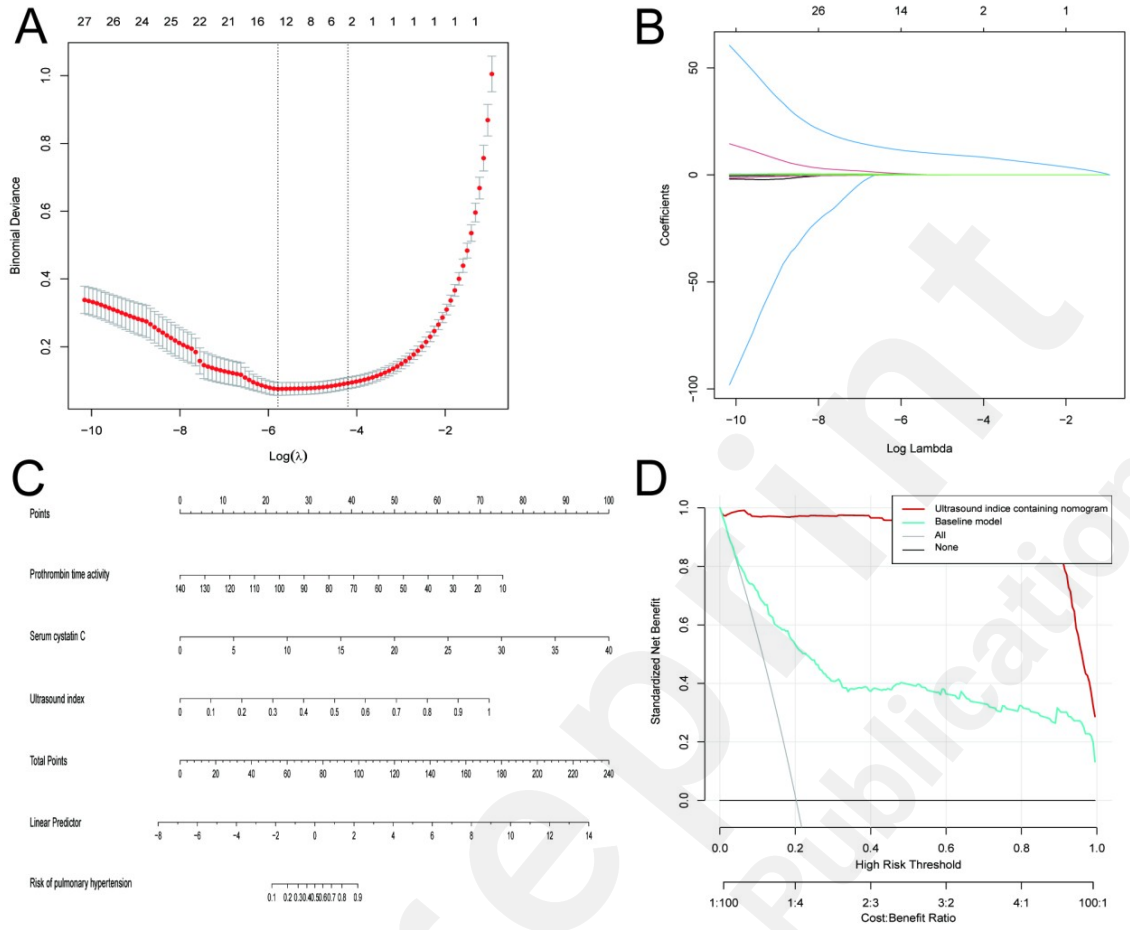


Figure 3: (A) Deviance Plot from Lasso Regression. (B) Coefficient Path for Lasso Regression. (C) Nomogram for Risk Prediction of Pulmonary Hypertension. (D) Decision Curve Analysis.

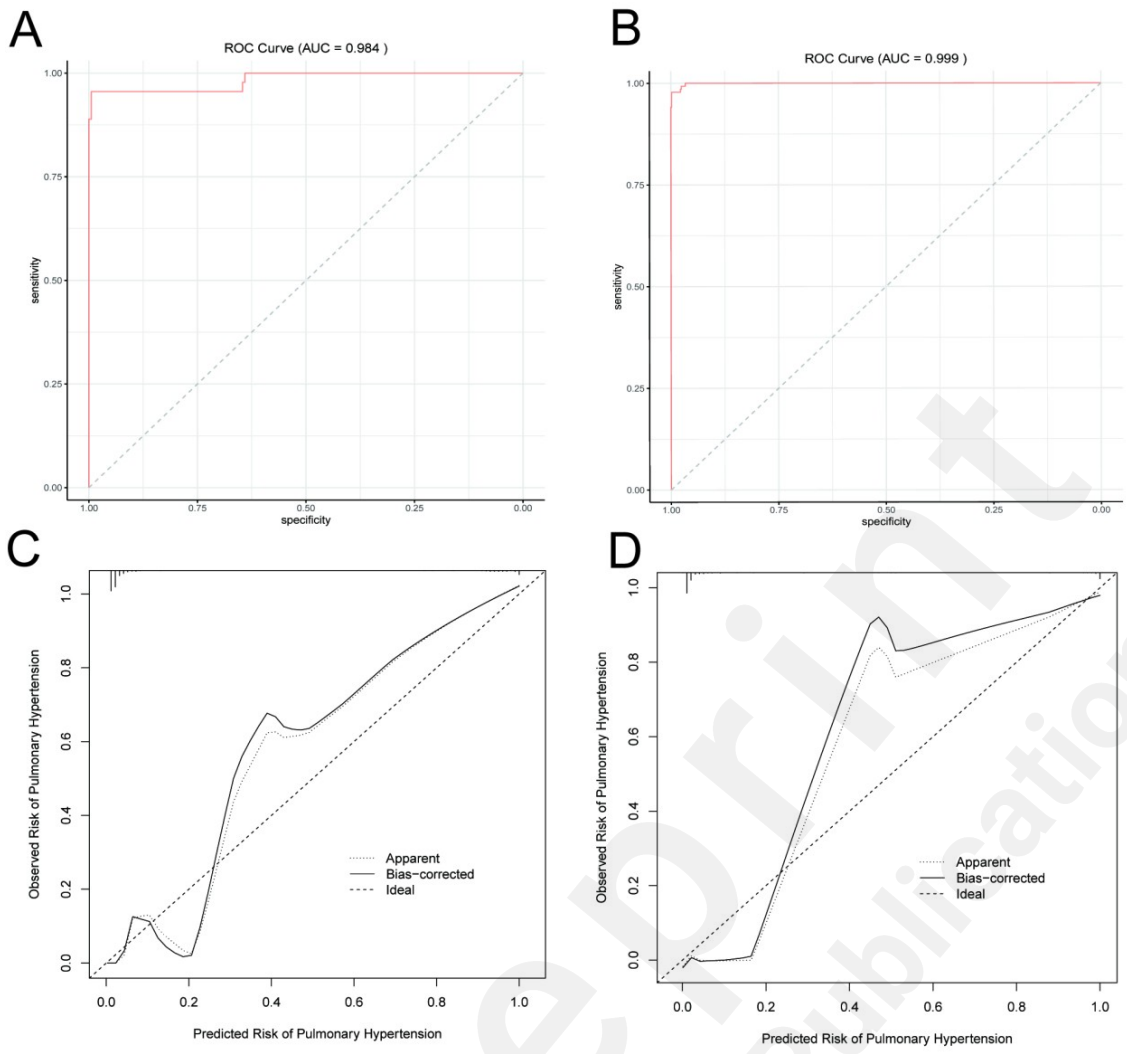
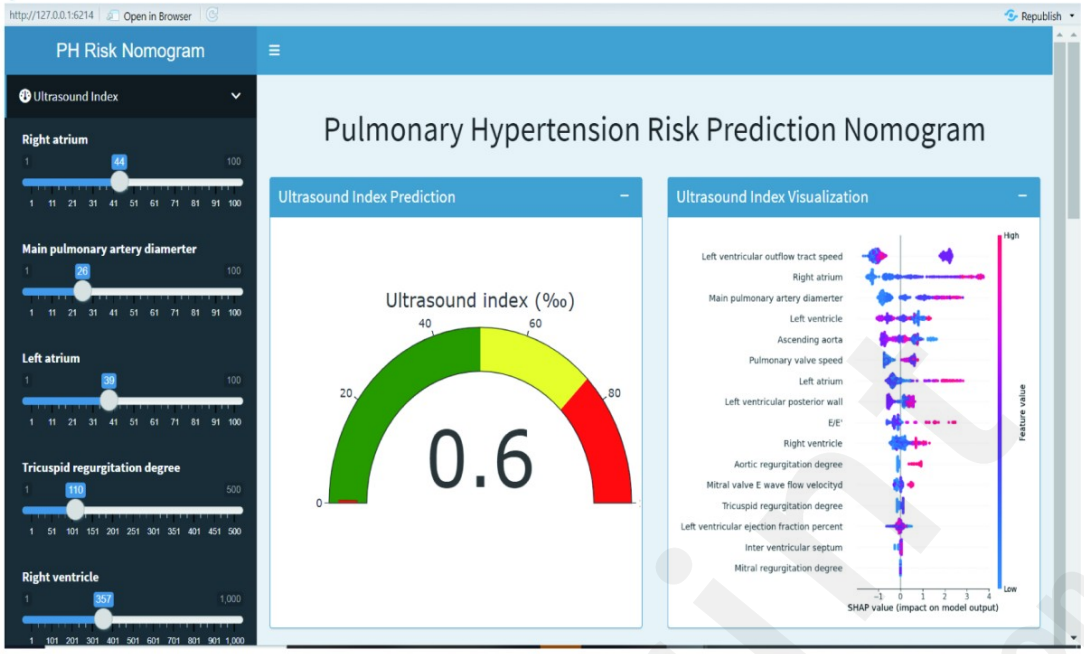


Figure 4: (A) ROC of the nomogram in the developing cohort. (B) ROC of the nomogram in validation cohort. (C) Calibration curve of the nomogram in the developing cohort. (D) Calibration curve of the nomogram in the validation cohort.

A



B

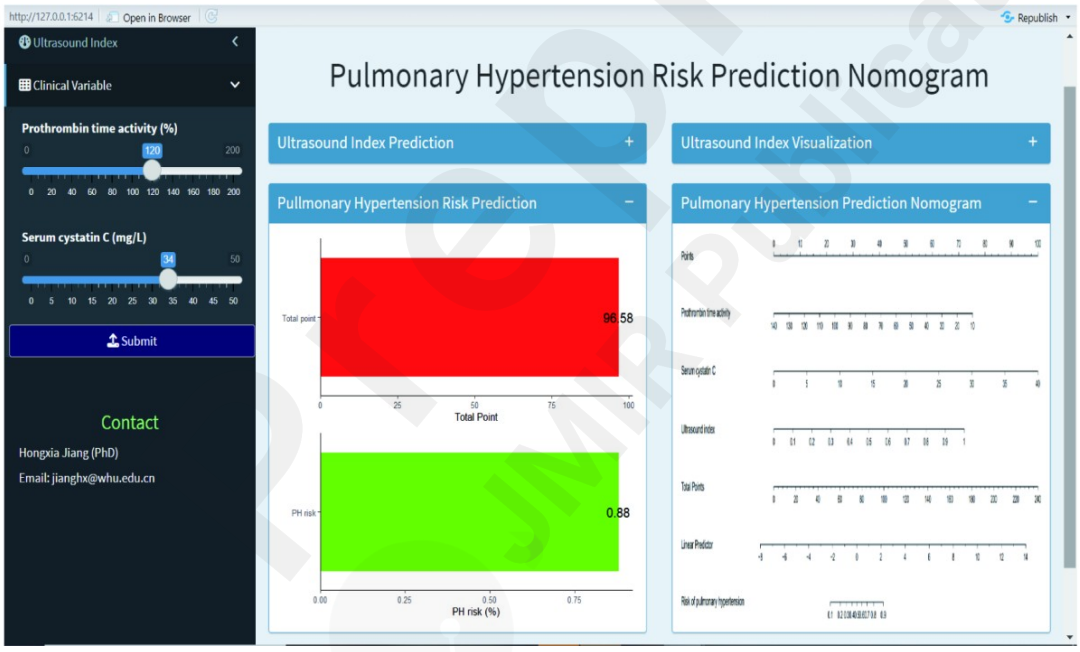


Figure 5: (A) Ultrasound-based pulmonary hypertension risk calculator showing index prediction and SHAP analysis explaining feature contributions to model output. The PH risk prediction nomogram uses clinical variables to calculate a total score, estimating the corresponding PH risk percentage.

Table 1 Comparison of Demographic and Echocardiographic Parameters Between Normal and Pulmonary Hypertension (PH) Patients

	Normal	PH (N=181)	P value	Overall(N=895)
Age				
Mean (SD)	52.0 (13.8)	53.2 (16.9)	0.38	52.3 (14.5)
Median [Min, Max]	53.0 [15.0, 88.0]	54.7 [16.0, 85.1]		54.0 [15.0, 88.0]
Sex				
Male	310 (43.4%)	81 (44.8%)	0.811	391 (43.7%)
Female	404 (56.6%)	100 (55.2%)		504 (56.3%)
AAO				
Mean (SD)	3.02 (0.280)	3.15 (0.540)	0.002	3.05 (0.352)
Median [Min, Max]	3.10 [2.20, 3.70]	3.20 [1.80, 4.90]		3.10 [1.80, 4.90]
LA				
Mean (SD)	3.18 (0.362)	4.22 (1.18)	<0.001	3.39 (0.746)
Median [Min, Max]	3.20 [1.80, 4.30]	4.00 [2.10, 7.70]		3.30 [1.80, 7.70]
LV				
Mean (SD)	4.41 (0.333)	4.98 (1.42)	<0.001	4.53 (0.740)
Median [Min, Max]	4.40 [3.20, 5.30]	4.60 [2.50, 9.90]		4.40 [2.50, 9.90]
IVS				
Mean (SD)	0.933 (0.112)	0.997 (0.210)	<0.001	0.946 (0.140)
Median [Min, Max]	0.900 [0, 1.30]	1.00 [0.500, 2.10]		0.900 [0, 2.10]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
LVPW				
Mean (SD)	0.908 (0.0988)	0.968 (0.254)	0.002	0.920 (0.146)

Table 1 Comparison of Demographic and Echocardiographic Parameters Between Normal and Pulmonary Hypertension (PH) Patients (continued)

Median [Min, Max]	0.900 [0.600, 1.20]	1.00 [0.500, 3.40]		0.900 [0.500, 3.40]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
RA				
Mean (SD)	30.8 (3.74)	47.2 (13.1)	<0.001	34.1 (9.45)
Median [Min, Max]	31.0 [20.0, 41.0]	46.0 [0, 107]		32.0 [0, 107]
RV				
Mean (SD)	3.01 (0.362)	4.33 (1.13)	<0.001	3.28 (0.799)
Median [Min, Max]	3.00 [1.90, 4.10]	4.15 [2.30, 7.60]		3.10 [1.90, 7.60]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
PA				
Mean (SD)	20.7	29.8	<0.001	22.5 (5.58)

	(2.08)	(8.44)		
Median [Min, Max]	21.0 [14.0, 31.0]	28.0 [0, 62.0]		21.0 [0, 62.0]
FSPercent				
Mean (SD)	36.0 (3.86)	32.6 (5.97)	<0.001	35.4 (4.50)
Median [Min, Max]	35.0 [22.0, 47.0]	33.0 [12.0, 44.0]		35.0 [12.0, 47.0]
Missing	0 (0%)	28 (15.5%)		28 (3.1%)
EFPercent				
Mean (SD)	65.7 (4.54)	57.1 (14.6)	<0.001	64.1 (8.34)
Median [Min, Max]	65.0 [53.0, 79.0]	62.0 [12.0, 75.0]		65.0 [12.0, 79.0]
Missing	0 (0%)	8 (4.4%)		8 (0.9%)
MVESpeed				
Mean (SD)	0.741 (0.195)	1.01 (0.488)	<0.001	0.796 (0.300)

Table 1 Comparison of Demographic and Echocardiographic Parameters Between Normal and Pulmonary Hypertension (PH) Patients (continued)

Median [Min, Max]	0.700 [0.300, 1.40]	0.900 [0.300, 3.00]		0.700 [0.300, 3.00]
Missing	0 (0%)	3 (1.7%)		3 (0.3%)
ASpeed				
Mean (SD)	0.778 (0.189)	0.782 (0.254)	0.86	0.779 (0.200)
Median [Min, Max]	0.800 [0.400, 1.40]	0.800 [0.300, 1.60]		0.800 [0.300, 1.60]
Missing	0 (0%)	52 (28.7%)		52 (5.8%)
LVOTSpeed				
Mean (SD)	0.912 (0.189)	0.739 (0.269)	0.002	0.906 (0.195)
Median [Min, Max]	0.900 [0.500, 1.80]	0.800 [0.300, 1.20]		0.900 [0.300, 1.80]
Missing	0 (0%)	153 (84.5%)		153 (17.1%)
AVSpeed				
Mean (SD)	1.22 (0.216)	1.32 (0.423)	0.004	1.24 (0.274)
Median [Min, Max]	1.20 [0.700, 2.40]	1.20 [0.500, 3.10]		1.20 [0.500, 3.10]
Missing	11 (1.5%)	2 (1.1%)		13 (1.5%)
PVSpeed				
Mean (SD)	0.953 (0.175)	1.09 (0.458)	<0.001	0.977 (0.256)
Median [Min, Max]	0.900 [0.600, 1.70]	1.00 [0.300, 4.10]		1.00 [0.300, 4.10]
Missing	0 (0%)	22 (12.2%)		22 (2.5%)
MRRefluxDegree				
Mean (SD)	0.296 (0.457)	1.08 (1.15)	<0.001	0.454 (0.729)
Median [Min, Max]	0 [0, 1.00]	1.00 [0, 4.00]		0 [0, 4.00]
ARRefluxDegree				

Table 1 Comparison of Demographic and Echocardiographic Parameters Between Normal and Pulmonary Hypertension (PH) Patients (continued)

Mean (SD)	0.0980 (0.298)	0.459 (0.619)	<0.001	0.171 (0.411)
Median [Min, Max]	0 [0, 1.00]	0 [0, 3.00]		0 [0, 3.00]
TRRefluxDegree				
Mean (SD)	0.396 (0.489)	1.85 (1.11)	<0.001	0.690 (0.882)
Median [Min, Max]	0 [0, 1.00]	2.00 [0, 4.00]		0 [0, 4.00]
Widening_of_ascending_aorta				
No	714 (100%)	142 (78.5%)	<0.001	856 (95.6%)
Yes	0 (0%)	39 (21.5%)		39 (4.4%)
Aortic_valve_thickening				
No	696 (97.5%)	147 (81.2%)	<0.001	843 (94.2%)
Yes	18 (2.5%)	34 (18.8%)		52 (5.8%)
Aortic_valve_echo_intensification				
No	695 (97.3%)	140 (77.3%)	<0.001	835 (93.3%)
Yes	19 (2.7%)	41 (22.7%)		60 (6.7%)
aortic_valve_calcification				
No	712 (99.7%)	175 (96.7%)	<0.001	887 (99.1%)
Yes	2 (0.3%)	6 (3.3%)		8 (0.9%)
Aortic_valve_closure_is_poor				
No	713 (99.9%)	119 (65.7%)	<0.001	832 (93.0%)
Yes	1 (0.1%)	62 (34.3%)		63 (7.0%)
Pulmonary_artery_widening				
No	714 (100%)	88 (48.6%)	<0.001	802 (89.6%)
Yes	0 (0%)	93 (51.4%)		93 (10.4%)
Poor_closure_of_the_pulmonary_valve				
No	713 (99.9%)	157 (86.7%)	<0.001	870 (97.2%)
Yes	1 (0.1%)	24 (13.3%)		25 (2.8%)

Table 1 Comparison of Demographic and Echocardiographic Parameters Between Normal and Pulmonary Hypertension (PH) Patients (continued)

Mitral_valve_insufficiency				
No	704 (98.6%)	83 (45.9%)	<0.001	787 (87.9%)
Yes	10 (1.4%)	98 (54.1%)		108 (12.1%)
tricuspid_insufficiency				
No	677 (94.8%)	33 (18.2%)	<0.001	710 (79.3%)
Yes	37 (5.2%)	148 (81.8%)		185 (20.7%)
The_ventricular_septum_thickened				
No	711 (99.6%)	147 (81.2%)	<0.001	858 (95.9%)
Yes	3 (0.4%)	34 (18.8%)		37 (4.1%)
Left_ventricular_posterior_wall_thickened				
No	714 (100%)	164 (90.6%)	<0.001	878 (98.1%)

Yes	0 (0%)	17 (9.4%)		17 (1.9%)
atrial_septal_defect				
No	714 (100%)	143 (79.0%)	<0.001	857 (95.8%)
Yes	0 (0%)	38 (21.0%)		38 (4.2%)
ventricular_septal_defect				
No	714 (100%)	171 (94.5%)	<0.001	885 (98.9%)
Yes	0 (0%)	10 (5.5%)		10 (1.1%)
ERatioE_yipie				
Mean (SD)	9.24 (1.99)	13.0 (5.43)	<0.001	9.97 (3.33)
Median [Min, Max]	9.00 [4.00, 18.0]	12.0 [5.00, 33.0]		10.0 [4.00, 33.0]
Missing	121 (16.9%)	39 (21.5%)		160 (17.9%)

Table 2 blood routine index in Normal and Pulmonary Hypertension (PH) Patients

	Normal	PH (N=181)	P value	Overall(N=895)
WBC				
Mean (SD)	6.16 (4.41)	6.15 (2.33)	0.972	6.16 (4.08)
Median [Min, Max]	5.67 [1.05, 103]	5.62 [2.50, 17.5]		5.65 [1.05, 103]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
RBC				
Mean (SD)	4.19 (0.660)	4.38 (0.905)	0.01	4.23 (0.720)
Median [Min, Max]	4.22 [1.76, 7.86]	4.27 [1.77, 8.48]		4.23 [1.76, 8.48]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Hb				
Mean (SD)	127 (19.4)	131 (26.1)	0.137	128 (20.9)
Median [Min, Max]	129 [52.8, 194]	130 [60.0, 252]		129 [52.8, 252]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Plt				
Mean (SD)	222 (80.4)	183 (75.3)	<0.001	214 (80.8)
Median [Min, Max]	215 [16.0, 575]	174 [34.0, 545]		206 [16.0, 575]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
neutrophilic_granulocyte_percentage				
Mean (SD)	60.9 (12.4)	64.0 (11.1)	0.002	61.5 (12.2)
Median [Min, Max]	61.0 [12.8, 97.0]	63.6 [32.8, 94.2]		61.5 [12.8, 97.0]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
lymphocytes_Percentage				
Mean (SD)	28.0 (10.6)	25.3 (10.1)	0.002	27.4 (10.5)
Median [Min, Max]	28.0 [2.00, 72.2]	26.0 [1.70, 53.4]		27.5 [1.70, 72.2]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Monocyte_percentage				
Mean (SD)	8.34 (3.94)	8.35 (2.81)	0.993	8.34 (3.74)
Median [Min, Max]	7.70 [0.800, 78.9]	8.20 [1.50, 19.4]		7.80 [0.800, 78.9]

Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Eosinophils_Percentage				
Mean (SD)	2.20 (2.22)	1.83 (3.11)	0.13	2.13 (2.43)
Median [Min, Max]	1.55 [0, 27.5]	1.20 [0, 37.0]		1.50 [0, 37.0]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Basophils_Percentage				
Mean (SD)	0.562 (0.375)	0.595 (0.368)	0.284	0.569 (0.373)
Median [Min, Max]	0.500 [0, 3.30]	0.500 [0, 2.60]		0.500 [0, 3.30]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Neutrophil_absolute_value				
Mean (SD)	3.85 (2.39)	4.03 (2.03)	0.304	3.89 (2.32)

Table 2 blood routine index in Normal and Pulmonary Hypertension (PH) Patients (continued)

Median [Min, Max]	3.40 [0.200, 28.7]	3.53 [1.38, 12.9]		3.40 [0.200, 28.7]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Lymphocyte_absolute_value				
Mean (SD)	1.57 (0.660)	1.48 (0.691)	0.105	1.55 (0.667)
Median [Min, Max]	1.50 [0.200, 6.79]	1.45 [0.180, 3.86]		1.50 [0.180, 6.79]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Monocyte_absolute_value				
Mean (SD)	0.582 (3.02)	0.500 (0.248)	0.474	0.565 (2.70)
Median [Min, Max]	0.410 [0, 80.9]	0.470 [0.100, 2.00]		0.430 [0, 80.9]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Eosinophils_absolute_value				
Mean (SD)	0.124 (0.134)	0.106 (0.242)	0.326	0.121 (0.161)
Median [Min, Max]	0.100 [0, 1.40]	0.100 [0, 3.04]		0.100 [0, 3.04]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Basophil_absolute_value				
Mean (SD)	0.0212 (0.0326)	0.0275 (0.0341)	0.027	0.0225 (0.0330)
Median [Min, Max]	0 [0, 0.200]	0.0200 [0, 0.110]		0 [0, 0.200]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
HCThematokrit				
Mean (SD)	38.4 (5.67)	40.1 (7.77)	0.005	38.7 (6.19)
Median [Min, Max]	38.8 [16.5, 61.8]	40.1 [17.6, 76.4]		38.9 [16.5, 76.4]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
mean_RBC_volume				
Mean (SD)	91.8 (6.48)	92.0 (7.42)	0.754	91.9 (6.67)
Median [Min, Max]	92.0 [63.2, 120]	92.8 [65.5, 123]		92.1 [63.2, 123]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Average_hemoglobin_content				
Mean (SD)	30.5 (2.52)	30.0 (3.08)	0.029	30.4 (2.65)
Median [Min, Max]	30.7 [18.1, 40.9]	30.8 [19.5, 40.3]		30.7 [18.1, 40.9]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Average_hemoglobin_concentration				
Mean (SD)	332 (8.33)	325 (14.1)	<0.001	331 (10.1)
Median [Min, Max]	332 [285, 364]	328 [257, 352]		332 [257, 364]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Erythrocyte_distribution_width_CV				

Mean (SD)	14.1 (1.94)	14.9 (2.28)	<0.001	14.3 (2.04)

Table 2 blood routine index in Normal and Pulmonary Hypertension (PH) Patients (continued)

Median [Min, Max]	13.5 [11.3, 28.9]	14.3 [11.9, 26.1]		13.6 [11.3, 28.9]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
mean_platelet_volume				
Mean (SD)	8.69 (1.17)	9.69 (1.49)	<0.001	8.89 (1.30)
Median [Min, Max]	8.50 [5.80, 14.5]	9.50 [7.10, 13.6]		8.70 [5.80, 14.5]
Missing	0 (0%)	6 (3.3%)		6 (0.7%)

Table 3 Indicators of coagulation function in Normal and Pulmonary Hypertension (PH) Patients

	Normal	PH (N=181)	P value	Overall(N=895)
PTprothrombin_time				
Mean (SD)	11.3 (0.992)	14.7 (7.36)	<0.001	12.0 (3.68)
Median [Min, Max]	11.1 [9.50, 19.7]	12.7 [9.30, 70.1]		11.3 [9.30, 70.1]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)

INR				
Mean (SD)	1.04 (0.0920)	1.35 (0.662)	<0.001	1.10 (0.332)
Median [Min, Max]	1.02 [0.870, 1.81]	1.16 [0.850, 6.08]		1.04 [0.850, 6.08]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
Prothrombin_time_activity				
Mean (SD)	96.2 (12.9)	75.3 (22.4)	<0.001	92.0 (17.4)
Median [Min, Max]	97.0 [43.0, 130]	80.0 [10.0, 136]		94.0 [10.0, 136]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
APTTactivated_partial_thromboplastin_time				
Mean (SD)	30.9 (3.93)	33.3 (7.91)	<0.001	31.4 (5.09)
Median [Min, Max]	30.5 [17.4, 74.2]	31.8 [24.3, 103]		30.6 [17.4, 103]
Missing	0 (0%)	1 (0.6%)		1 (0.1%)
TTthrombin_time				
Mean (SD)	14.5 (1.76)	16.3 (16.3)	0.14	14.9 (7.46)
Median [Min, Max]	14.5 [11.0, 46.7]	15.0 [12.2, 231]		14.6 [11.0, 231]
Missing	0 (0%)	4 (2.2%)		4 (0.4%)
Fibrinogen_content				
Mean (SD)	328 (86.0)	297 (75.9)	<0.001	322 (84.9)
Median [Min, Max]	314 [97.0, 743]	286 [141, 536]		308 [97.0, 743]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
Ddimer				
Mean (SD)	401 (1130)	1420 (5630)	0.018	602 (2720)
Median [Min, Max]	136 [0, 15900]	189 [1.00, 63200]		141 [0, 63200]
Missing	0 (0%)	5 (2.8%)		5 (0.6%)

Table 4 Liver and kidney function indicators in Normal and Pulmonary Hypertension (PH) Patients

	Normal	PH (N=181)	P value	Overall(N=895)
ALT				
Mean (SD)	26.0 (45.1)	84.0 (394)	0.091	35.2 (163)
Median [Min, Max]	18.0 [3.00, 867]	19.5 [1.00, 3330]		18.0 [1.00, 3330]
Missing	0 (0%)	47 (26.0%)		47 (5.3%)
AST				
Mean (SD)	24.6 (33.2)	95.0 (556)	0.092	38.7 (252)
Median [Min, Max]	19.0 [4.00, 574]	22.0 [9.00, 6160]		20.0 [4.00, 6160]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
ASTratioALT				
Mean (SD)	1.22 (0.611)	8.42 (16.9)	<0.001	2.65 (8.06)
Median [Min, Max]	1.11 [0.270, 6.86]	1.44 [0.350, 133]		1.17 [0.270, 133]
Missing	0 (0%)	4 (2.2%)		4 (0.4%)
TBIL				
Mean (SD)	12.4 (12.4)	16.4 (17.0)	0.003	13.2 (13.6)
Median [Min, Max]	10.6 [1.50, 275]	12.0 [0.540, 103]		10.6 [0.540, 275]

Missing	0 (0%)	2 (1.1%)		2 (0.2%)
DBIL				
Mean (SD)	4.36 (12.3)	8.86 (8.70)	<0.001	5.26 (11.8)
Median [Min, Max]	2.70 [0.300, 231]	5.60 [0.600, 51.1]		2.90 [0.300, 231]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
IDBIL				
Mean (SD)	8.66 (5.09)	12.3 (11.1)	<0.001	9.39 (6.90)
Median [Min, Max]	7.70 [0.800, 44.1]	9.40 [1.10, 94.1]		7.90 [0.800, 94.1]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
TotalProtein				
Mean (SD)	68.6 (9.60)	54.3 (25.4)	<0.001	65.7 (15.3)
Median [Min, Max]	68.9 [3.04, 105]	64.9 [3.20, 88.5]		68.2 [3.04, 105]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
Alb				
Mean (SD)	40.9 (5.48)	45.9 (12.9)	<0.001	41.9 (7.82)
Median [Min, Max]	40.9 [15.7, 76.5]	41.5 [22.1, 86.5]		41.0 [15.7, 86.5]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
Glb				
Mean (SD)	31.0 (26.0)	30.6 (6.03)	0.698	31.0 (23.4)
Median [Min, Max]	28.3 [11.7, 319]	30.1 [16.3, 47.0]		28.7 [11.7, 319]

Table 4 Liver and kidney function indicators in Normal and Pulmonary Hypertension (PH) Patients (continued)

Missing	0 (0%)	2 (1.1%)		2 (0.2%)
AlbratioGlb				
Mean (SD)	1.71 (2.31)	8.35 (12.7)	<0.001	3.04 (6.61)
Median [Min, Max]	1.46 [0.450, 28.6]	1.51 [0.730, 48.9]		1.46 [0.450, 48.9]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
gamma_GGT				
Mean (SD)	37.8 (62.7)	44.9 (79.8)	0.273	39.3 (66.5)
Median [Min, Max]	21.5 [0.630, 927]	23.0 [0.600, 752]		22.0 [0.600, 927]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
ALP_alkaline_phosphatase				
Mean (SD)	80.3 (53.7)	73.1 (63.9)	0.167	78.9 (56.0)
Median [Min, Max]	72.0 [3.59, 612]	64.0 [11.0, 768]		70.0 [3.59, 768]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
TBA_total_bile_acid				
Mean (SD)	6.84 (22.1)	26.1 (37.4)	<0.001	10.7 (27.0)
Median [Min, Max]	3.10 [0.100, 303]	6.60 [0.400, 204]		3.50 [0.100, 303]
Missing	0 (0%)	2 (1.1%)		2 (0.2%)
BUN				
Mean (SD)	6.58 (10.0)	7.44 (9.33)	0.279	6.75 (9.91)
Median [Min, Max]	5.14 [1.20, 105]	5.82 [0.400, 112]		5.25 [0.400, 112]
Missing	0 (0%)	3 (1.7%)		3 (0.3%)
creatinine				
Mean (SD)	80.9 (104)	62.0 (43.5)	<0.001	77.1 (95.6)
Median [Min, Max]	65.0 [2.23, 1140]	62.6 [1.95, 255]		64.2 [1.95, 1140]
Missing	0 (0%)	4 (2.2%)		4 (0.4%)
Uric_Acid				
Mean (SD)	327 (103)	331 (193)	0.796	328 (126)

Median [Min, Max]	317 [0.760, 839]	335 [44.8, 1000]		321 [0.760, 1000]
Missing	0 (0%)	4 (2.2%)		4 (0.4%)
Carbon_dioxide				
Mean (SD)	23.9 (4.21)	110 (171)	<0.001	41.0 (83.5)
Median [Min, Max]	23.8 [1.00, 34.9]	24.8 [12.3, 842]		23.9 [1.00, 842]
Missing	0 (0%)	4 (2.2%)		4 (0.4%)
Serum_cystatin_C				
Mean (SD)	1.04 (0.794)	6.99 (10.5)	<0.001	2.18 (5.20)
Median [Min, Max]	0.900 [0.360, 7.59]	1.11 [0.510, 38.1]		0.910 [0.360, 38.1]

Table 4 Liver and kidney function indicators in Normal and Pulmonary Hypertension (PH) Patients (continued)

Missing	0 (0%)	12 (6.6%)		12 (1.3%)

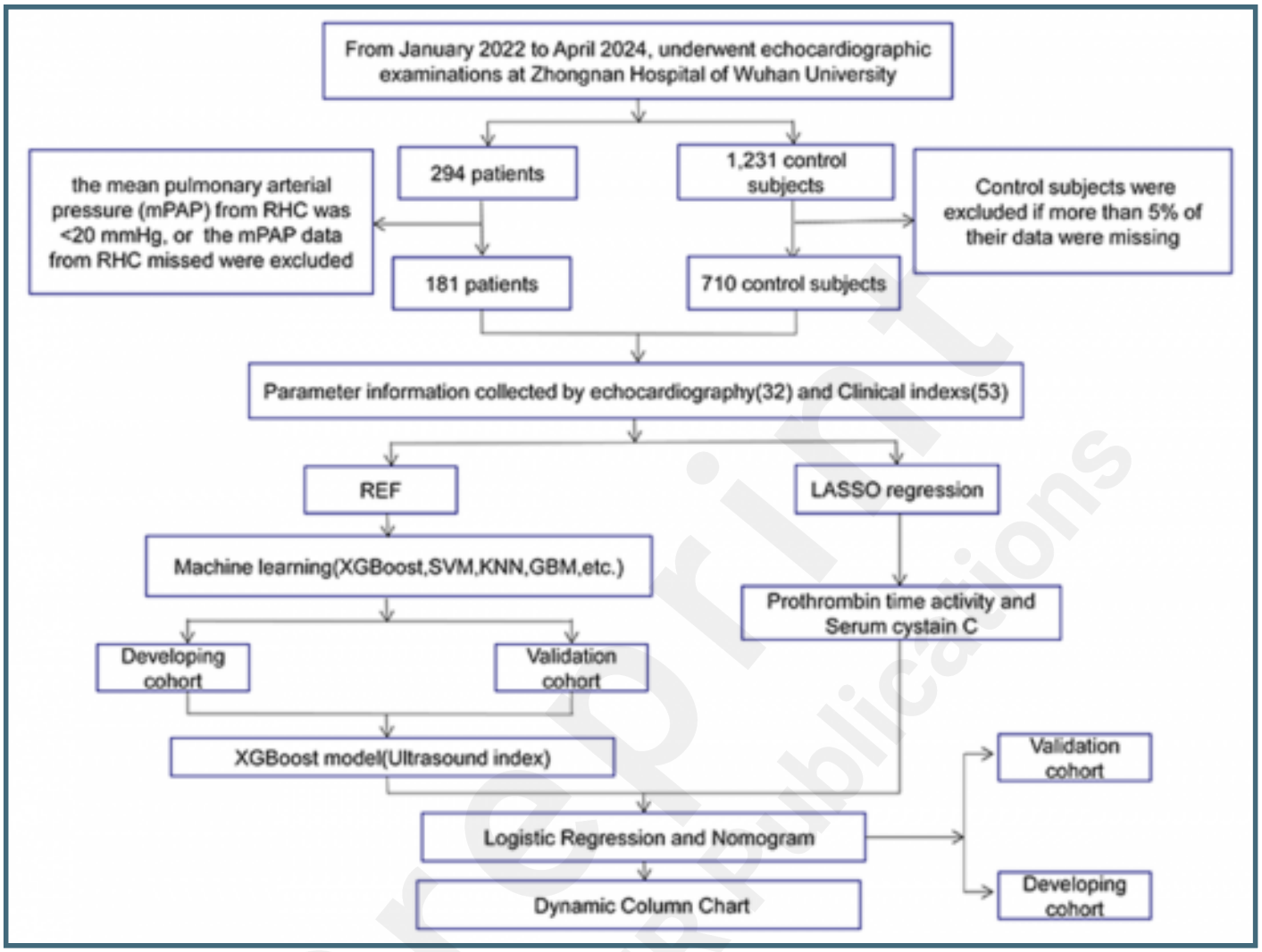
Table 5 Electrolyte level in Normal and Pulmonary Hypertension (PH) Patients

	Normal	PH (N=181)	P value	Overall(N=895)
Potassium				
Mean (SD)	3.94 (0.462)	3.42 (1.22)	<0.001	3.84 (0.707)
Median [Min, Max]	3.94 [0.990, 6.27]	3.84 [0.720, 5.72]		3.93 [0.720, 6.27]
Missing	0 (0%)	10 (5.5%)		10 (1.1%)
Sodium				
Mean (SD)	138 (13.9)	108 (57.5)	<0.001	132 (30.8)
Median [Min, Max]	140 [0.940, 148]	138 [2.93, 155]		140 [0.940, 155]
Missing	0 (0%)	6 (3.3%)		6 (0.7%)
Chlorine				
Mean (SD)	103 (11.7)	113 (15.0)	<0.001	105 (13.0)
Median [Min, Max]	105 [2.22, 114]	106 [91.0, 146]		105 [2.22, 146]
Missing	0 (0%)	6 (3.3%)		6 (0.7%)
Calcium				
Mean (SD)	2.57 (5.43)	26.0 (43.1)	<0.001	7.19 (21.8)
Median [Min, Max]	2.30 [0.390, 105]	2.32 [1.78, 111]		2.30 [0.390, 111]
Missing	0 (0%)	6 (3.3%)		6 (0.7%)
Magnesium				
Mean (SD)	2.40 (26.1)	1.20 (0.596)	0.22	2.17 (23.5)
Median [Min, Max]	0.880 [0.380, 667]	0.890 [0.530, 2.54]		0.880 [0.380, 667]
Missing	0 (0%)	14 (7.7%)		14 (1.6%)
Phosphorus				
Mean (SD)	5.38 (45.6)	1.16 (0.281)	0.014	4.57 (41.0)
Median [Min, Max]	1.18 [0.310, 784]	1.17 [0.660, 2.32]		1.17 [0.310, 784]
Missing	0 (0%)	11 (6.1%)		11 (1.2%)

Supplementary Files

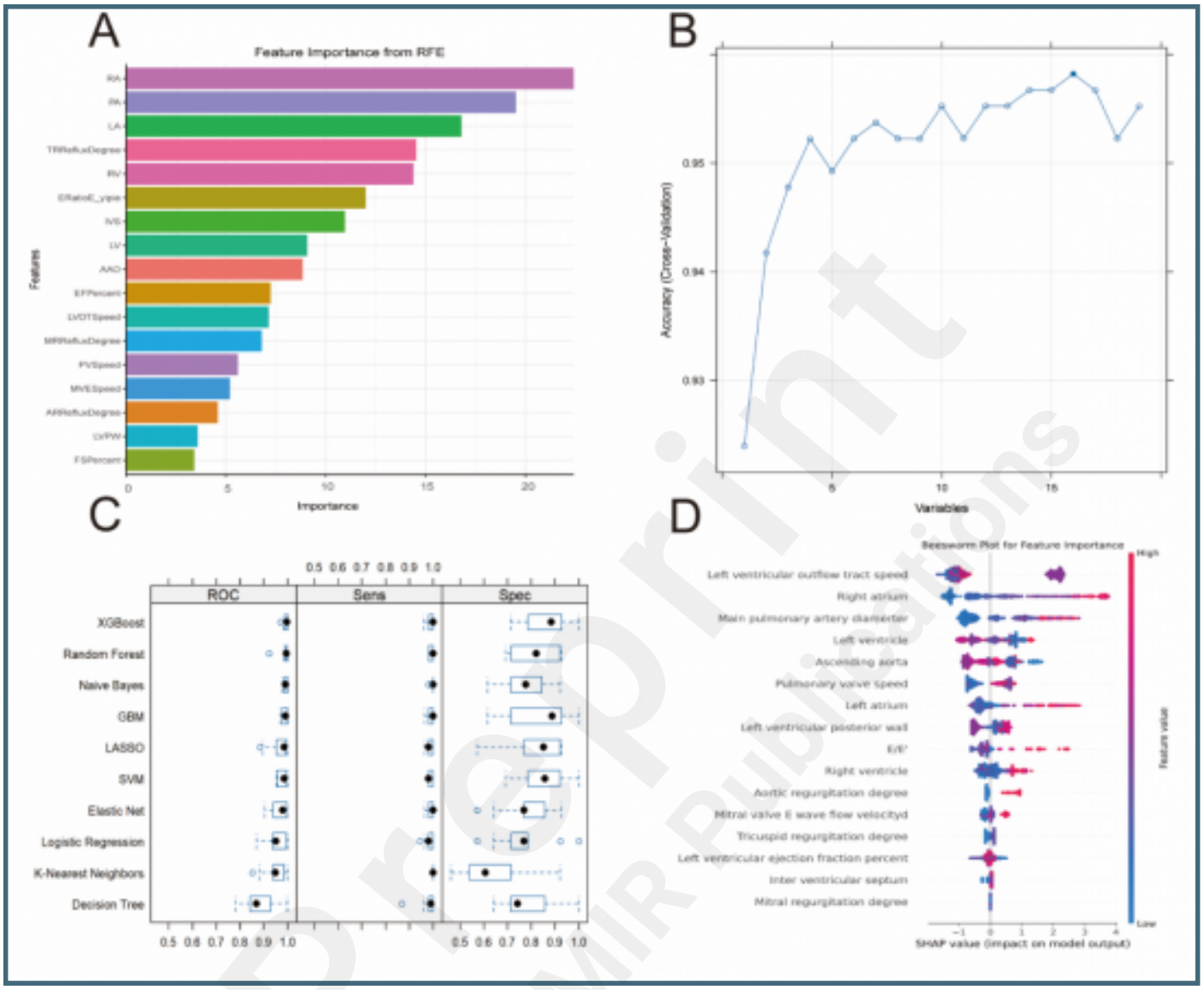
??s

Workflow of the analysis.

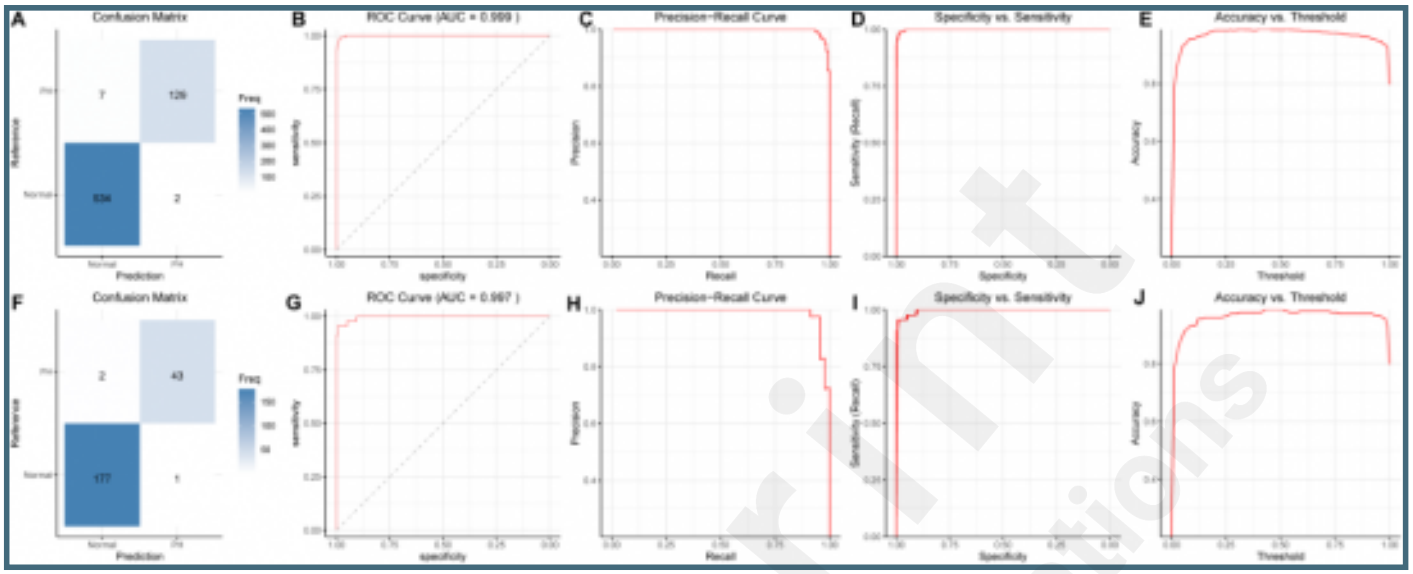


Figures

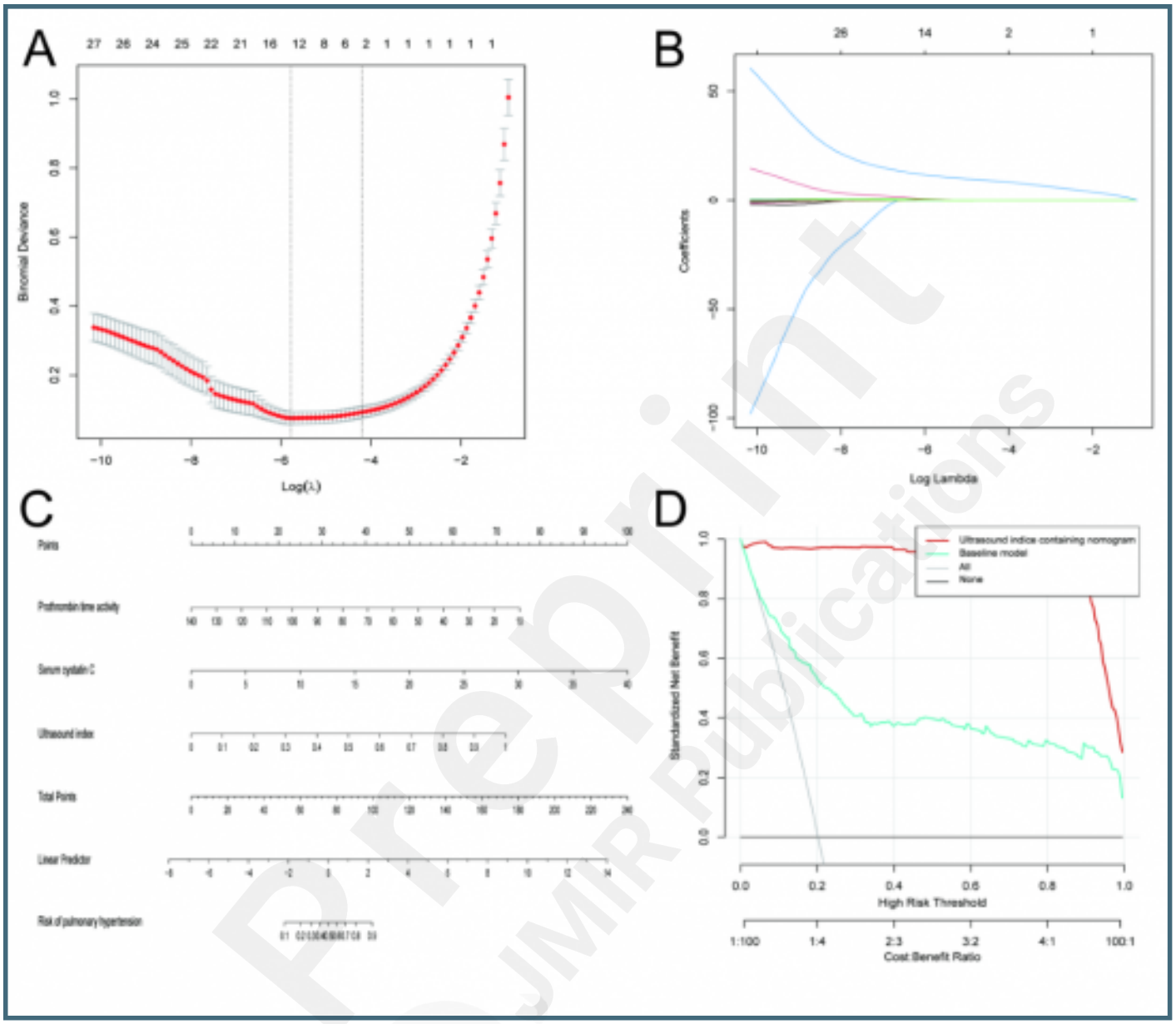
(A) eDeviance Plot from Lasso Regression. (B) Coefficient Path for Lasso Regression. (C) Nomogram.



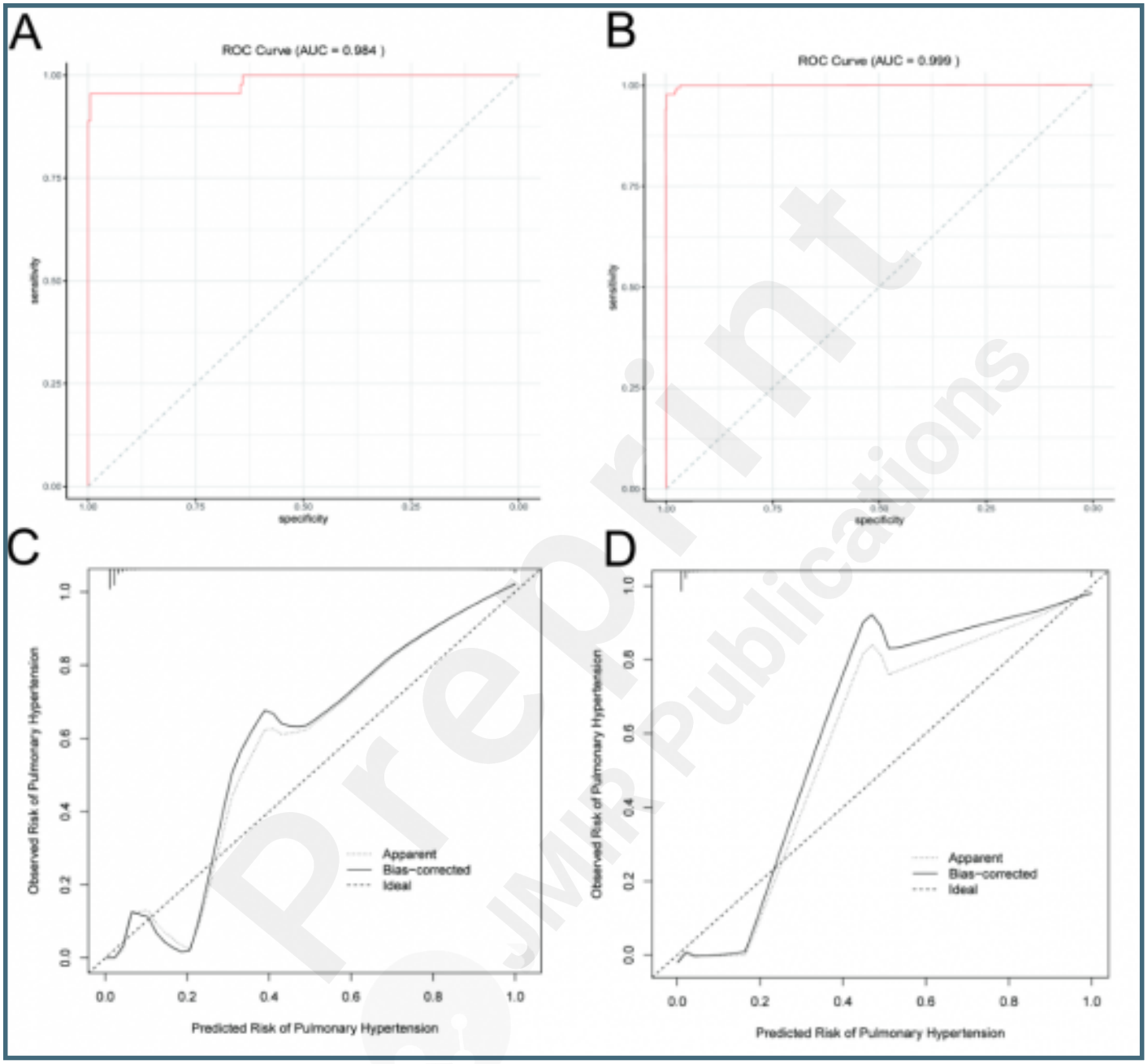
(A) Confusion Matrix for Developing cohort. (B) ROC Curve for Developing cohort (AUC = 0.999). (C) Precision-Recall Curve for Developing cohort. (D) Specificity vs Sensitivity for Developing cohort. (E) Accuracy vs Threshold for Developing cohort. (F) Confusion Matrix for Validation cohort. (G) ROC Curve for Validation cohort (AUC = 0.997). (H) Precision-Recall Curve for Validation cohort. (I) Specificity vs Sensitivity for Validation cohort. (J) Accuracy vs Threshold for Validation cohort.



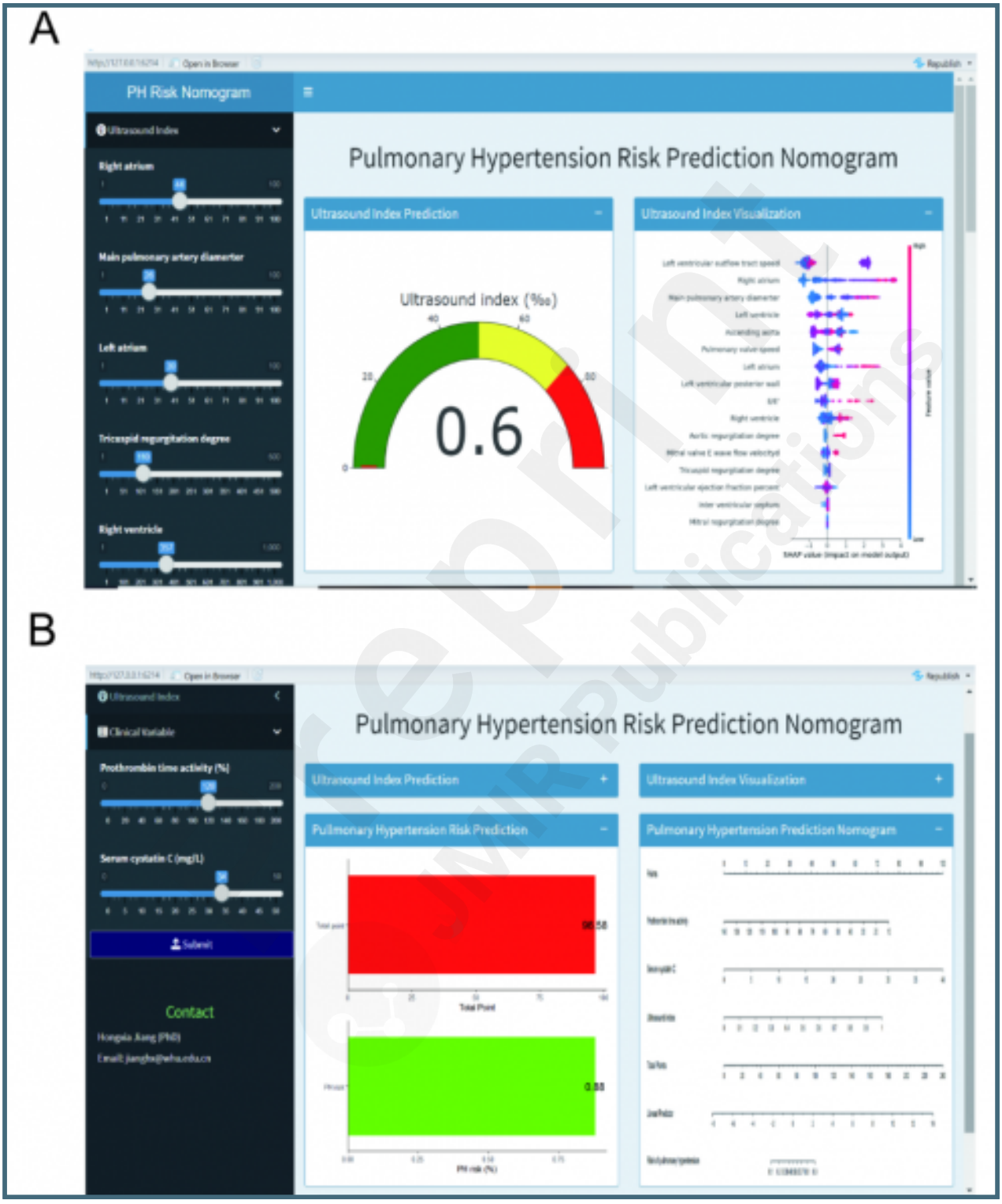
(A) Deviance Plot from Lasso Regression. (B) Coefficient Path for Lasso Regression. (C) Nomogram for Risk Prediction of Pulmonary Hypertension. (D) Decision Curve Analysis.



(A) ROC of the nomogram in the developing cohort. (B) ROC of the nomogram in validation cohort. (C) Calibration curve of the nomogram in the developing cohort. (D) Calibration curve of the nomogram in the validation cohort.



(A) Ultrasound-based pulmonary hypertension risk calculator showing index prediction and SHAP analysis explaining feature contributions to model output. The PH risk prediction nomogram uses clinical variables to calculate a total score, estimating the corresponding PH risk percentage.



Multimedia Appendixes

Table 4. Liver and kidney function indicators in normal and pulmonary hypertension (PH) patients.
URL: http://asset.jmir.pub/assets/f7167c5dab11dca841127642ade2ad2b.docx

Table 1.Comparison of demographic and echocardiographic parameters between normal and pulmonary hypertension (PH) patients.
URL: http://asset.jmir.pub/assets/2d382dae87a0fd92bd53d9637eef154e.doc

Table 2. blood routine index in normal and pulmonary hypertension (PH) patients.
URL: http://asset.jmir.pub/assets/b59e9b694844cebabe8e478c4060bd51.docx

Table 3. Indicators of coagulation function in normal and pulmonary hypertension (PH) patients.
URL: http://asset.jmir.pub/assets/ad4069a2930b166644b02d0035587fce.docx

Table 5. Electrolyte level in normal and pulmonary hypertension (PH) patients.
URL: http://asset.jmir.pub/assets/86b8168403be4d24afccf0bff2760efb.docx