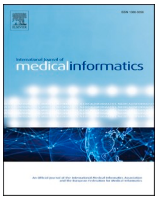
International Journal of Medical Informatics 195 (2025) 105741

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Development of machine learning-based models to predict congenital heart disease: A matched case-control study

Shutong Zhang [a], Chenxi Kang [a], Jing Cui [a], Haodan Xue [a], Shanshan Zhao [a], Yukui Chen [a], Haixia Lu [a], Lu Ye [b], Duolao Wang [c,d], Fangyao Chen [a], Yaling Zhao [a], Leilei Pei [a,*], Pengfei Qu [e,f,**]

[a] *Department of Epidemiology and Health Statistics, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi 710061, China*
[b] *Shaanxi Eye Hospital, Xi'an People's Hospital (Xi'an Fourth Hospital), Xi'an, China*
[c] *Biostatistics Unit, Department of Clinical Sciences, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK*
[d] *Department of Neurology, Guangdong Key Laboratory of Age-Related Cardiac and Cerebral Diseases, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China*
[e] *Translational Medicine Center, Northwest Women's and Children's Hospital, Xi'an 710061, China*
[f] *Central Laboratory, Beijing Obstetrics and Gynecology Hospital, Capital Medical University, Chaoyang, Beijing 100026, China*

ARTICLE INFO

Keywords:
Congenital heart disease
Prediction
Risk score
Machine learning
Web tool

ABSTRACT

Background: The current congenital heart disease (CHD) prediction tools lack adequate interpretability and convenience, hindering the development of personalized CHD management strategies. We developed a machine learning-based risk stratification model for CHD prediction.

Methods: This study utilized data from 1,759 participants in a case-control study of CHD conducted across six birth defects surveillance hospitals located in Xi'an, Shaanxi Province, Northwest China, spanning from January 2014 to December 2016. The data was partitioned into training and testing datasets with a ratio of 7:3. Predictors were selected from a total of 47 input variables through the Least Absolute Shrinkage and Selection Operator (LASSO). Five machine learning algorithms were used to build the CHD risk prediction models. Model performance was assessed based on a range of learning metrics, including the area under the receiver operating characteristic curve (AUROC), F1 score, and Brier score. Permutation feature importance was employed to elucidate the prediction model. The best-performing model was used to conduct the risk scores.

Results: The eXtreme Gradient Boosting (XGB) model demonstrated superior performance among CHD prediction models, achieving an AUROC of 0.772 (95 % CI 0.728, 0.817) in the testing dataset and 0.738 (0.699, 0.775) in the external validation dataset. The pivotal predictors (top 3) identified by the model included living in rural areas, the low wealth index, and folic acid supplements (<90 days). The resultant risk score exhibited robust calibration capabilities. Utilizing the risk scores, participants were stratified into low, moderate, and high-risk categories, signifying substantial variations in CHD risk.

Conclusion: This study underscores the feasibility and efficacy of employing a machine learning-based approach for CHD prediction. The risk scores exhibited potential in identifying pregnant women at high risk for fetal CHD, offering valuable insights for guiding primary prevention and CHD management.

1. Introduction

Congenital heart disease (CHD) significantly contributes to infant and child mortality and morbidity, and is the most common birth defect accounting for one-third of all congenital abnormalities worldwide.[1,2] According to the Global Burden of Disease (GBD) study, it was estimated that approximately 3 million newborns were born with congenital heart anomalies in 2019 worldwide.[3] In China, the prevalence of CHD was 8.94 per 1000 live births in 2014,[4] and the cumulative lifetime economic burden associated with new CHD cases exceeded 2 billion USD.

[*] Corresponding author: Leilei Pei, Department of Epidemiology and Health Statistics, School of Public Health, Xi'an Jiaotong University Health Science Center, Xi'an, Shaanxi 710061, China.

[**] Corresponding author: Pengfei Qu, Translational Medicine Center, Northwest Women's and Children's Hospital, China.

E-mail addresses: pll_paper@126.com (L. Pei), xinxi3057@163.com (P. Qu).

https://doi.org/10.1016/j.ijmedinf.2024.105741
Received 13 December 2023; Received in revised form 8 October 2024; Accepted 30 November 2024
Available online 2 December 2024

[5] Approximately 9 % to 18 % of CHD cases are attributed to simple chromosomal aberrations or gene mutations. More cases are caused by environmental factors or the gene–environment interaction. While the etiology of CHD remains poorly understood, the role of the environment in CHD is increasingly recognized, such as alcohol, tobacco, loud noise, and drugs, as well as biological factors. [6–8] From a public health perspective, it stands as a paramount concern for the primary prevention of CHD. Therefore, it is imperative, based on CHD risk factors, to predict individual CHD risk to provide more detailed screening and preventive interventions for CHD.

Machine learning (ML) methods are extensively employed for disease prediction by applying computer algorithms to large datasets containing a multitude of multidimensional variables to capture high-dimensional nonlinear relationships among clinical features for data-driven outcome prediction. [9,10] Most studies used echocardiography, MRI and other medical images to build deep learning models for CHD prediction, achieving better accuracy and sensitivity. [11,12] Using convolutional neural networks, Arnout et al.[13] differentiated between normal hearts and complex CHD using echocardiograms. However, deep learning models often lack interpretability. Support vector machine (SVM), random forests (RF), and logistic regression (LR) were utilized by Luo et al.[14] to predict CHD risk. They categorized risk factors into nine crucial indicator variables and summed these factors to reduce data dimensionality. Li et al.[15] developed a prediction model for CHD using artificial neural networks, based on a case-control study involving 358 subjects. Kuar et al.[16] enhanced random forest-based CHD prediction by employing unsupervised learning clustering, using the same data as Luo et al. Despite the application of ML in these studies for CHD prediction, traditional statistical methods like logistic regression, which are susceptible to feature correlation and nonlinear relationships, continued to be widely used. [17,18].

These studies did not evaluate the clinical utility of their predictive models, and the lack of access to personalized management strategies remains an unmet need. To advance this research, we built upon the foundations laid by these studies and proposed a novel ML-based model for CHD prediction. We evaluated multiple models for discrimination, calibration, and clinical utility, and constructed an interpretable risk score. Our model serves as an initial screening tool to identify high-risk fetal CHD early, guiding healthcare professionals in prenatal management and prevention, thereby furthering the current understanding and clinical application of CHD prediction.

2. Methods

2.1. Study design and participants

The study utilized data from a case-control study of CHD conducted across six birth defects surveillance hospitals in Xi'an, Shaanxi Province, Northwest China, from January 2014 to December 2016. Approval was obtained from the Human Research Ethics Committee of Xi'an Jiaotong University Health Sciences Center (No. 20120008). Participants were fully informed about the study and signed an informed consent form before the investigation.

Case inclusion criteria were: termination of pregnancy from January 2014 to December 2016; perinatal infants (both live and stillborn) diagnosed with CHD according to ICD-10 classification criteria from 28 weeks gestation to 7 days post-birth, and fetuses under 28 weeks diagnosed with CHD by ultrasound and other examinations in the hospital; malformed fetuses other than CHD were excluded. In the control group, singleton newborns without birth defects in the same hospital were selected in a 1:2 matching method according to age and birth date. Subjects were excluded if the perinatal diagnosis was unclear or if the parents could not answer the questionnaire accurately because of psychiatric symptoms or serious illness.

2.2. Data collection

A face-to-face survey, administered by trained personnel from Xi'an Jiaotong University Health Sciences Center, employed a standardized questionnaire developed by the university. The questionnaire covered socio-demographics, lifestyle, environmental factors, nutritional supplementation, medication use, disease, and pregnancy history. Pretested in a pilot study, and detailed interviewer guides were developed.

Prenatal diagnostic, clinic results, physical examination, ultrasound reports, and medical history were collected through hospital medical records. Cardiovascular epidemiologists, obstetricians, pediatricians, and imaging physicians reviewed the questionnaires and made clinical diagnoses of cases. Family surveys were conducted when necessary, ensuring comprehensive information. Telephone follow-ups within a year post-birth confirmed diagnoses. Participants were excluded if data for the variables of interest were missing. Fig. 1 illustrates the sample selection process.

2.3. Feature selection and model development

Based on the literature review and data availability,[6,19,20] we initially selected 47 variables encompassing socio-demographic characteristics, family history of CHD, pregnancy history, and periconceptional environmental exposures. The definition of each variable is detailed in the online supplemental Text 1.

In this study, model development and reporting adhered to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) + Artificial intelligence (AI) statement.[21] The data were divided into training and testing datasets at a 7:3 ratio. The Least Absolute Shrinkage and Selection Operator (LASSO) method was employed for feature selection, executed on the training dataset to prevent data leakage and result bias. To model the risk prediction of CHD, five ML classification models—LR, SVM, RF, eXtreme Gradient Boosting (XGB), and Neural Network (NN)—were selected. These risk prediction models were constructed using the selected features. Additionally, hyperparameters were tuned through ten-fold cross-validation and GridSearch within each ML model. Cut-off values were adjusted based on the maximum Youden index observed in the Receiver Operating Characteristic (ROC) curve of each ML model.

The test data exclusively served the purpose of assessing the final performance of the classifiers. The evaluation of model performance was executed using various metrics (supplemental Text 2), mainly considering: the area under the ROC curve (AUROC), F1 score, and Brier score. The predictive accuracy was represented by ROC curves and precision-recall (P-R) curves, and it was quantified using AUROC and average precision (AP, i.e., the area under the P-R curve). To gauge the agreement between predicted and observed risk, we employed the Brier score and calibration plots. Furthermore, a decision curve analysis (DCA) was conducted to measure the net benefit (NB) associated with identifying and intervening in genuinely high-risk patients. Ultimately, the best-performing model was determined. An external dataset was used to validate the optimal model's performance (supplementary Text 3).

2.4. Feature importance

To identify the main predictors of CHD in the study population, we measured the importance of each permutation feature in the optimal model. Permutation feature importance was ascertained by evaluating the model's increase in prediction error resulting from the permutation of a feature's value. [22] Intuitively, if a variable's value is randomly permuted while holding all other variables constant, predictions are then generated based on the modified dataset. A substantial decrease in the model's predictive performance indicates a higher relative importance for that variable. As relative importance is not a fixed scaled value, the results are expressed in terms of scaled importance, meaning the ratio between the relative importance of each variable and the highest

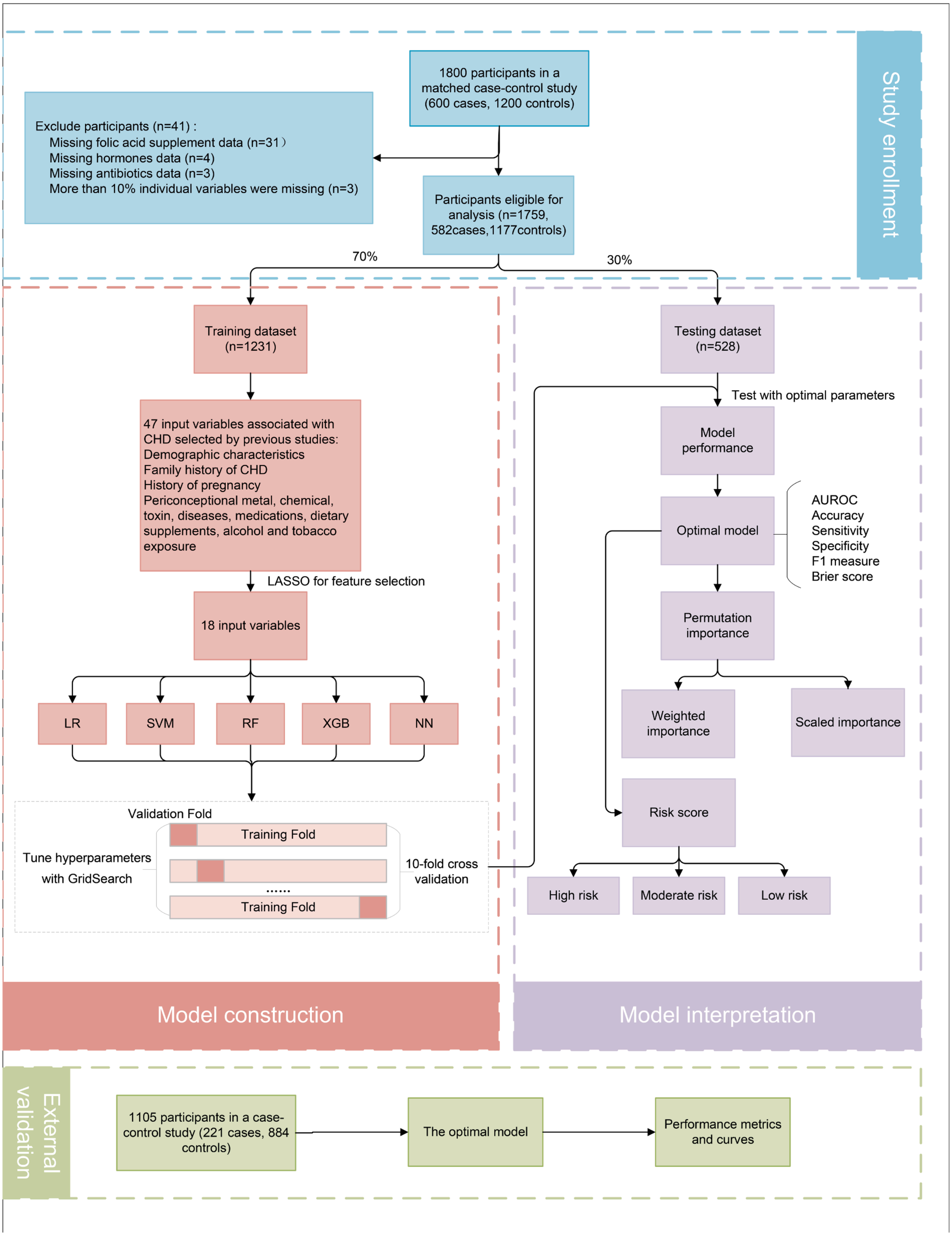

Study enrollment
1800 participants in a matched case-control study (600 cases, 1200 controls)
Exclude participants (n=41) :
Missing folic acid supplement data (n=31）
Missing hormones data (n=4)
Missing antibiotics data (n=3)
More than 10% individual variables were missing (n=3)
Participants eligible for analysis (n=1759, 582cases,1177controls)
70%
30%
Training dataset (n=1231)
47 input variables associated with CHD selected by previous studies:
Demographic characteristics
Family history of CHD
History of pregnancy
Periconceptional metal, chemical, toxin, diseases, medications, dietary supplements, alcohol and tobacco exposure
LASSO for feature selection
18 input variables
LR
SVM
RF
XGB
NN
Validation Fold
Training Fold
Tune hyperparameters with GridSearch
......
Training Fold
10-fold cross validation
Model construction
Testing dataset (n=528)
Test with optimal parameters
Model performance
Optimal model
AUROC
Accuracy
Sensitivity
Specificity
F1 measure
Brier score
Permutation importance
Weighted importance
Scaled importance
Risk score
High risk
Moderate risk
Low risk
Model interpretation
External validation
1105 participants in a case-control study (221 cases, 884 controls)
The optimal model
Performance metrics and curves

Fig. 1. Flowchart.

relative importance value. The scaled importance provides direct comparisons between the importance of each variable and the most important variable.

2.5. *Deciles of risk score and risk groups*

Risk scores were derived from the optimal model and then divided into deciles for both the training and testing datasets. The observed CHD frequency against the average risk score within each decile was plotted. Participants were subsequently grouped into three risk groups: low, moderate, and high, based on outcomes observed in the training dataset. Additionally, calibration of the risk score was assessed within age subgroups across all participants, exploring the performance of the risk score in distinct age groups (<35 years and ≥ 35 years).

2.6. Statistical analysis

Categorical variables were presented as counts (%) and compared between groups using the chi-square test.[23] Continuous variables were expressed as mean (SD) or median (IQR), and were compared between groups using Student's *t*-test[24] when normally distributed, and Wilcoxon rank sums test[25] when non-normally distributed. Furthermore, we developed a web application using the Streamlit package[26] in Python. A significance level of $p < 0.05$ denoted statistical significance. All statistical analyses were executed employing Python version 3.9.12.

3. Results

3.1. Characteristics of subjects

A total of 1759 participants were included in the study, comprising 582 CHD cases and 1177 controls. Participants were randomly divided into a training dataset (n = 1231) and a testing dataset (n = 528). Essential demographic characteristics and the available predictors are presented in Table 1. Out of the participants in the training and testing datasets, 407 and 175 individuals respectively suffered from CHD, with a proportion of 33.1 %.

Table 1
Baseline characteristics of study participants.

	Training dataset (n = 1231)	**Testing dataset (n = 528)**
Socio-demographic characteristics		
Maternal age, years	28 (25–31)	28 (26–30)
Paternal age, years	29 (27.5–31)	29 (28–32)
Maternal nationality		
Han	1215 (98.7)	522 (98.9)
Minorities	16 (1.3)	6 (1.1)
Paternal nationality		
Han	1208 (98.1)	524 (99.2)
Minorities	23 (1.9)	4 (0.8)
Maternal residence		
Urban	682 (55.4)	297 (56.2)
Rural	549 (44.6)	231 (43.8)
Maternal education		
College or above	859 (69.8)	381 (72.2)
Senior high school or below	372 (30.2)	147(27.8)
Household wealth index		
Moderate or high	775 (63.0)	357 (67.6)
low	456 (37.0)	171 (32.4)
Pregnancy history		
Parity		
<2	855 (69.5)	392 (74.2)
≥2	376 (30.5)	136 (25.8)
Labor induction	53 (4.3)	17 (3.2)
Periconceptional environmental exposures		
Living near mines or factories	77 (6.3)	36 (6.8)
Noise	222 (18.0)	88 (16.7)
Chemicals and toxins	103 (8.4)	48 (9.1)
Negative emotions	168 (13.6)	70 (13.3)
Cold	311 (25.3)	133 (25.2)
Anemia	67 (5.4)	20 (3.8)
Antibiotics	44 (3.6)	19 (3.6)
Hormones	38 (3.0)	20 (3.8)
Folic acid supplement		
<90 days	553 (44.9)	238 (45.1)
≥90 days	678 (55.1)	290 (54.9)
Multivitamin supplement	111 (9.0)	46 (8.7)
Iron supplement	60 (4.9)	26 (4.9)
Alcohol	16 (1.3)	10 (1.9)
Passive smoking	588 (47.8)	244 (46.2)
CHD	407 (33.1)	175 (33.1)

Data are n (%), or median (IQR).
Abbreviations: CHD, congenital heart disease.

3.2. Selected predictors and model performance

In the study, 18 predictors were chosen from 47 candidate features through LASSO, including 3 predictors from socio-demographic characteristics, 2 predictors from the pregnancy history, and 13 predictors from periconceptional environmental exposures. Table 2 shows the performance metrics of five prediction models constructed based on these 18 predictors. The XGB model performed the best among all algorithms with better evaluation scores (AUROC = 0.772, F1 score = 0.610, Brier score = 0.174).

The discriminative performance of the predictive models was visualized through ROC curves and P-R curves (Fig. 2a, b). XGB exhibited the highest AUROC (0.772, 95 % CI 0.728, 0.817; Table 2), and the best AP (0.631, 95 % CI 0.527, 0.680; Fig. 2b). As for the calibration performance, XGB achieved the lowest Brier score, and its calibration curve was close to the diagonal line (Fig. 2c). The DCA showed that the clinical NB of XGB was higher compared to several competing intervention strategies. At reasonable threshold probabilities (e.g., Pt = 0.4), compared with the nontreatment strategy, XGB outperformed the other models, yielding the highest NB ($NB_{XGB} = 0.125$) while the RF model yielded the lowest ($NB_{RF} = 0.074$). These results indicated that for every 100 subjects, the XGB identified 12 true-positive subjects who should receive the intervention, whereas the RF model only identified 7 true-positive subjects (Fig. 2d).

Performance metrics for the XGB model on the external validation dataset are detailed in supplementary Text 3. When assessed on this external dataset, the model achieved an AUC of 0.738 (95 % CI 0.699, 0.775).

3.3. Importance of features

Fig. 3 shows the scaled importance of the main predictors (top 8), living in a rural area (1.00), low wealth index (0.63), folic acid supplement < 90 days (0.54), parity ≥ 2 (0.39), negative emotions (0.37),

Table 2
Performance metrics comparison of five ML models in the testing dataset.

	LR	SVM	RF	XGB	NN
Threshold	0.441	0.277	0.466	0.295	0.367
AUROC (95 % CI)	0.767 (0.727, 0.807)	0.754 (0.709, 0.780)	0.760 (0.709, 0.780)	0.772 (0.728, 0.817)	0.758 (0.711, 0.804)
Accuracy, % (95 % CI)	68.94 (64.96, 72.73)	67.05 (63.26, 71.02)	71.59 (67.80, 75.19)	70.45 (66.48, 74.05)	71.59 (67.61, 75.00)
Balanced accuracy, % (95 % CI)	69.57 (65.16, 73.68)	68.15 (64.20, 72.30)	70.40 (66.17, 74.44)	70.27 (66.10, 74.43)	68.38 (64.22, 72.18)
Sensitivity, % (95 % CI)	71.43 (64.67, 77.89)	71.43 (64.94, 77.92)	66.86 (59.78, 73.42)	69.71 (62.92, 75.95)	58.86 (51.74, 65.50)
Specificity, % (95 % CI)	67.71 (62.64, 72.21)	64.87 (59.77, 69.47)	73.94 (69.32, 78.28)	70.82 (65.93, 75.21)	77.90 (73.01, 82.27)
F1 score (95 % CI)	0.604 (0.545, 0.659)	0.590 (0.532, 0.640)	0.609 (0.552, 0.660)	0.610 (0.556, 0.659)	0.579 (0.158, 0.202)
Brier score (95 % CI)	0.191 (0.174, 0.209)	0.182 (0.164, 0.200)	0.191 (0.178, 0.205)	0.174 (0.155, 0.195)	0.181 (0.158, 0.202)
ECI	0.830	0.915	0.828	0.927	0.944

Abbreviations: ML, machine learning; LR, logistic regression; SVM, support vector machine; RF, random forest; XGB, extreme gradient boosting; NN, neural network; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; ECI, estimated calibration index.

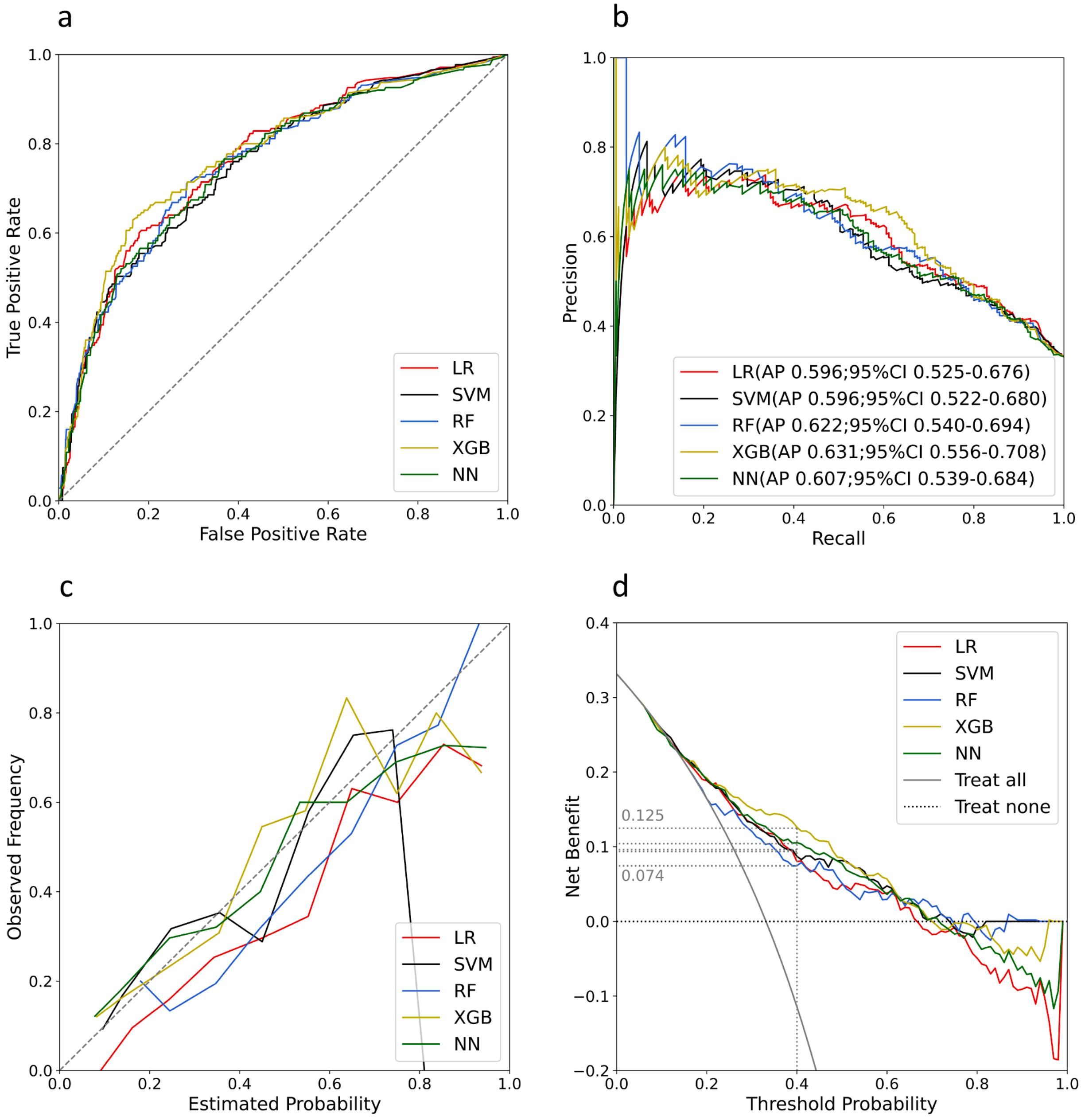


Fig. 2. Model performance evaluation and comparisons in the testing dataset. (a) ROC curves for five ML models. (b) P-R curves for five ML models. (c) Calibration curves for five ML models. (d) Decision curves for five ML models. Abbreviations: ROC, receiver operating characteristic; P-R, precision-recall; AP, average precision, the average precision was defined as the average of precisions across all recall (i.e., sensitivity) values, and equaled the area under the precision-recall curve in the current study; LR, logistic regression; SVM, support vector machine; RF, random forest; XGB, extreme gradient boosting; NN, neural network.

passive smoking (0.36), chemicals and toxins (0.29) and cold (0.26). All predictors' importance is presented in the supplemental Table 3.

3.4. Risk score and classes of risk

Fig. 4 shows the calibration plots of predicted CHD risk score by decile in the training and testing datasets. For further insight, supplemental Table 4 outlines the difference between the predicted risk score and observed CHD frequency by decile in the testing dataset. The risk of CHD is underestimated in most deciles. The degree of underestimation or overestimation was generally limited to within 0.06, except in the ninth and tenth deciles, where the difference exceeded 0.1.

Furthermore, the participants of the training dataset were divided into three risk groups, including low risk (first to sixth deciles), moderate risk (seventh to eighth deciles), and high risk (ninth to tenth deciles). The threshold value for each risk group was determined as the risk score corresponding to their respective deciles. Participants in the testing dataset were also classified into three risk groups according to the determined thresholds (Fig. 5). Out of the 528 participants, 332 (62.9 %) were in the low-risk group, 110 (20.8 %) were in the moderate-risk group, and 86 (16.3 %) were in the high-risk group. In the three risk groups, the proportion of CHD was 17.8 % in the low-risk group, 49.1 % in the moderate-risk group, and 72.1 % in the high-risk group respectively, with a gradually increasing trend across the three risk groups. Supplemental Table 5 showed that participants classified as moderate-risk or high-risk had a higher CHD proportion than low-risk ($p < 0.001$). The moderate-risk group exhibited a 2.8-fold increase, while the high-risk group exhibited a 4.1-fold increase.

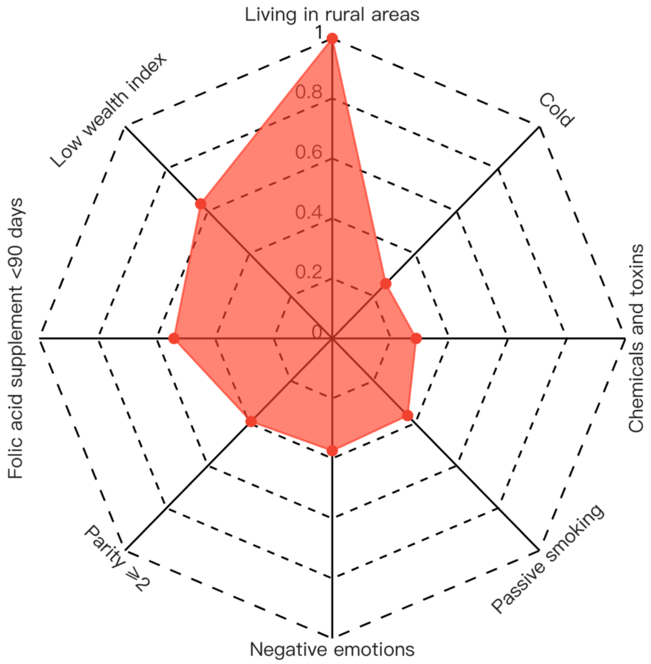


Fig. 3. Radar plot for the eight most important features of CHD.

3.5. Subgroup analysis of maternal age

Predicted risk scores and observed CHD frequencies across the deciles of risk score by maternal age group were presented in supplemental Table 6. Overall, we found slight overestimations of risk in both the < 35 years and ≥ 35 years age groups, with the highest level of overestimation occurring in the ≥ 35 years group (0.103 vs. 0.046 in the < 35 years group). Similarly, the most significant underestimation was also more pronounced in the ≥ 35 years group compared to the < 35 years group (0.078 vs. 0.049).

3.6. Web Deployment tool

We integrated the risk score into a web application designed for individual risk prediction using input predictors. This tool calculates CHD risk and categorizes individuals into risk groups. The web application is accessible online (https://chd-prediction-eenqive3mktxsnuappfheb5.streamlit.app/).

4. Discussion

In this study, we used data from 1,759 pregnant women participating in a case-control study of CHD to develop five ML-based prediction models for assessing fetal CHD risk. Eighteen predictors were identified from health-related candidate variables using LASSO. The XGB model performed the best in prediction discrimination, calibration, and clinical utility. Utilizing the risk score derived from the XGB model, participants were stratified into clinically significant risk groups—low, moderate, and high—for fetal CHD. Risk score can be readily and effectively applied to predicting CHD, particularly in resource-constrained regions where comprehensive pregnancy monitoring systems may be lacking.

Our optimal model, the XGB model, exhibited an AUROC of 0.772 (95% CI, 0.728–0.817), with several distinctions from the best-performing models developed in other pregnant women samples.[14,15] Notably, the best-performing model (AUROC = 0.819) of Luo et al[14] considered nine indicator variables. Each continuous indicator variable covered multiple risk factor items summed to create a "total risk factor score" to reduce data dimensionality. In contrast, our model prioritizes interpretability, retaining information on each risk factor, employing LASSO for feature selection, and calculating permutation feature importance for every feature. Li et al[15] found superior performance (AUROC = 0.87 [95 %CI, 0.75–0.98]) in their NN prediction model, yet our testing dataset's larger sample size (175 CHD cases) compared to their 18 CHD cases provides a more precise performance estimate. Overall, these studies demonstrates the efficacy of ML models in predicting CHD across diverse samples and research teams.

It was found that maternal residence in rural areas was the most important characteristic of fetal CHD. Previous studies in Inner Mongolia, China, and Ontario, Canada, have shown that living in rural areas can increase the risk of CHD.[27,28]Rural environments

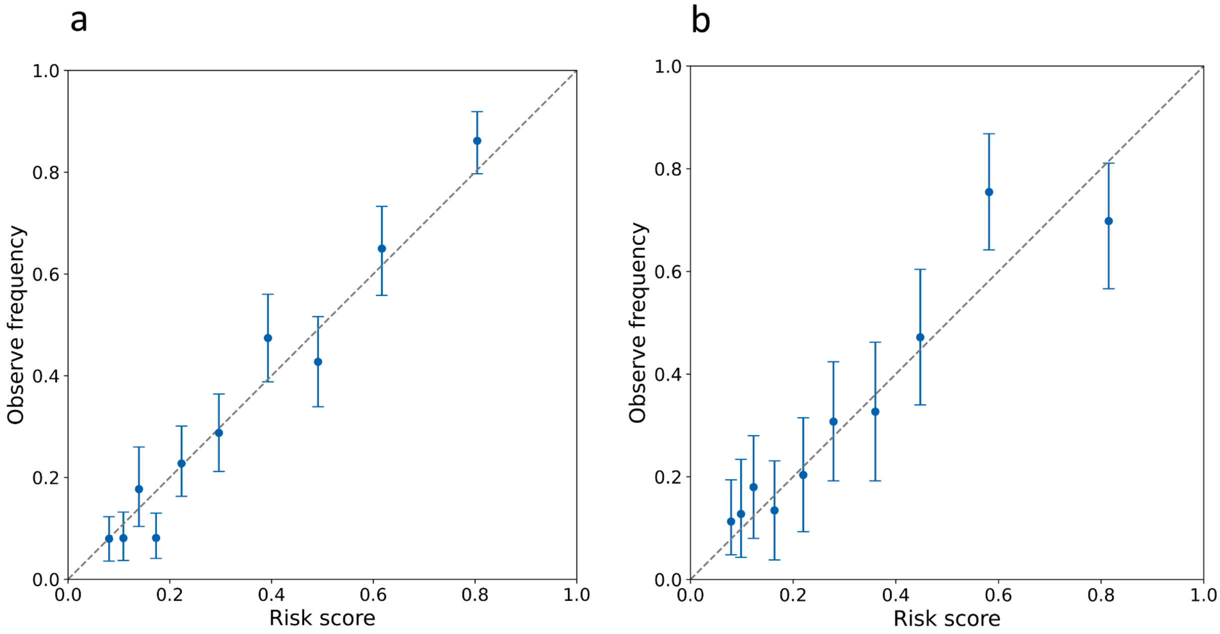


Fig. 4. Calibration plots for CHD risk across the deciles of risk score. Observed versus predicted CHD risk in the training dataset (a) and the testing dataset (b). Data points represent means and error bars represent 95% CI.

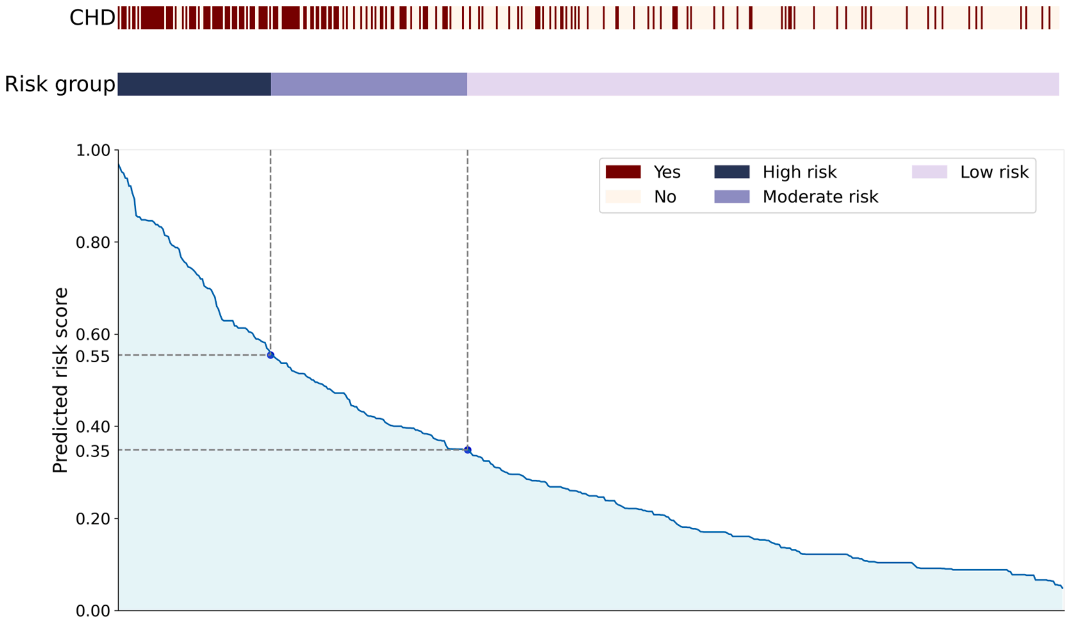


Fig. 5. Distribution of predicted risk scores and risk groups among participants. Risk score, risk group, and observed CHD in the testing dataset.

potentially expose individuals to CHD-related substances absent in urban settings, e.g. chemicals employed in agricultural and gardening practices.[29] Additionally, rural isolation and limited access to health services may be potential reasons for CHD among rural residents. [30] Furthermore, the low wealth index, indicative of impoverished household economic conditions, is also an important feature of the prediction model. This aligns with the results obtained from the CHD prediction model derived from the China Birth Cohort, which identified higher annual household income as a protective factor against CHD in offspring. [17] Household economic status is directly related to material conditions, such as living conditions, health care, and lifestyle.[31,32] The third most important feature of the predictive model is folic acid supplements. Studies have affirmed adequate folic acid intake plays a crucial role in the primary prevention of CHD.[33–35] Parity, negative emotions, passive smoking, chemicals and toxins, and cold are also important predictors in the predictive model consistent with previous studies. [18,36–39]

Risk scores were derived from the XGB model and divided into deciles. Subsequently, we compared the discrepancies between the observed proportion of CHD in each decile and the predicted risk score. The observed CHD proportion exhibited minor underestimations of less than 0.04 across the majority of deciles in the testing dataset. However, a more substantial underestimation of 0.17 was observed in the ninth decile, contrasted by an overestimation of 0.12 in the tenth decile. From a clinical perspective, these differences may be considered negligible, given that most pregnant women will continue to receive recommended CHD risk management. Moreover, deviations in risk prediction for high-risk populations are unlikely to lead to missed or unnecessary medical interventions since all pregnant women at high risk for fetal CHD should be advised to receive the same preventive management. Nevertheless, when we stratified by maternal age, the calibration appeared less robust for pregnant women aged 35 years or older, possibly attributable to the smaller sample size.

While the increase in event risk is progressive throughout the risk scores, we recommend categorizing pregnant women into three risk tiers—namely, low, moderate, and high. This stratification is intended to underscore the clinical significance of each risk value generated by the model. Based on such stratification, pregnant women identified as high risk for CHD in their offspring during the periconceptional period should receive more comprehensive monitoring. While our study did not specifically address the issue of prenatal screening and counseling for CHD, we posit that this may constitute one of the most substantial implications of the risk score. The classification of pregnant women into three risk categories could potentially rationalize the planning of detailed fetal ultrasound examinations and avoid one-size-fits-all.

Nonetheless, this study has several limitations. Firstly, although we selected strong predictors of CHD using the LASSO method, relevant biochemical indicators and genetic factors could not be incorporated because they were not available in the epidemiologic data used for model development. Secondly, the data were collected through a case-control study, which is a retrospective observational study prone to recall bias. However, it's essential to note that the survey questions were highly specific and unambiguous, the questionnaire underwent meticulous pretesting, and the data were diligently sourced from multiple outlets, thereby enhancing statistical power and mitigating bias. Thirdly, although external validation was performed using data from a case-control study, additional external validation is imperative to comprehensively assess the model's generalizability across diverse populations.

5. Conclusions

This study developed a ML-based risk-scoring tool for predicting the CHD risk in fetuses. It demonstrated that a ML-based approach is feasible and effective in this domain, offering significant insights for guiding preventive primary care.

6. Summary table

What was already known about the topic?

- Several prediction models for congenital heart disease (CHD) have been developed, with a predominant reliance on traditional statistical methods, notably logistic regression, which limits their predictive capacities.
- Currently, there is a gap in the development of interpretable and user-friendly CHD risk prediction tools based on machine learning.

What has this study added to our knowledge?

- Among the five machine learning prediction models, eXtreme Gradient Boosting (XGB) emerged as the top-performing model for CHD prediction.
- We developed a risk score, a tool based on XGB, designed to forecast the risk of CHD in fetuses of pregnant women during early pregnancy.
- To underscore the clinical significance of the model-estimated risks, we categorized the risks into three classes (low, moderate, and high).

Contributors

LP and SZ designed the study and drafted the manuscript. SZ, CK, JC, HX, SZ, YC and HL contributed to data acquisition, analysis, and interpretation. SZ, LP, PQ, LY, YZ, FC and DW participated in the critical revision of the manuscript for important intellectual content. All authors contributed to the critical revisions and approved the final version of the manuscript.

Ethics approval

The study was approved by the ethics committee of Xi'an Jiaotong University Health Science Center (No. 2012008).

CRediT authorship contribution statement

Shutong Zhang: Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Chenxi Kang:** Validation, Software, Data curation. **Jing Cui:** Validation, Software, Data curation. **Haodan Xue:** Validation, Methodology, Data curation. **Shanshan Zhao:** Validation, Software, Data curation. **Yukui Chen:** Validation, Software, Data curation. **Haixia Lu:** Validation, Methodology, Data curation. **Lu Ye:** Writing – review & editing, Supervision, Funding acquisition. **Duolao Wang:** Writing – review & editing, Supervision. **Fangyao Chen:** Writing – review & editing, Supervision. **Yaling Zhao:** Writing – review & editing, Supervision. **Leilei Pei:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Pengfei Qu:** Writing – review & editing, Supervision, Methodology, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank all medical staff involved in the study for recruiting the participants. The authors also thank all mothers and infants who participated in the study and all investigators who contributed to data collection.

Funding

All the work was supported by the National Natural Science Foundation of China (grant numbers 72174167, 82103924); Natural Science Foundation of Shaanxi (grant number 2021JM-031); Key Research and Development Projects of Shaanxi Province (grant number 2022ZDLSF03-10, 2024SF-YBXM-238); National Key Research and Development Program of China (grant numbers 2017YFC0907200, 2017YFC0907201); Xi'an Science and Technology Plan Project (grant number 22YXYJ0095).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2024.105741.

References

[1] A.C. Fahed, B.D. Gelb, J.G. Seidman, C.E. Seidman, Genetics of congenital heart disease: the glass half empty, Circ. Res. 112 (2013) 707–720.

[2] J.I. Hoffman, S. Kaplan, The incidence of congenital heart disease, J. Am. Coll. Cardiol. 39 (2002) 1890–1900.

[3] G.A. Roth, G.A. Mensah, C.O. Johnson, G. Addolorato, E. Ammirati, L.M. Baddour, N.C. Barengo, A.Z. Beaton, E.J. Benjamin, C.P. Benziger, A. Bonny, M. Brauer, M. Brodmann, T.J. Cahill, J. Carapetis, A.L. Catapano, S.S. Chugh, L.T. Cooper, J. Coresh, M. Criqui, N. DeCleene, K.A. Eagle, S. Emmons-Bell, V.L. Feigin, J. Fernández-Solà, G. Fowkes, E. Gakidou, S.M. Grundy, F.J. He, G. Howard, F. Hu, L. Inker, G. Karthikeyan, N. Kassebaum, W. Koroshetz, C. Lavie, D. Lloyd-Jones, H. S. Lu, A. Mirijello, A.M. Temesgen, A. Mokdad, A.E. Moran, P. Muntner, J. Narula, B. Neal, M. Ntsekhe, G. Moraes de Oliveira, C. Otto, M. Owolabi, M. Pratt, S. Rajagopalan, M. Reitsma, A.L.P. Ribeiro, N. Rigotti, A. Rodgers, C. Sable, S. Shakil, K. Sliwa-Hahnle, B. Stark, J. Sundström, P. Timpel, I.M. Tleyjeh, M. Valgimigli, T. Vos, P.K. Whelton, M. Yacoub, L. Zuhlke, C. Murray, V. Fuster, Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study, J. Am. Coll. Cardiol. 76 (2020) 2982–3021.

[4] Q.M. Zhao, X.J. Ma, X.L. Ge, F. Liu, W.L. Yan, L. Wu, M. Ye, X.C. Liang, J. Zhang, Y. Gao, B. Jia, G.Y. Huang, Pulse oximetry with clinical assessment to screen for congenital heart disease in neonates in China: a prospective study, Lancet 384 (2014) 747–754.

[5] Q. He, Z. Dou, Z. Su, H. Shen, T.N. Mok, C.J.P. Zhang, J. Huang, W.K. Ming, S. Li, Inpatient costs of congenital heart surgery in China: results from the National Centre for Cardiovascular Diseases, Lancet Reg. Health West Pac. 31 (2023) 100623.

[6] R. Boyd, H. McMullen, H. Beqaj, D. Kalfa, Environmental Exposures and Congenital Heart Disease, Pediatrics 149 (2022).

[7] Q. Liang, W. Gong, D. Zheng, R. Zhong, Y. Wen, X. Wang, The influence of maternal exposure history to virus and medicine during pregnancy on congenital heart defects of fetus, Environ. Sci. Pollut. Res. Int. 24 (2017) 5628–5632.

[8] S. Liu, K.S. Joseph, S. Lisonkova, J. Rouleau, M. Van den Hof, R. Sauve, M. S. Kramer, Association between maternal chronic conditions and congenital heart defects: a population-based cohort study, Circulation 128 (2013) 583–589.

[9] N. Schwalbe, B. Wahl, Artificial intelligence and the future of global health, Lancet 395 (2020) 1579–1586.

[10] F. D'Ascenzo, O. De Filippo, G. Gallone, G. Mittone, M.A. Deriu, M. Iannaccone, A. Ariza-Sole, C. Liebetrau, S. Manzano-Fernandez, G. Quadri, T. Kinnaird, G. Campo, J.P. Simao Henriques, J.M. Hughes, A. Dominguez-Rodriguez, M. Aldinucci, U. Morbiducci, G. Patti, S. Raposeiras-Roubin, E. Abu-Assi, G.M. De Ferrari, P.s. group, Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets, Lancet 397 (2021) 199–207.

[11] R. Arnaout, L. Curran, E. Chinn, Y. Zhao, A. Moon-Grady, Deep-learning models improve on community-level diagnosis for common congenital heart disease lesions, arXiv preprint arXiv:1809.06993, (2018).

[12] H. Chen, L. Wu, Q. Dou, J. Qin, S. Li, J.Z. Cheng, D. Ni, P.A. Heng, Ultrasound Standard Plane Detection Using a Composite Neural Network Framework, IEEE Trans. Cybern. 47 (2017) 1576–1586.

[13] R. Arnaout, L. Curran, Y. Zhao, J.C. Levine, E. Chinn, A.J. Moon-Grady, An ensemble of neural networks provides expert-level prenatal detection of complex congenital heart disease, Nat. Med. 27 (2021) 882–891.

[14] Y. Luo, Z. Li, H. Guo, H. Cao, C. Song, X. Guo, Y. Zhang, Predicting congenital heart defects: A comparison of three data mining methods, PLoS One 12 (2017) e0177811.

[15] H. Li, M. Luo, J. Zheng, J. Luo, R. Zeng, N. Feng, Q. Du, J. Fang, An artificial neural network prediction model of congenital heart disease based on risk factors: A hospital-based case-control study, Medicine (Baltimore) 96 (2017) e6090.

[16] I. Kaur, T. Ahmad, A cluster-based ensemble approach for congenital heart disease prediction, Comput. Methods Programs Biomed. 243 (2024) 107922.

[17] M. Zhang, Y. Sun, X. Zhao, R. Liu, B.Y. Yang, G. Chen, W. Zhang, G.H. Dong, C. Yin, W. Yue, How Parental Predictors Jointly Affect the Risk of Offspring Congenital Heart Disease: A Nationwide Multicenter Study Based on the China Birth Cohort, Front. Cardiovasc. Med. 9 (2022) 860600.

[18] Y. Liang, X. Li, X. Hu, B. Wen, L. Wang, C. Wang, A predictive model of offspring congenital heart disease based on maternal risk factors during pregnancy: a hospital based case-control study in Nanchong City, Int. J. Med. Sci. 17 (2020) 3091–3097.

[19] R. Sun, M. Liu, L. Lu, Y. Zheng, P. Zhang, Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments, Cell Biochem. Biophys. 72 (2015) 857–860.

[20] T. van der Bom, A.C. Zomer, A.H. Zwinderman, F.J. Meijboom, B.J. Bouma, B. J. Mulder, The changing epidemiology of congenital heart disease, Nat. Rev. Cardiol. 8 (2011) 50–60.

[21] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. van Smeden, A.-L. Boulesteix, J. C. Camaradou, L.A. Celi, S. Denaxas, A.K. Denniston, B. Glocker, R.M. Golub, H. Harvey, G. Heinze, M.M. Hoffman, A.P. Kengne, E. Lam, N. Lee, E.W. Loder, L. Maier-Hein, B.A. Mateen, M.D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, P. Logullo, TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378.

[22] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.

[23] K. Pearson, On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling, in: S. Kotz, N.L. Johnson (Eds.),

Breakthroughs in Statistics: Methodology and Distribution, Springer New York, New York, NY, 1992, pp. 11–28.

[24] D. Kalpić, N. Hlupić, M. Lovrić, Student's t-Tests, in: M. Lovric (Ed.), International Encyclopedia of Statistical Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 1559–1563.

[25] T.P. Hettmansperger, J.W. McKean,Robust nonparametric statistical methods, CRC press2010.

[26] Streamlit • a Faster Way to Build and Share Data Apps.

[27] X. Zhang, S. Li, S. Wu, X. Hao, S. Guo, K. Suzuki, H. Yokomichi, Z. Yamagata, Prevalence of birth defects and risk-factor analysis from a population-based survey in Inner Mongolia, China, BMC Pediatr. 12 (2012) 125.

[28] Q. Miao, S. Dunn, S.W. Wen, J. Lougheed, F. Sharif, M. Walker, Associations of congenital heart disease with deprivation index by rural-urban maternal residence: a population-based retrospective cohort study in Ontario, Canada, BMC Pediatr 22 (2022) 476.

[29] G.A. Heeren, J. Tyler, A. Mandeya, Agricultural chemical exposures and birth defects in the Eastern Cape Province, South Africa: a Case-Control Study, Environ Health 2 (2003) 11.

[30] Z.C. Luo, R. Wilkins, Degree of rural isolation and birth outcomes, Paediatr. Perinat. Epidemiol. 22 (2008) 341–349.

[31] Q. Miao, S. Dunn, S.W. Wen, J. Lougheed, J. Reszel, C. Lavin Venegas, M. Walker, Neighbourhood maternal socioeconomic status indicators and risk of congenital heart disease, BMC Pregnancy Childbirth 21 (2021) 72.

[32] R.E. Portilla, V. Harizanov, K. Sarmiento, J. Holguín, G. Gracia, P. Hurtado-Villa, I. Zarante, Risk factors characterisation for CHD: a case-control study in Bogota and Cali, Colombia, 2002-2020, Cardiol. Young (2023) 1–5.

[33] H. Chen, Y. Zhang, D. Wang, X. Chen, M. Li, X. Huang, Y. Jiang, Y. Dou, Y. Wang, X. Ma, W. Sheng, B. Jia, W. Yan, G. Huang, Periconception Red Blood Cell Folate and Offspring Congenital Heart Disease : Nested Case-Control and Mendelian Randomization Studies, Ann. Intern. Med. 175 (2022) 1212–1220.

[34] A. Xu, X. Cao, Y. Lu, H. Li, Q. Zhu, X. Chen, H. Jiang, X. Li, A Meta-Analysis of the Relationship Between Maternal Folic Acid Supplementation and the Risk of Congenital Heart Defects, Int. Heart J. 57 (2016) 725–728.

[35] S. Liu, K.S. Joseph, W. Luo, J.A. León, S. Lisonkova, M. Van den Hof, J. Evans, K. Lim, J. Little, R. Sauve, M.S. Kramer, Effect of Folic Acid Food Fortification in Canada on Congenital Heart Disease Subtypes, Circulation 134 (2016) 647–655.

[36] L. Zhao, L. Chen, T. Yang, L. Wang, T. Wang, S. Zhang, L. Chen, Z. Ye, Z. Zheng, J. Qin, Parental smoking and the risk of congenital heart defects in offspring: An updated meta-analysis of observational studies, Eur. J. Prev. Cardiol. 27 (2020) 1284–1293.

[37] R. Adams, A. Porras, G. Alonso, Essential role of p38alpha MAP kinase in placental but not embryonic cardiovascular development, Mol. Cell 102 (2000) 131–134.

[38] T. Lai, L. Xiang, Z. Liu, Y. Mu, X. Li, N. Li, S. Li, X. Chen, J. Yang, J. Tao, J. Zhu, Association of maternal disease and medication use with the risk of congenital heart defects in offspring: a case-control study using logistic regression with a random-effects model, J. Perinat. Med. 47 (2019) 455–463.

[39] L. Pei, Y. Kang, Y. Zhao, H. Yan, Prevalence and risk factors of congenital heart defects among live births: a population-based cross-sectional survey in Shaanxi province, Northwestern China, BMC Pediatr 17 (2017) 18.