Original Article

Identifying the Factors Affecting the Incidence of Congenital Heart Disease Using Support Vector Machine and Particle Swarm Optimization

Bahar Dehghan[1], Mohammad Reza Sabri[1], Alireza Ahmadi[1], Mehdi Ghaderian[1], Chehreh Mahdavi[1], Davood Ramezani Nejad[1], Mohammad Sattari[2]

[1]Pediatric Cardiovascular Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran, [2]Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

Abstract

Background: Congenital malformations are defined as "any defect in the structure of a person that exists from birth". Among them, congenital heart malformations have the highest prevalence in the world. This study focuses on the development of a predictive model for congenital heart disease in Isfahan using support vector machine (SVM) and particle swarm intelligence.

Materials and Methods: It consists of four parts: data collection, preprocessing, identify target features, and technique. The proposed technique is a combination of the SVM method and particle swarm optimization (PSO).

Results: The data set includes 1389 patients and 399 features. The best performance in terms of accuracy, with 81.57%, is related to the PSO-SVM technique and the worst performance, with 78.62%, is related to the random forest technique. Congenital extra cardiac anomalies are considered as the most important factor with averages of 0.655.

Conclusion: Congenital extra cardiac anomalies are considered as the most important factor. Detecting more important feature affecting congenital heart disease allows physicians to treat the variable risk factors associated with congenital heart disease progression. The use of a machine learning approach provides the ability to predict the presence of congenital heart disease with high accuracy and sensitivity.

Keywords: Congenital heart disease, incidence, particle swarm optimization, risk factors, support vector machine

Address for correspondence: Dr. Mohammad Sattari, Department of Health Information Technology and Management, School of Medical Management and Information Sciences, Isfahan University of Medical Sciences, Hezarjerib Avenue, P.O. Box: 81745-346, Isfahan, Iran.
E-mail: msattarimng.mui@gmail.com
Submitted: 16-Feb-2022; **Revised:** 09-Oct-2022; **Accepted:** 12-Oct-2022; **Published:** 19-May-2023

Introduction

Congenital malformations are defined as "any defect in the structure of a person that exists from birth". Among them, congenital heart disease (CHD) has the highest prevalence in the world.[1] CHD is the most common type of birth defect, which accounts for about 25% of all congenital anomalies. It is characterized by a wide range of clinical symptoms.[2] According to studies, more than 8 cases of every 1000 births has CHD.[3] In severe cases, the disease can lead to child mortality.[4] These diseases can lead to developmental disability in adulthood as well.[5] CHD can cause many problems and impose high costs on households.[6] Various factors can be effective in the incidence of this disease. Also, several genetic factors have been considered as effective factors.[7] Yet, the causes of these congenital diseases are not clearly known.[8,9]

Many factors have made this disease the most important congenital anomaly. These factors are the burden of not recognizing and treating, the economic, psychological costs of society, and health system and families' problems. The risk

Access this article online

Quick Response Code:

Website: www.advbiores.net

DOI: 10.4103/abr.abr_54_22

How to cite this article: Dehghan B, Sabri MR, Ahmadi A, Ghaderian M, Mahdavi C, Ramezani Nejad D, *et al.* Identifying the factors affecting the incidence of congenital heart disease using support vector machine and particle swarm optimization. Adv Biomed Res 2023;12:130.

factors for congenital diseases such as CHD are different in various regions of Iran and the world. Identifying these factors and their impact on the incidence of CHD can play an important role in preventing these diseases.[10]

Data mining techniques have been used as one of the effective methods in diagnosing and identifying medical patterns in recent years. Data mining is a set of techniques that receives a set of data as input and produces a set of patterns (knowledge) as output. By using these techniques, it is possible to identify the effect of the combination of different factors on the occurrence of CHD at a lower cost. Luo *et al.*[11] used three data mining techniques: weighted support vector machine (WSVM), weighted random forest (WRF), and logistic regression to predict CHD defects. It was suggested that these three techniques be used as a tool to identify high-risk groups. Rani *et al.*[12] used simple Bayesian and neural network techniques to predict CHDs. The results showed the better performance of the neural network technique and suggested the use of this technique. Hoodbhoy *et al.*[13] evaluated the accuracy of machine learning techniques in the diagnosis of CHD. The investigation showed the appropriate performance of the neural network technique. Chu investigated the complications of CHD for pregnant women. He used support vector machine (SVM), random forest, AdaBoost, decision tree, k-nearest neighbor, simple Bayesian, and multilayer perceptron techniques to create the model and stated that women with previous congenital disease should be consulted about pregnancy during puberty.[14] By examining this point, in previous studies, various topics in CHD have been discussed in different collections, and it is also necessary to mention that the Isfahan CHD dataset has not been investigated for the discussion of machine learning. So, this study focuses on the development of a predictive model for CHD in Isfahan using SVM and particle swarm intelligence.

Materials and Methods

The protocol of study approved by Ethics committee of Isfahan University of Medical Sciences (Ethical Code: IR.MUI.NUREMA.REC.1400.222). It consists of four parts: data collection, preprocessing, identify target features, and technique.

Data collection

The datasets play the most significant role in data analysis. In this paper, information about all patients in Isfahan Pediatric Cardiovascular Research Center will be considered. The data set includes 1389 patients and 399 features (a noticeable or important characteristic or part). Some of these are positive family history of CHD, maternal illness (diabetes), maternal illness (hypothyroidism), hypertension (mother), thyroid medications, insulin, metformin, aspirin, maternal addiction, and genetic abnormalities.

Preprocessing

The first part of preprocessing is data conversion. The nationality attribute was removed from the set of values. Considering that all these patients were Iranian. The days and months of birth were removed from the patients' data set. The second part of preprocessing is handling missing attributes. Among the available features, features that have more than 75% of the rows without value are removed from the set of features. The third part is data split. Also, the ages of patients' fathers and mothers are obtained based on the date of referral and their date of birth. For the feature of parental occupation, two features were considered, so that a person can have two jobs. For better processing, adjectives father jobs and mother jobs were used. The fourth part is identifying the related attributes by experts. After this part, a total of 21 attributes are remained. These 21 attributes included congenital extra cardiac anomalies, Positive family history of CHD, maternal illness (diabetes), maternal illness (hypothyroidism), hypertension (mother), thyroid medications, insulin, metformin, aspirin, maternal addiction, genetic abnormalities, sex, parent's consanguinity, birth conditions, referral time, patient's age at the time of registry (year), gestational age, mother's and father's ages, father' and mother's jobs.

Identify target features

Target feature values include ventricular septal defect, atrial septal defect, patent ductus arteriosus, double outlet right ventricle, congenital insufficiency of aortic valves, congenital pulmonary valve stenosis, and coarctation of aorta, atrioventricular septal defect, and tetralogy of fallot.

Techniques

The proposed technique is a combination of SVM method and particle swarm optimization (PSO). SVM can be utilized for classification problems. It is a supervised machine learning algorithm. The SVM algorithm considers any data item as a data point in n-dimensional space (n = the number of features). By receiving labeled training data (supervised training), the SVM algorithm presents a hyperplane. It uses this hyperplane to classify new instances. The SVM technique works in such a way that it is used for binary target features. Multiclasses become binary problems. The backup vector machine has a problem. The problem is that the parameters of this attribute need to be determined by the user and the results depend on these parameters. To deal with this problem, the evolutionary method of PSO has been proposed. This method is a collective search algorithm modeled on the social behavior of flocks of birds. PSO algorithms use particles moving in a later space to look for solutions to the variable function optimization problem. The performance of this technique is compared with the SVM technique and the random forest technique.

Results

Evaluation criteria

This study used test data sets to test methods. First, the complexity matrix is calculated, which includes various criteria such as TP, TN, FP, and FN. The TP identifies the number of records that the group has correctly placed as low. TN identifies records that have been correctly identified as Medium or high.

FN identifies records that have been incorrectly identified as low, and FP identifies the number of records that have been incorrectly identified as medium or high.

The target group consists of seven classes with three values: low, normal, and high. In fact, the techniques will be used separately for each class. Different criteria are used for evaluation. One of these criteria is accuracy[15], that the closer the accuracy of this criterion, the better the result. This criterion is calculated based on the following formula:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

The other criteria are sensitivity, specificity[16], and confidence.[17] The confidence is between 0 and 1. The closer to one the confidence is, the better the performance of the technique.

Results

Thirty-three adjectives are considered as predictors, some of which are listed in Table 1. According to Table 1, 156 people had a preterm birth. Among which 504 are men and 513 are women. Sixteen parents were cousins. Also, 382 of their fathers' jobs are workers and 54 are drivers. Five hundred and eighty-eight people are less than 40 years old.

According to Table 2, all three techniques performed well and were able to achieve at least 78% accuracy. The best performance in terms of accuracy, with 81.57%, is related to the particle swarm optimization support vector machine (PSO-SVM) technique and the worst performance, with 78.62%, is related to the random forest technique.

The principle component analysis technique extracted 12 factors affecting the incidence of heart disease. Then, the importance of these factors in different techniques was extracted. The output is shown in Table 3. In all three techniques, congenital extra cardiac anomalies are considered as the most important factor. This factor averages 0.655. The second factor in terms of the mean value is chromosomal/genetic abnormalities, which has been more than 50% significant in the SVM technique. The third factor based on the average value is weight, which is more than 50% of the importance in both random forest and PSO-SVM techniques.

Table 1: Attributes, quantities, and numbers related to the congenital heart patient database

Attributes	Values (number)
Congenital extra cardiac anomalies	Yes (186) No (830)
Positive Family history of CHD	Yes (116) No (900)
Maternal Illness (diabetes)	Yes (34) No (982)
Maternal Illness (hypothyroidism)	Yes (81) No (935)
Hypertension (mother)	Yes (8) No (1008)
Thyroid medications	Yes (78) No (938)
Insulin	Yes (16) No (1000)
Metformin	Yes (18) No (998)
Aspirin	Yes (18) No (998)
Maternal addiction	Yes (28) No (988)
Chromosomal/genetic abnormalities	Yes (186) No (830)
Sex	Male (504) Female (513)
Parent's consanguinity	Cousin (16)
	Fourth-degree communication (96)
	Third-party communication (93)
	Fifth-degree communication (247)
Birth condition	Term (850)
	Early (156)
Referral time	2018 (501)
	2017 (492)
Patient's age at the time of registry (year)	1-10 (766)
	11-20 (303)
	21-30 (41)
	31-40 (3)
Mother's age (years)	Below than 20 (11)
	20-30 (167)
	30-40 (598)
	40-50 (7)
Father's age (years)	Below than 20 (7)
	20-30 (36)
	30-40 (470)
	40-50 (305)
Gestational age (weeks)	Less than 28 (7)
	28-32 (24)
	33-36 (86)
	37-40 (70)
Father's job	Workers (382)
	Drivers (54)
	Unemployed (23)
	Retired (15)
	Employees (157)
	Engineer (17(
Mother's job	Housewives (902)
	Working outdoors (115)

Discussion

To the best of our knowledge, this is the first study that has applied data mining to identify some factors involved in the occurrence of CHDs in Iran. Our study showed that among 12 factors affecting the disease, in CHD patients, any association with other congenital extracardiac anomalies is the most important factor affecting the incidence of CHD. Other important factors are the presence of genetic disorders and birth weight.

Several studies have evaluated the influence of factors on the incidence of CHD; yet, the results are not the same. Liu *et al.*[18] have reported that the mother's education level, neonatal asphyxia, number of previous pregnancies, maternal infections, and her mental stress during early pregnancy are the environmental risk factors associated with CHD. In another study, Jin *et al.*[19] found out living suburban, and having a previous child were negatively associated with CHD. Yet, twin pregnancies, maternal illness in the first trimester of pregnancy, a family history of CHD, and having another child with a congenital anomaly were positively associated with

CHD. In another study, Bassili *et al.*,[20] have been reported that high age in both parents, consanguinity, positive family history of CHD, female babies, irradiation, some maternal occupations, diabetes, and suburban or rural residence before and during the pregnancy are the most effective risk factors for any type of CHD. Vasanthanageswari assesses the risk factor of congenital heart defect using the association rule mining technique. Many number of patients affected by rheumatoid, and symptom affected the minimum number of patients is body temperature.[21]

In our study, we could include multiple known affecting factors and evaluate them with a data mining method. Our findings were more similar to those obtained by Hassan *et al.*[22] who have reported that chromosomal abnormalities, especially Trisomy 21, are the most common association with CHDs; nonetheless, history of abortions or stillbirths, parental consanguinity, and maternal diabetes did not have any significant influence on their survey.

Due to the preventability of many factors affecting CHDs and to decrease the infantile mortality rate, many studies have been established to find more risk factors that may prevent CHD during pregnancy. The investigations are still ongoing in this regard. Abqari *et al.*[23] have reported a positive association between the occurrence of CHD and high parental age, history of a bad pregnancy, infections during pregnancy, and folic acid deficiency.

Nowadays, data mining techniques are currently used widely in preventive medicine to identify important risk factors. Luo *et al.*[11] surveyed to improve and validate machine learning models to predict the risks of bearing a child with CHD for mothers. Three models have been used: WSVM, WRF, and logistic regression (Logit) for nine maternal variables. Authors could confirm the predictive ability of these models. We used SVM, PSO-SVM, and random forest techniques, and as mentioned, our results were significant too.

A lot of information has been obtained with machine learning applications to optimize tools for the timely diagnosis of CHDs.[24] Rani *et al.*[12] used neural networks and Bayesian techniques to predict CHD. The results showed better performance of the neural network technique. He suggested the use of this technique for predicting this disease. Hoodbhoy *et al.*[13] evaluated the accuracy of machine learning techniques in diagnosing CHD. The study showed the proper performance of the neural network technique. Chu studied the complications of CHD for pregnant women. He used SVM, random forest, AdaBoost, decision tree, k-nearest neighbor, Naive-Bayes, and multilayer perception to create the model. It stated that women with a previous congenital disease should be consulted about pregnancy during puberty.[14]

In our study, SVM + PSO was used to identify factors affecting CHD. The SVM gets high performance in many pattern recognition techniques.[25] This technique has a remarkable ability to generalize unseen test data. Moreover, PSO has a good ability to solve the optimization problem.

The limitations of this study are that due to a large amount of missing data in some features (for example, information about the area where mother has lived before and during pregnancy which has shown a positive correlation with the occurrence of CHD), these features were removed. Another limitation is that the dataset has many features.

Table 2: Accuracy of different techniques (SVM-PSO-SVM-random forest)

Techniques	Accuracy
SVM	79.61%
PSO-SVM	81.57%
Random Forest	78.62%

Conclusion

In this study, it was shown that machine learning algorithms can be used with high accuracy to detect factors that may influence on the occurrence of CHD. It was also shown that SVM+PSO detects more accurately than other techniques.

Table 3: Factors affecting the incidence of congenital heart disease with the importance

Factors	Significance by Random Forest (Gini index)	Significance by PSO-SVM (Gain information)	Significance by support vector machine	Mean
Congenital extra cardiac anomalies	0.718	0.590	0.657	0.655
Chromosomal/genetic abnormalities	0.453	0.399	0.575	0.475
Hypothyroidism	0.124	0.140	0.248	0.170
thyroid drugs	0.08	0.141	0.219	0.146
Father age	0.221	0.202	0.380	0.267
Region	0.254	0.155	0.544	0.317
Positive Family history of CHD	0.112	0.178	0.333	0.207
Gender	0.160	0.189	0.220	0.189
Father's job	0.182	0.186	0.317	0.228
Mother's Job	0.363	0.141	0.479	0.327
Height	0.457	0.383	0.134	0.324
Weight	0.558	0.526	0.250	0.444

Detecting more important features of CHD allows physicians to treat the variable risk factors associated with CHD progression. The use of a machine learning approach provides the ability to predict the presence of CHD with high accuracy. Thus, it allows physicians to perform timely preventive treatment in patients with CHD.

Acknowledgements

We thank the support and cooperation of the Health Information Technology Research Center and the Pediatric Cardiovascular Research Center.

Financial support and sponsorship

This study is supported by Isfahan University of Medical Sciences.

Conflicts of interest

There are no conflicts of interest.

References

1. World Health Organization. Community genetics services: report of a WHO consultation on community genetics in low and middle-income countries. Available from: http://www.who.int/iris/handle/10665/44532. [Last accessed on 2011 Sep].
2. Correa-Villaseñor A, Cragan J, Kucik J, O'Leary L, Siffel C, Williams L. The metropolitan Atlanta congenital defects program: 35 Years of birth defects surveillance at the centers for disease control and prevention. Birth Defects Res A Clin Mol Teratol 2003;67:617-24.
3. Bernier PL, Stefanescu A, Samoukovic G, Tchervenkov CI. The challenge of congenital heart disease worldwide: Epidemiologic and demographic facts. Semin Thorac Cardiovasc Surg Pediatr Card Surg Annu 2010;13:26-34.
4. Abd MA, Sarhat AR. Congenital heart diseases in consanguineous marriages in Tikrit-Iraq. Indian J Public Health Res Dev 2019;1070-1072.
5. Jordan LC, Siciliano RE, Cole DA, Lee CA, Patel NJ, Murphy LK, *et al.* Cognitive training in children with hypoplastic left heart syndrome: A pilot randomized trial. Prog Pediatr Cardiol 2020;57:101185.
6. Oyen N, Poulsen G, Boyd HA, Wohlfahrt J, Jensen PK, Melbye M. National time trends in congenital heart defects, 1977-2005. Am Heart J 2009;157:467-73.e1.
7. Yang Q, Chen H, Correa A, Devine O, Mathews TJ, Honein MA. Racial differences in infant mortality attributable to birth defects in the United States, 1989-2002. Birth Defects Res A Clin Mol Teratol 2006;76:706-13.
8. Pierpont ME, Basson CT, Benson DW Jr, Gelb BD, Giglia TM, Goldmuntz E, *et al.* American Heart Association Congenital Cardiac Defects Committee, Council on Cardiovascular Disease in the Young. Genetic basis for congenital heart defects: Current knowledge: A scientific statement from the American heart association congenital cardiac defects committee, council on cardiovascular disease in the young: Endorsed by the American Academy of Pediatrics. Circulation 2007;115:3015-38.
9. Taheri M, Dehghani A, Lotfi MH, Noori shadkam M, Fallahzadeh H. Study of maternal risk factors associated with the incidence of congenital heart disease: A case-control study. J Pediatr Nurs 2015;2:70-8.
10. Dehghan B, Sabri MR, Hosseinzadeh M, Ahmadi AR, Ghaderian M, Sarrafzadegan N, *et al.* The commencement of congenital heart diseases registry in Isfahan, Iran: Methodology and design. ARYA Atheroscler 2020;16:244-7.
11. Luo Y, Li Z, Guo H, Cao H, Song C, Guo X, *et al.* Predicting congenital heart defects: A comparison of three data mining methods. PLoS One 2017;12:e0177811.
12. Rani S, Masood S. Predicting congenital heart disease using machine learning techniques. J Discret Math Sci Cryptogr 2020;23:293-303.
13. Hoodbhoy Z, Jiwani U, Sattar S, Salam R, Hasan B, Das JK. Diagnostic accuracy of machine learning models to identify congenital heart disease: A meta-analysis. Front Artif Intell 2021;4:708365.
14. Chu R, Chen W, Song G, Yao S, Xie L, Song L, *et al.* Predicting the risk of adverse events in pregnant women with congenital heart disease. J Am Heart Assoc 2020;9:e016371.
15. Mastrogiannis N, Boutsinas B, Giannikos I. A method for improving the accuracy of data mining classification algorithms. Comput Oper Res 2009;36:2829-39.
16. Cortez P, Embrechts MJ. Using sensitivity analysis and visualization techniques to open black box data mining models. Inf Sci 2013;225:1-17.
17. Zaki MJ, Parthasarathy S, Li W, Ogihara M. Evaluation of sampling for data mining of association rules. In Proceedings Seventh International Workshop on Research Issues in Data Engineering. High Performance Database Management for Large-Scale Applications. IEEE; 1997.p. 42-50.
18. Liu S, Liu J, Tang J, Ji J, Chen J, Liu C. Environmental risk factors for congenital heart disease in the Shandong Peninsula, China: A hospital-based case–control study. J Epidemiol 2009;19:122-30.
19. Jin X, Ni W, Wang G, Wu Q, Zhang J, Li G, *et al.* Incidence and risk factors of congenital heart disease in Qingdao: A prospective cohort study. BMC Public Health 2021; 21:1044.
20. Bassili A, Mokhtar SA, Dabous NI, Zaher SR, Mokhtar MM, Zaki A. Risk factors for congenital heart diseases in Alexandria, Egypt. Eur J Epidemiol 2000;16:805-14.
21. Vasanthanageswari S, Vanitha M. Predicting risk factor of congenital heart defect using association rule mining technique. Int J Pure Appl Math 2018;118:399-404.
22. Hassan I, Haleem AA, Bhutta ZA. Profile and risk factors for congenital heart disease. J Pak Med Assoc 1997;47:78-81.
23. Abqari S, Gupta A, Shahab T, Rabbani MU, Manazir A, Firdaus U. Profile and risk factors for congenital heart defects: A study in a tertiary care hospital. Ann Pediatr Cardiol 2016;9:216-21.
24. Helman SM, Herrup EH, Christopher AB, Al-Zaiti SS. The role of machine learning applications in diagnosing and assessing critical and non-critical CHD: A scoping review. Cardiol Young 2021;31:1770-80.
25. Byun H, Lee S-W. A survey on pattern recognition applications of support vector machines. Int J Pattern Recognit Artif Intell 2003;17:459-86.