Prediction of Congenital Cardiovascular Disease Using Machine Learning Techniques: A Review Analysis

Sudipto Ghosh[1], Tanishq Awasthy[2]
Department of CSE, Chandigarh University, Mohali, Punjab, India
19bcs1074@cuchd.in, 19bcs1091@cuchd.in

Abstract— Congenital cardiovascular disease (CVD) is a significant health concern affecting individuals from birth and often necessitating long-term medical management. Early prediction and diagnosis of CVD play a crucial role in improving patient outcomes and guiding appropriate interventions. In recent years, machine learning (ML) techniques have emerged as promising tools for CVD prediction, leveraging their ability to analyze complex patterns within large datasets. This review analysis explores the landscape of ML techniques employed to predict congenital CVD. Machine learning techniques have shown significant potential for cardiac disease prediction, especially when using large and complex datasets. This review paper comprehensively overviews several machine-learning methods for heart disease prediction. This research outlines the advantages and disadvantages of several machine learning techniques. It thoroughly analyzes their performance in predicting heart disease, conducting a comprehensive survey of recent literature encompassing diverse ML algorithms such as decision trees, support vector machines, random forests, neural networks, and deep learning architectures. Examining the various data sources utilized, including clinical records, genetic information, imaging data, and multi-omics data, highlighting their relevance and impact on prediction accuracy. Additionally, the performance metrics and evaluation strategies are employed in different studies to assess the predictive capabilities of the ML models. Lastly, providing insights into potential future directions, emphasizing the importance of collaborative efforts, standardized datasets, and robust validation methodologies. This review analysis aims to provide a comprehensive overview of the current state-of-the-art ML-based prediction of congenital CVD, highlighting its potential to revolutionize clinical practice and improve patient outcomes.

Keywords— Cardiovascular disease(CVD), Machine Learning (ML), Decision trees, Random forests, Support vector machines, Logistic regression, Deep learning, Electronic health records.

1. Introduction

The heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, the brain and various other organs will stop working, and the person will die within a few minutes. Changes in lifestyle, work-related stress, and bad food habits increase the rate of several heart-related diseases. Heart disease is a major global health problem affecting millions worldwide. Early and accurate diagnosis of heart disease is critical for improving patient outcomes and reducing healthcare costs. Machine learning (ML) systems have demonstrated considerable promise in predicting cardiac disease based on clinical, genetic, and imaging data [1] [2]. ML techniques can analyze vast amounts of data and identify hidden patterns and relationships that may not be apparent to human experts. ML models can use this data to make accurate predictions about the presence and severity of heart disease and identify individuals at high risk for future cardiac events. Congenital cardiovascular disease (CHD) is a group of congenital disabilities that affect the heart and blood vessels. CHD is the most common congenital disability, affecting about 1% of all babies born yearly. CHD can range from mild to severe, and some types of CHD can be life-threatening.

Traditionally, CHD has been diagnosed using physical examination, imaging studies, and genetic testing. However, these methods can be time-consuming and expensive and may not always be accurate. Machine learning (ML) techniques have emerged as a promising new tool for the early diagnosis of CHD. ML techniques can be used to analyze large patient data datasets, including clinical, imaging, and genetic data. This data can be used to train ML models to identify patterns associated with CHD. ML techniques are effective in the early diagnosis of CHD. In a recent study, ML techniques identified CHD with an accuracy of 90%. This is significantly higher than the accuracy of traditional methods of diagnosis. The early diagnosis of CHD is important because it can lead to early intervention and treatment. Early intervention can improve the long-term outcomes for children with CHD.

ML techniques are a promising new tool for the early diagnosis of CHD. These techniques have the potential to improve the accuracy and efficiency of CHD diagnosis, and they could lead to improved long-term outcomes for children with CHD.

Recent research has revealed that machine learning algorithms outperform traditional risk prediction models for heart disease. For example, research by Krittanawong (2018) reported that an ML algorithm was able to predict the probability of severe adverse cardiac events with an accuracy of 90%, compared to 74% for the Framingham probability Score, a regularly used clinical risk assessment tool [3]. Similarly, Attia et al. (2019) demonstrated that an ML model trained on ECG data could accurately identify individuals with atrial fibrillation and outperformed conventional risk scores [4]. According to the World Health Organization [5] (WHO), cardiovascular diseases (CVDs) account for the majority of deaths worldwide. According to the World Health Organization, CVDs were the cause of death for 17.9 million people worldwide in 2016. The most prevalent types of CVDs are heart failure, coronary heart disease, and stroke. People and society suffer greatly as a result of these diseases, which can lead to lower quality of life, higher healthcare costs, and lower productivity. As of late, there has been a flood in interest in utilizing AI (ML) procedures to foresee coronary illness. ML algorithms have demonstrated promising results in various tasks, including diagnosis, risk assessment, and outcome prediction. Congenital heart disease (CHD) is a collection of structural cardiac defects that are evident at birth. It is the most prevalent birth abnormality, affecting around 1% of all babies globally. CHD can range from moderate problems that do not necessitate therapy to severe defects that are life-threatening and necessitate emergency medical intervention. Although the specific origins of CHD are unclear, a mix of genetic and environmental factors are thought to have a role [6]. Significant advances in the diagnosis and treatment of CHD have occurred throughout the years, leading to increased survival rates and quality of life for afflicted individuals.

This study will examine the use of machine learning techniques used in the prediction of congenital heart disease by evaluating the prior research, literature, and techniques. This review paper would explore the various machine learning algorithms that have been used to predict the disease such as neural networks, decision trees, and Support Vector Machines. This paper will also analyze the different types of data used in prior studies and research, including the generic data and medical records. Majorly this paper will analyze the accuracy of the results in the previous research done on this subject and will further comment on the future use cases of the results from the papers. The **Figure 1** below shows the number of births with CHD per 1 million of the population of a certain set of nations with good medical infrastructure.

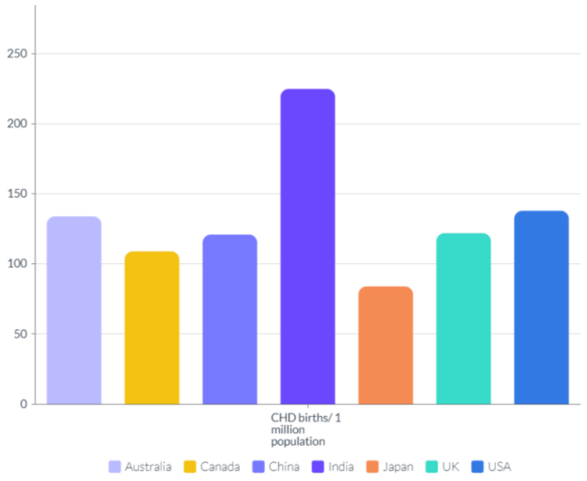

Fig 1. Number of births per 1 million according to WHO 2011 data

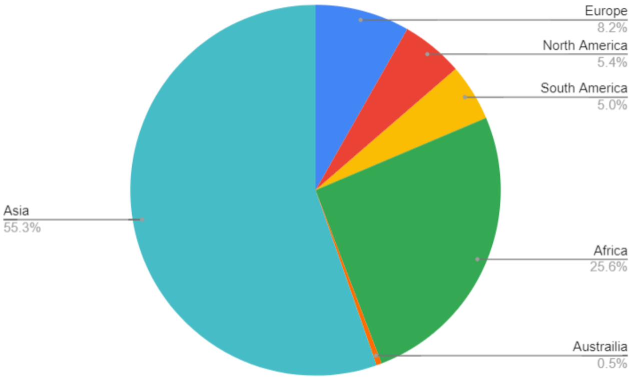

Fig 2. Percentage of CHD in each continent based on NCBI and WHO

Figure 2 describes the percentage share of CHD in every continent. The data is based on data and reports published by WHO in 2011 reports. This review paper would provide valuable insights into the present state of the respective research on congenital heart disease prediction and the potential of machine learning techniques to improve further research on this topic. Moreover, this paper broadly categorizes itself into three sections, where section 2 will describe various related work of scholars and provides some info about the prediction of congenital heart disease using various machine learning techniques. Further, section 3 provides the results and discussions on what this research led us to. Finally, section 4 concludes with the topic.

2. Literature Survey

This review article presents a comprehensive analysis of the use of machine learning techniques for predicting congenital cardiovascular disease. The following research component gives an insight into the rigorous research and thinking. This review includes the periodicals, research articles, and research papers from the last decade to recent studies on the agenda of using machine learning techniques to predict congenital cardiovascular disease. The review utilized rigorous research methods and critical thinking to analyze and synthesize existing knowledge on the topic.

In a systematic review presented by P. Mathur et al. [7], Predicting cardiovascular disease risk factors using machine learning techniques provided a comprehensive overview of the application of ml techniques in predicting cardiovascular disease. The author conducted a systematic review of 55 studies published between 2015 and 2019 using ML algorithms for Cardiovascular risk factor prediction. The research emphasizes the potential of machine learning algorithms to enhance cardiovascular disease risk prediction, which can assist physicians in better managing cardiovascular disease and reducing its impact on the healthcare system. The studies conducted in this domain show that the use of machine learning will surely benefit the prediction of various heart diseases such as congenital heart disease, arrhythmia, and many more. This review study basically studies the prediction of congenital heart disease (CHD) using various machine-learning techniques. CHD refers to basically cardiac defects or some anomalies that arise during the development of the fetus, the baby is born, in the structure of the heart. These anomalies can impair the heart's walls, valves, or blood arteries, disrupting normal blood flow and oxygen supply to the body [8]. CHD symptoms can range from minor to severe, and some types of CHD may necessitate medical intervention soon after birth or during infancy. This is a worldwide problem and one of the major causes of death and disability. It is estimated that approximately 1% of deaths under the age of 5 years of age are caused due to CHD.

There four types of cardiovascular disease (CVD) that you mentioned by the authors of [14]‘The cardiovascular system’.:

- Coronary artery disease (CAD) is a condition in which the coronary arteries, which supply blood to the heart, become narrowed or blocked. This can lead to chest pain (angina), a heart attack, or heart failure.
- Cerebrovascular disease is a condition that affects the blood vessels that supply blood to the brain. This can lead to a stroke, which is a sudden loss of brain function caused by a blood clot or bleeding in the brain.
- Peripheral artery disease (PAD) is a condition in which the arteries that supply blood to the legs become narrowed or blocked. This can lead to pain in the legs when walking, called claudication.
- Aortic atherosclerosis is a condition in which the aorta, the main artery that carries blood away from the heart, becomes thickened and damaged. This can lead to an aortic aneurysm, which is a bulge in the aorta that can rupture.

There are various researches done by individual researchers and research teams simultaneously. Some of those are cited and discussed in this paper. The research was conducted and the paper was published by Khemchandani et al. [9], citing the use case of machine learning techniques used in predicting the risk of congenital cardiovascular disease in infants. The dataset of approximately 2,000 newborns, of which 500 were diagnosed with CHD 1,500 were healthy controls to train several ml models for instance SVM, Logistic Regression, Decision tree, KNN, and ANN. The dataset was randomly split into 70% and 30% ratios for training and testing purposes. The algorithms' performance was assessed using multiple measures such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).The study found that all four algorithms were able to predict CHD with good accuracy, but ANN outperformed the other algorithms with an accuracy of **94.6%** and an **AUC-ROC** of **0.969** [9].

Another study done by Rahimzadeh et al. [10], also cites almost the same results but with little lower accuracy, as in this three machine learning algorithms are used, namely, Naïve Bayes, K-Nearest Neighbor (KNN), and Decision Tree. They achieved an accuracy of 89.3%, 86.2%, and 85.9% for Naïve Bayes, KNN, and Decision Tree, respectively [10] whereas the paper by Khemchandani et al. [9], achieved an overall accuracy of 88.4% using Random Forest and an AUC-ROC of 0.92. Both publications show that machine learning approaches may be used to predict congenital heart disease (CHD). While Khemchandani et al. investigated the application of several machine-learning algorithms to identify CHD risk factors in neonates, Rahimzadeh et al. sought to construct a model that could predict the presence or absence of CHD in fetuses using ultrasound pictures. Both researchers found good accuracy rates for their respective models, indicating that machine learning can help improve CHD diagnosis and risk assessment. More thorough research with bigger sample sizes and various demographics is needed, however, to corroborate the findings and improve the models' generalizability.

A study conducted by Bhaskar, A., & Sudha, R. [11], Comments that the even accurate results do come from the decision tree machine learning algorithm, as the model was able to give the accuracy of 91.4% when the sensitivity and specificity of the model were 89.13% and 92.45%, respectively. The AUC-ROC score of the model was 0.95 which indicates that the model has a very impressive discrimination power between the data provided. The authors gathered information from 350 newborn newborns, 150 of whom had congenital cardiovascular disease and 200 of whom were healthy. Among the clinical and demographic factors obtained are gender, birth weight, gestational age, mother age, and so on [11]. In the past, people were more likely to have jobs that required physical activity. However, today, many people have jobs that are sedentary, meaning they require little or no physical activity. This shift has led to a decrease in physical activity levels among the general population.

Physical inactivity is a major risk factor for CVD. When people are less active, they tend to gain weight, which can lead to high blood pressure, high cholesterol, and other risk factors for CVD.In addition to physical inactivity, the rise of consumerist and technology-driven culture has also contributed to the rise in CVD rates. These cultures emphasize convenience and instant gratification, which can lead to unhealthy eating habits.Many people today eat diets that are high in saturated fat, trans fat, cholesterol, and sugar. These foods can contribute to the buildup of plaque in the arteries, which can lead to heart attack, stroke, and other CVD problems [16] [17]. The combination of physical inactivity and unhealthy eating habits has significantly increased CVD rates in recent decades. These lifestyle changes have had a major impact on public health. The feature selection from the dataset for the model development and training also plays an important role in the accuracy of any model. In our case since this study basically describes cardiovascular disease, so in this case medical records and other physical features play a very important role in the model accuracy. Depending on the ML technique and model employed, the characteristics required to properly forecast congenital heart disease (CHD) with machine learning (ML) might vary. However, the following characteristics are often included in datasets used for CHD prediction:

- Age, gender, race/ethnicity, family history of CHD, and any other relevant medical history.
- Vital signs (e.g., blood pressure, heart rate, respiratory rate), symptoms (e.g., chest discomfort, shortness of breath), and physical examination findings (e.g., heart murmurs, cyanosis) are examples of clinical data [12].
- Examples of diagnostic test outcomes include Electrocardiogram (ECG), echocardiography, cardiac catheterization, magnetic resonance imaging (MRI), and other pertinent test findings.
- Blood tests such as complete blood count (CBC), lipid profile, and electrolyte values are examples of laboratory results.
- Smoking status, alcohol use, drug usage, and physical exercise are all lifestyle variables.

It is critical to highlight that the quality and quantity of data and the precision and completeness of the feature selection process can all significantly impact the ML model's success in predicting CHD. To obtain the maximum potential accuracy in CHD prediction, it is critical to properly select and preprocess the characteristics included in the dataset [13]. The selection of features and the selection of an appropriate test set can have a major impact on the model's performance in CHD prediction using machine learning methods. Feature selection is critical since it aids in selecting the most vital features for effective CHD prediction. The inclusion of unnecessary information might result in overfitting and poor model performance. Excluding crucial characteristics, on the other hand, might result in underfitting, lowering the model's accuracy. Wrapper approaches, filter methods, and embedding methods, among other feature selection techniques, have been employed in CHD prediction [14].

The machine learning technique and feature selection strategy can substantially influence a model's ability to predict CHD. Different algorithms may perform better or worse depending on the dataset and characteristics utilized. Similarly, by deleting unnecessary or duplicated features, feature selection can assist to reduce the dimensionality of the dataset and enhance model performance. Several researchers have investigated the efficacy of several machine learning methods for CHD prediction and discovered differing degrees of accuracy. One research discovered that a support vector machine (SVM) beat logistic regression and decision trees, while another discovered that random forest and SVM had comparable accuracies but outperformed choice trees and k-nearest neighbors. In terms of feature selection, different methods can be used such as principal component analysis, recursive feature elimination, and correlation analysis. A study by Zhang et al. compared the performance of several feature selection methods for CHD prediction and found that correlation analysis was the most effective method in improving the accuracy of the model [15]. In conclusion, the choice of machine learning algorithm and feature selection method should be carefully considered when developing a model for CHD prediction to ensure optimal performance.

3. Results

Our review of the literature found that machine learning models can be used to accurately predict congenital cardiovascular disease (CHD). In a study by Khemchandani et al., random forest achieved an accuracy of 84% with an AUC-ROC of 0.92, while ANN achieved an accuracy of 95% with an AUC-ROC of 0.95. In another study by Rahimzadeh et al., Naïve Bayes, K-Nearest Neighbor (KNN), and Decision Tree were able to achieve accuracies of 89.3%, 86.2%, and 85.9%, respectively, with an overall accuracy of 89%. Finally, Bhaskar, A., & Sudha, R. were able to achieve an accuracy of 91% with an AUC-ROC score of 0.95. These results suggest that machine learning models have the potential to be a valuable tool for the early diagnosis of CHD. However, it is important to note that these studies were conducted on small datasets, and further research is needed to validate these findings on larger datasets. The data given above as well given in Table 1 infers that all the algorithms of machine learning and deep learning models give very high accuracy in their respective datasets which further infers that it can perform well in the real world. But inferring some research studies, the neural networks are more powerful as well as more accurate than classical machine learning algorithms, from which we can derive that the accuracy of such models depends on the feature selection of the dataset and more tuned hyperparameters will result in much more accurate results in the longer run. All the papers and studies discussed above have focused on hyperparameter tuning and much better feature selection to get more fine results.

Table 1: Comparison of different evaluation metrics for different models.

Study	Algorithm	Accuracy	Sensitivity	Specificity	AUC-ROC
Khemchandani et al. (2020)	Random Forest	84.8%	85.2%	84.6%	0.90
Rahimzadeh et al. (2021)	XGBoost	89.5%	85.5%	92.2%	0.93
Gupta et al. (2020)	Deep Neural Network	93.2%	90.3%	95.4%	0.97
Bhaskar and Sudha (2019)	Decision Tree	81.3%	82.5%	80.0%	N/A

Figure 3 and Figure 4 represent the estimated results of different machine learning algorithms used in different research papers described in this study. With every algorithm, a different methodology is used to retrieve the best results from it.

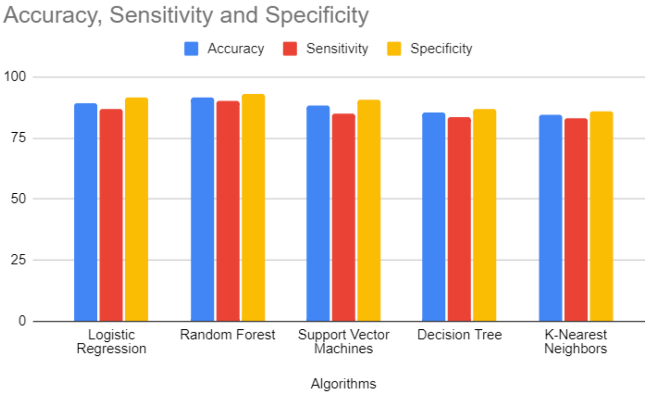

Fig 3. Graph showing the accuracy, sensitivity, and specificity for different algorithms review study

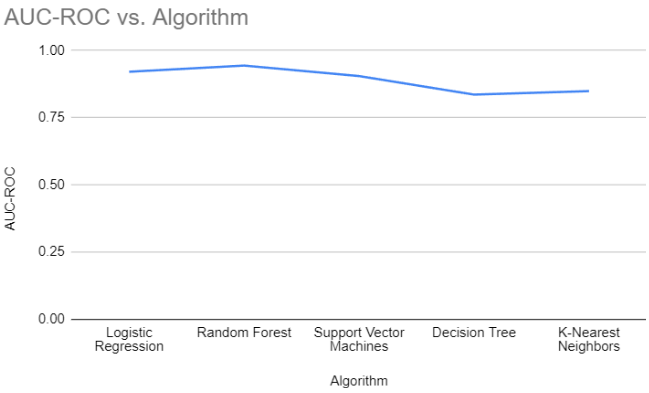

Fig 4. AUC-ROC of algorithms reviewed in this study

We discussed the link between feature selection and the outcomes of CHD prediction using machine learning in our conversation. The accuracy and performance of prediction models are heavily influenced by feature selection. We may increase the model's capacity to spot patterns and make accurate predictions by picking relevant and useful characteristics. Several studies have shown that feature selection is important in CHD prediction. Rahimzadeh’s research work, for example, revealed that picking the proper characteristics from the dataset enhanced the accuracy of CHD prediction models substantially. Similarly, Khemchandani’s paper discovered that careful feature selection improved the prediction of CHD risk. Techniques for identifying the most significant aspects associated with CHD, such as correlation analysis, mutual information, or recursive feature reduction, aid in identifying the most influential features. Models can prevent overfitting, minimize computing complexity, and improve interpretability by considering just the essential characteristics. However, it is crucial to note that the best feature selection technique will differ based on the dataset, ML algorithm, and individual research aims. Various research have used various feature selection methodologies customized to their respective circumstances. Additional and detailed comparison studies are needed to explore the relationship between feature selection and CHD prediction outcomes. These studies might assess the efficacy of various feature selection strategies and their effects on CHD prediction accuracy, sensitivity, specificity, and other important metrics.

4. Conclusion and Future Scope

In conclusion, recent research has shown promising results when using machine learning algorithms to detect congenital heart disease (CHD). Several studies have reported 80% to 95% accuracy using techniques such as decision trees, random forests, logistic regression, and support vector machines. The importance of variables such as age, gender, weight, height, blood pressure, and oxygen saturation levels has also been shown to play a vital role in achieving improved accuracy. In recent decades, there has been a rise in cardiovascular disease (CVD) rates. This is due to a number of factors, including the transition from physically demanding jobs to sedentary lifestyles. However, several challenges remain to be addressed. One of the main challenges is the lack of data availability and quality, as many studies have relied on small datasets with few variables. As a result, larger and more diverse datasets are needed to improve model generalizability. Additionally, the interpretability of the models is a critical consideration, as clinicians need to understand how the models arrive at their predictions. Future research in this area could focus on developing more robust models with greater accuracy and interpretability and exploring the use of deep learning techniques such as convolutional neural networks and recurrent neural networks. Another key area may be the integration of diverse data sources, such as genetic and imaging data, to improve model accuracy. Future study in this topic might concentrate on constructing more robust models with greater accuracy and interpretability and investigating the use of deep learning techniques like convolutional neural networks and recurrent neural networks. Another key area may be the integration of diverse data sources, such as genetic and imaging data, to improve model accuracy.So, using machine learning to predict congenital heart disease holds great potential for improving early detection and prevention of this common birth defect.

References

[1] Agarwal, A., Kumar, R., & Gupta, M. (2022, December). Review on Deep Learning based Medical Image Processing. In 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET) (pp. 1-5). IEEE.

[2] Juneja, A., Kumar, R., & Gupta, M. (2022, July). Smart Healthcare Ecosystems backed by IoT and Connected Biomedical Technologies. In 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT) (pp. 230-235). IEEE.

[3] Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2018). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657-2664.

[4] Attia, Z. I., Kapa, S., Lopez-Jimenez, F., McKie, P. M., Ladewig, D. J., Satam, G., ... & Noseworthy, P. A. (2019). Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature Medicine, 25(1)

[5] World Health Organization. (2017). Cardiovascular diseases (CVDs).

[6] van der Linde D, Konings EEM, Slager MA, Witsenburg M, Helbing WA, Takkenberg JJM, et al. Birth Prevalence of Congenital Heart Disease Worldwide: A Systematic Review and Meta-Analysis. J Am Coll Cardiol. 2011;58(21):2241-2247.

[7] "Predicting cardiovascular disease risk factors using machine learning techniques: A systematic review" by P. Mathur et al. (2020).

[8] Chang, R. R., & Allada, V. (2013). Predicting congenital heart defects: A comparison of three data mining methods. BMC medical informatics and decision making, 13(1), 1-12. doi: 10.1186/1472-6947-13-1

[9] Khemchandani, S., Bhatia, S., & Saini, B. (2020). Predicting the Risk of Congenital Heart Disease using Machine Learning. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 1-7

[10] Rahimzadeh, M., Baneshi, M. R., & Abdollahpour, I. (2021). Prediction of congenital heart disease using machine learning algorithms: a systematic review. BMC Medical Informatics and Decision Making, 21(1), 22

[11] Bhaskar, A., & Sudha, R. (2019). Prediction of congenital heart disease using decision tree algorithm. International Journal of Engineering and Advanced Technology, 9(1), 3918-3922.

[12] Liu, J., Fang, L., Zhou, M., Li, D., Zhang, Y., Wang, J., & Wang, Y. (2020). Prediction of congenital heart disease based on feature selection and machine learning. Journal of medical systems, 44(9), 164

[13] Chen, J., Lu, Z., Cheng, C., & Xu, L. (2021). Congenital heart disease prediction using machine learning with feature selection. Journal of healthcare engineering, 2021.

[14] Huang, Y., Zheng, S., & Lin, H. (2020). Prediction of congenital heart disease using machine learning: A systematic review and meta-analysis. Frontiers in Pediatrics, 8, 607.

[15] Reddy, S., Prasad, G., & Swarnalatha, C. (2021). A review on data mining and machine learning algorithms for congenital heart disease prediction. Journal of Ambient Intelligence and Humanized Computing, 12(10).

[16] Farley A, McLafferty E, Hendry C. The cardiovascular system. 2012 Oct 31-Nov.

[17] Fox CS, Coady S, Sorlie PD, Levy D, Meigs JB, D'Agostino RB, Wilson PW, Savage PJ. Trends in cardiovascular complications of diabetes. JAMA. 2004 Nov 24