U.PORTO

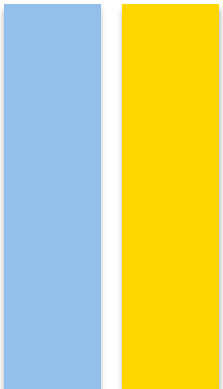
2º CICLO DE ESTUDOS
MESTRADO EM INFORMÁTICA MÉDICA

Diagnosing Congenital Heart Disease Using Clinical and Auscultation Data: A Bayesian Approach

Maria João Veiga Brito

M

2026



U.PORTO

2º CICLO DE ESTUDOS
MESTRADO EM INFORMÁTICA MÉDICA

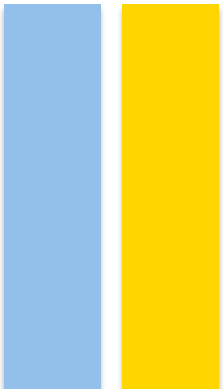
Diagnosing Congenital Heart Disease Using Clinical and Auscultation Data: A Bayesian Approach

Maria João Veiga Brito

Advisors:
Professor Doutor Pedro Pereira Rodrigues
Professor Paulo Jorge Fernandes Dias da Costa

M

2026



Acknowledgments

This dissertation has been the greatest academic challenge I have ever undertaken, but it has also been a testament to perseverance. More than an academic achievement, it is a personal reminder to never give up.

I would like to express my sincere gratitude to everyone who has been part of this journey, whether directly or indirectly.

To my family, thank you for always stepping in whenever life became overwhelming, and for being my greatest supporters and constant source of encouragement.

To Professor Pedro Pereira Rodrigues, thank you for accepting this challenge, for continually pushing me to do better, and for never giving up on this project, even when I was close to giving up myself.

To Professor Paulo Costa, thank you for sharing your knowledge, but above all for your genuine interest, guidance, and constant availability throughout this work.

To my friends, thank you for your encouragement, your patience, and for always being there, even when my free time became increasingly scarce. Your friendship, support, and ability to make even the most stressful moments a little lighter meant more than you know.

Abstract

Introduction: Congenital heart disease (CHD) is the most common group of congenital anomalies worldwide and remains a major cause of morbidity and mortality in paediatric populations. Early detection is essential to improve outcomes, yet access to specialised diagnostic tools such as echocardiography is often limited, particularly in resource-constrained settings. There is, therefore, a need for reliable and interpretable decision-support tools to support clinical screening.
Aim: To develop and evaluate interpretable probabilistic models for paediatric CHD screening using routinely collected clinical and auscultatory data.
Methods: A retrospective modelling study was conducted using data from 865 paediatric subjects enrolled in a cardiovascular screening programme. Following outcome definition, 701 subjects with sufficient diagnostic information to define the outcome were included in supervised analyses. Clinically relevant predictors available at the time of screening were selected and discretised. Bayesian Network (BN), Naïve Bayes (NB), and Tree-Augmented Naïve Bayes (TAN) models were developed and evaluated using cross-validation. Model performance was assessed using area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, precision, and F1-score. Threshold optimisation was performed to explore clinical applicability in a screening context.
Results: All models demonstrated comparable discriminative performance for binary CHD classification. The Naïve Bayes model achieved the highest observed AUC (0.781), closely followed by the TAN model (0.778) and the core Bayesian Network model (0.772). Expanding the predictor set did not improve performance and reduced the effective sample size due to missing data. Murmur-related variables consistently emerged as the most informative predictors. Threshold optimisation increased sensitivity to clinically relevant levels (≥85%), at the cost of reduced specificity and higher referral rates. Four-class diagnostic classification showed substantially lower performance.
Discussion: Probabilistic models provide an interpretable and flexible framework for supporting congenital heart disease screening using routinely collected clinical, physiological, and auscultation data. The findings suggest that carefully selected clinical predictors may have a greater influence on screening performance than increasing model complexity. These models show potential as decision-support tools for early risk stratification and referral, particularly in settings with limited access to specialised care.

Keywords: *congenital heart disease; paediatric screening; Bayesian networks; auscultation.*

Resumo

Introdução: A cardiopatia congénita é o grupo mais frequente de anomalias congénitas a nível mundial e continua a ser uma das principais causas de morbilidade e mortalidade na população pediátrica. A deteção precoce é essencial para melhorar os resultados clínicos, mas o acesso a meios diagnósticos especializados, como a ecocardiografia, é frequentemente limitado. Torna-se, assim, necessária a existência de ferramentas de apoio à decisão clínica que sejam simultaneamente fiáveis e interpretáveis.

Objetivos: Desenvolver e avaliar modelos probabilísticos interpretáveis para o rastreio de cardiopatia congénita em idade pediátrica, com base em dados clínicos e de auscultação recolhidos por rotina.

Métodos: Foi realizado um estudo retrospetivo de modelação com dados de 865 participantes pediátricos incluídos num programa de rastreio cardiovascular. Após definição da variável-alvo, 701 participantes com informação diagnóstica suficiente foram incluídos nas análises supervisionadas. Os preditores clinicamente relevantes disponíveis no momento do rastreio foram selecionados e discretizados. Foram desenvolvidos modelos de Redes Bayesianas (BN), Naïve Bayes (NB) e Tree-Augmented Naïve Bayes (TAN), avaliados através de validação cruzada. O desempenho foi analisado através da área sob a curva ROC (AUC), exatidão, sensibilidade, especificidade, precisão e F1-score. Foi ainda realizada otimização de limiares para avaliar a aplicabilidade clínica em contexto de rastreio.

Resultados: Todos os modelos apresentaram um desempenho discriminativo moderado e globalmente semelhante na classificação binária de cardiopatia congénita. O modelo Naïve Bayes apresentou a AUC mais elevada (0,781), seguido de perto pelo modelo TAN (0,778) e pelo modelo de Rede Bayesiana base (0,772). A inclusão de variáveis adicionais não melhorou o desempenho e reduziu o tamanho efetivo da amostra devido à ausência de dados. As variáveis relacionadas com sopros cardíacos destacaram-se consistentemente como os preditores mais informativos. A otimização de limiares permitiu atingir níveis de sensibilidade clinicamente relevantes (≥85%), com redução da especificidade e aumento das taxas de referenciação. A classificação em quatro categorias diagnósticas apresentou desempenho substancialmente inferior.

Discussão: Os modelos probabilísticos constituem uma abordagem interpretável e flexível para apoiar o rastreio de cardiopatia congénita com base em dados recolhidos por rotina. Os resultados

sugerem que modelos mais simples podem apresentar desempenho semelhante a estruturas mais complexas, desde que os preditores sejam cuidadosamente selecionados. Estes modelos apresentam potencial como ferramentas de apoio à decisão para estratificação precoce do risco e orientação da referenciação, particularmente em contextos com acesso limitado a cuidados especializados.

Palavras-Chave: *cardiopatia congénita; rastreio pediátrico; redes Bayesianas; auscultação.*

Preamble

This dissertation was developed within the context of a Master's degree in Medical Informatics and is grounded in the intersection between clinical practice and data-driven decision support. It was motivated by direct experience in paediatric healthcare settings, where early identification of congenital heart disease remains both essential and, in many cases, challenging.

In routine practice, congenital heart disease is often first suspected during initial clinical assessment, frequently before access to specialised diagnostic resources. In these situations, clinicians must rely on incomplete and sometimes heterogeneous information, making decisions under uncertainty. This highlights the importance of tools that can support, rather than replace, clinical judgement, while remaining transparent and clinically interpretable.

The aim of this work is to explore how probabilistic modelling approaches, particularly Bayesian networks, can be applied to routinely collected clinical and auscultatory data to support paediatric CHD screening.

This work reflects an iterative process of exploration and refinement. Rather than presenting a definitive diagnostic solution, it aims to contribute to the development of interpretable and practically applicable decision-support tools in paediatric healthcare.

Table of Contents

Acknowledgments i

Abstract iii

Resumo v

Preamble vii

Table of Contents ix

List of Figures xiii

List of Tables xv

Acronyms xvii

Dissertation Organisation xix

1 Introduction 21

2 Aim 23

3 Background 25

3.1 Epidemiology and Clinical Burden of Congenital Heart Disease 25

3.2 Aetiology and Risk Factors 26

3.3 Diagnosis Pathways and Global Disparities 26

3.3.1 Prenatal Diagnosis 26

3.3.2 Postnatal Diagnosis 27

3.3.3 Global Disparities in Detection and Management 28

3.4 Artificial Intelligence and Machine Learning in Healthcare 29

3.4.1 Overview of Artificial Intelligence and Machine Learning 29

3.4.2 Early machine-learning applications 30

3.4.3 Auscultation and Phonocardiogram-Based Diagnostics 30

3.4.4 Prenatal and Epidemiological Risk Prediction 31

3.4.5 Data Quality, Imbalance, and Methodological Challenges 32

3.5 Interpretability and Bayesian Modelling 33
3.5.1 Probabilistic reasoning in medical diagnostics 33
3.5.2 Bayesian networks as interpretable decision-support tools 34
3.5.3 Bridging machine learning and Bayesian reasoning 34
3.6 Synthesis and Methodological Motivation 34
4 Methods 37
4.1 Study Design and Modelling Framework 37
4.2 Data Sources and Study Population 37
4.3 Outcome Definition (Reference Standard) 38
4.4 Predictor Selection and Variable Preprocessing 39
4.5 Missing Data Handling 40
4.6 Bayesian Network Model Development 41
4.7 Model Evaluation 42
5 Results 45
5.1 Study Population and Outcome Distribution 45
5.2 Missing Data Analysis 46
5.3 Model BN-V1: Core Clinical Predictor Model 46
5.4 Model BN-V2: Expanded Predictor Model Including Blood Pressure 49
5.5 Model BN-V3: Murmur Characterisation Model 50
5.6 Analysis of Predictor Strength Across Bayesian Network Models 52
5.7 Model NB-V1: Naïve Bayes Classification Model 53
5.8 Model TAN-V1: Tree-Augmented Naïve Bayes Classification Model 55
5.9 Comparative Analysis of Binary Outcome Models 56
5.10 Models Using Four-Class Outcome (CHD_4Class) 58
5.10.1 Model BN-V1-4Class: Bayesian Network with Hill-Climbing Structure Learning 58
5.10.2 Model NB-V1-4Class: Naïve Bayes Classification Model 60
5.10.3 Summary of Four-Class Modelling 60
6 Discussion 61
6.1 Overview of the main findings 61
6.2 Interpretation of model performance for binary CHD detection 62
6.3 Clinical Interpretation of the Most Informative Predictors 63

6.3.1 Auscultation and Clinical Examination Predictors 64

6.3.2 Secondary Predictors and Predictor Selection 64

6.4 Why Expanded Predictor Sets Did Not Improve Performance 65

6.5 Comparison Between General Bayesian Networks, Naïve Bayes, and Tree-Augmented Naïve Bayes Models 66

6.6 Threshold Optimisation and Clinical Utility in Screening Contexts 67

6.7 Interpretation of the Multi-Class Modelling Results 68

6.7.1 Challenges of Multi-Class Classification Using Routine Screening Variables 69

6.7.2 Class Imbalance and Clinical Implications 69

6.8 Strengths of the Study 70

6.9 Limitations 71

7 Conclusion 73

8 Implications for Future Research 75

9 References 77

10 Appendices 83

Appendix A – Variable Definitions and Preprocessing Summary 84

List of Figures

Figure 5.1 – Consensus Bayesian Network structure for Model BN-V1 derived using bootstrap resampling (R = 500) and arc strength threshold of 0.7 47
Figure 5.2 – Receiver operating characteristic curve for Model BN-V1 (AUC = 0.772) 48
Figure 5.3 – Consensus Bayesian Network structure for Model BN-V2 derived using bootstrap resampling (R = 500) and arc strength threshold of 0.7 49
Figure 5.4 – Receiver operating characteristic curve for Model BN-V2 (AUC = 0.759) 50
Figure 5.5 – Consensus Bayesian Network structure for Model BN-V3 derived using bootstrap resampling (R = 500) and arc strength threshold of 0.7 51
Figure 5.6 – Receiver operating characteristic curve for Model BN-V3 (AUC = 0.741) 52
Figure 5.7 – Bootstrap arc strengths of the strongest predictors directed toward CHD_Binary across Bayesian Network models BN-V1, BN-V2, and BN-V3 with 500 bootstrap samples 53
Figure 5.8 – Receiver operating characteristic curve for Model NB-V1 (AUC = 0.781) 54
Figure 5.9 – Tree-Augmented Naïve Bayes network structure for Model TAN-V1 55
Figure 5.10 – Receiver operating characteristic curve for Model TAN-V1 (AUC = 0.778) 56
Figure 5.11 – Bayesian Network structure for Model BN-V1-4Class 59

List of Tables

Table 5.1 – Study Population Characteristics 45
Table 5.2 – Predictive performance of Model BN-V1 at standard and referral-optimised probability thresholds 48
Table 5.3 – Predictive performance of Model BN-V2 at standard and referral-optimised probability thresholds 50
Table 5.4 – Predictive performance of Model BN-V3 at standard and referral-optimised probability thresholds 52
Table 5.5 – Predictive performance of Model NB-V1 at standard and referral-optimised probability thresholds 54
Table 5.6 – Predictive performance of Model TAN-V1 at standard and referral-optimised probability thresholds 56
Table 5.7 – Comparative performance of Bayesian Network models for binary congenital heart disease classification 58
Table 5.8 – Confusion matrix for Model BN-V1-4Class 59
Table 5.9 – Confusion matrix for Model NB-V1-4Class 60
Table A.1 – Outcome Variables Used in Model Development 84
Table A.2 – Predictor Variables Included in Each Model 85
Table A.3 – Age-Specific Heart Rate Thresholds Used to Derive Heart Rate Class 88
Table A.4 – Age-Specific Systolic Blood Pressure Thresholds Used to Derive SBP Class 89
Table A.5 – Age-Specific Diastolic Blood Pressure Thresholds Used to Derive DBP Class 90

Acronyms

AHA American Heart Association

AI Artificial Intelligence

AUC Area Under the Curve

BDeu Bayesian Dirichlet Equivalent Uniform

BN Bayesian Network

CHD Congenital Heart Disease

ECG Electrocardiogram

HICs High-Income Countries

LMICs Low- and Middle-Income Countries

MACS Munich Age Classification System

ML Machine Learning

NB Naïve Bayes

PALS Pediatric Advanced Life Support

PCG Phonocardiogram

QT QT interval

ROC Receiver Operating Characteristic

RRNN Residual-Recurrent Neural Network

SVM Support-Vector Machine

TAN Tree-Augmented Naïve Bayes

WHO World Health Organization

Dissertation Organisation

Chapter 1 (Introduction) presents the clinical and methodological context of paediatric congenital heart disease screening. It outlines the challenges associated with early detection and introduces the motivation for developing interpretable decision-support tools, particularly in real-world and resource-limited settings.

Chapter 2 (Aim) defines the main objectives of the dissertation, focusing on the development and evaluation of interpretable probabilistic models to support CHD screening in paediatric populations.

Chapter 3 (Background) reviews the clinical and methodological concepts necessary to contextualise the study. This includes an overview of congenital heart disease, current screening approaches, auscultation findings, and the role of artificial intelligence and Bayesian modelling in healthcare.

Chapter 4 (Methods) describes the study design and analytical approach. It details the data sources, cohort selection, outcome definition, predictor preprocessing, and handling of missing data. It also presents the modelling strategies used, including Bayesian Networks, Naïve Bayes, and Tree-Augmented Naïve Bayes, as well as the evaluation framework based on cross-validation and performance metrics.

Chapter 5 (Results) presents the findings of the modelling process. This includes descriptive analysis of the study population, assessment of missing data, learned network structures, and predictive performance of each model. A comparative analysis of the different modelling approaches is also provided.

Chapter 6 (Discussion) interprets the results in light of existing literature and clinical knowledge. It explores the role of key predictors, discusses the performance of different modelling strategies, and evaluates the potential clinical utility of the proposed approach, along with its limitations.

Chapter 7 (Conclusion) summarises the main findings of the dissertation and reflects on their clinical and methodological implications.

Chapter 8 (Implications for Future Research) outlines potential directions for further work, including model refinement, incorporation of additional data sources, and opportunities for validation in different clinical settings.

1 Introduction

Congenital heart disease (CHD) represents the most prevalent group of congenital anomalies worldwide and remains a leading cause of morbidity and mortality among the paediatric population, particularly during the neonatal and infant stages (Cheng et al., 2024; van der Linde et al., 2011; World Health Organization, 2023). Despite major advances in diagnostic imaging, surgical techniques, and perioperative care, timely identification of clinically significant heart defects continues to play a critical role in improving long-term outcomes (Meller et al., 2020; Reddy et al., 2022). Delayed or missed diagnoses may lead to avoidable complications, while unnecessary referrals can place additional strain on already limited healthcare resources (Cheng et al., 2024; Oster et al., 2025).

Access to paediatric cardiology services and diagnostic tools, such as echocardiography, remains uneven across regions. In many settings, particularly those that are resource-limited or geographically remote, specialised equipment and trained personnel are not readily available (Cheng et al., 2024). As a result, early detection of CHD often relies on clinical screening methods, including patient history, physical examination, vital signs, and cardiac auscultation (Hoffman & Kaplan, 2002). These approaches are non-invasive, cost-effective, and widely used in practice, but their effectiveness depends heavily on the clinician's experience and confidence in interpreting findings (Oliveira et al., 2022; Oster et al., 2025).

Cardiac auscultation remains the cornerstone of paediatric cardiac assessment and is frequently the first indicator of underlying heart disease. However, interpretation of heart sounds and murmurs is inherently subjective and can vary considerably between observers, particularly among non-specialists or clinicians working outside tertiary care centres (Mangione & Nieman, 1997; Zhao et al., 2019). This variability is further compounded by limited opportunities for training and supervised practice, as many healthcare professionals encounter relatively few children with congenital heart disease during routine clinical work, making it difficult to develop and maintain proficiency in paediatric cardiac auscultation (Mangione & Nieman, 1997; Oliveira et al., 2022). In practice, these challenges may contribute to delayed diagnosis, missed cases of clinically relevant CHD, and unnecessary referrals for benign findings. These limitations highlight the need for tools that can support more consistent and clinically meaningful interpretation of screening data.

In recent years, data-driven and machine learning approaches have been increasingly explored to support CHD screening and clinical decision-making. These methods have been applied to tasks such as murmur classification, analysis of phonocardiographic signals, and risk stratification using clinical variables (Clifford et al., 2016; Oliveira et al., 2022). While many approaches report promising predictive performance, a substantial proportion rely on complex "black-box" models, such as deep neural networks, which offer limited interpretability. In clinical settings, a lack of transparency can reduce trust, limit adoption, and complicate integration into routine workflows (Rudin, 2019).

Another challenge in real-world clinical datasets is the presence of missing or incomplete data. Vital signs may not always be recorded, anthropometric measurements may be unavailable, and auscultation findings may be incomplete or uncertain (Kyrimi et al., 2021). Models that do not explicitly account for missingness may produce biased predictions or demonstrate reduced generalisability when applied in external clinical settings (Van Calster et al., 2026). Addressing missing data in a principled and transparent way is therefore essential when developing clinically applicable decision-support tools.

Bayesian networks offer a probabilistic modelling framework that is well-suited to these challenges (Kyrimi et al., 2021). By representing variables as nodes connected through conditional dependencies, they allow explicit modelling of relationships between demographic factors, clinical findings, physiological measurements, and disease states (Korb & Nicholson, 2011). Their graphical structure supports interpretability, enabling clinicians and researchers to inspect and assess the relationships the model has learned against clinical knowledge. In addition, Bayesian networks can naturally accommodate uncertainty and missing data, allowing inference even when some variables are unobserved (Kyrimi et al., 2021; Lucas et al., 2004).

In the context of paediatric CHD screening, Bayesian networks provide a way to integrate heterogeneous sources of information, such as clinical examination findings and auscultatory features, within a single, interpretable framework (Kyrimi et al., 2021). Rather than replacing specialist evaluation or diagnostic imaging, such models can function as decision-support tools, assisting clinicians in early risk stratification, prioritising referrals, and improving the efficiency of screening programmes (Oster et al., 2025). This approach is particularly relevant in settings where access to paediatric cardiology expertise is limited and where screening decisions are often made under uncertainty (Cheng et al., 2024; Kyrimi et al., 2021).

This dissertation focuses on the development and evaluation of interpretable probabilistic models for paediatric congenital heart disease screening using routinely collected clinical and auscultatory data. Particular emphasis is placed on careful outcome definition, clinically grounded variable selection, transparent handling of missing data, and avoidance of information leakage. By aligning methodological choices with the realities of real-world clinical practice, this work aims to contribute a robust and interpretable decision-support framework with potential applicability across diverse screening contexts.

2 Aim

The main objective of this dissertation is to develop and evaluate interpretable probabilistic models to support paediatric congenital heart disease screening using routinely collected clinical, physiological, and auscultation data. The proposed approach aims to assist in identifying and prioritising children who may benefit from further cardiological evaluation, particularly in screening and pre-referral contexts.

Specifically, this work explores Bayesian network-based modelling approaches, including general Bayesian Networks and classification-oriented variants such as Naïve Bayes and Tree-Augmented Naïve Bayes. These models are designed to integrate heterogeneous clinical information, handle missing data in a principled manner, and reflect clinically plausible relationships between variables.

By doing so, the models aim to support earlier identification of children at increased risk of CHD, facilitate more informed referral decisions, and provide transparent probabilistic outputs that can be meaningfully interpreted by clinicians.

3 Background

3.1 Epidemiology and Clinical Burden of Congenital Heart Disease

Congenital heart disease (CHD) encompasses a broad range of structural abnormalities of the heart and great vessels that arise during foetal development and are present at birth (Meller et al., 2020; Zaidi & Brueckner, 2017). These defects may involve the cardiac septum, valves, or outflow tracts, leading to altered haemodynamics and impaired systemic or pulmonary circulation (Meng et al., 2024). Clinically, CHD is often stratified into simple and complex forms according to anatomical severity and expected management complexity (Zaidi & Brueckner, 2017). Presentations vary widely, ranging from asymptomatic or incidental heart murmurs in mild lesions to cyanosis, respiratory distress, or overt heart failure in critical cases (Kaur & Ahmad, 2024; Kumar et al., 2025). Cyanosis typically becomes apparent when arterial oxygen saturation falls below approximately 80%, although detection may be more challenging in neonates with darker skin tones (Kumar et al., 2025). Other frequent manifestations include tachycardia, poor feeding, diaphoresis, and fatigue (Meng et al., 2024).

CHD is the most common major congenital anomaly worldwide, affecting approximately one in every 100 live births and accounting for nearly one-third of all congenital malformations (Cheng et al., 2024; Helman et al., 2021; Zaidi & Brueckner, 2017). The World Health Organization (2023) identifies congenital disorders as a substantial global public health concern, contributing significantly to perinatal morbidity, long-term disability, and mortality. CHD is estimated to account for around 10% of infant deaths and nearly half of deaths associated with congenital malformations (Chami et al., 2023; Kumar et al., 2025; Lopes et al., 2018), highlighting its substantial impact on child health.

Despite advances in paediatric cardiology and surgical care, CHD remains a leading cause of paediatric morbidity (Zaidi & Brueckner, 2017). Many affected children experience chronic complications, including heart failure, arrhythmias, pulmonary hypertension, and growth or neurodevelopmental delays, which contribute to lifelong healthcare needs and reduced quality of life (Kaur & Ahmad, 2024; Meller et al., 2020; Meng et al., 2024). Approximately half require at least one surgical or catheter-based intervention during their lifetime (Cheng et al., 2024). Without timely diagnosis and appropriate treatment, many children with severe forms of CHD do not survive beyond early childhood (Lopes et al., 2018).

Taken together, the high prevalence, substantial mortality, and enduring morbidity associated with CHD highlight the critical importance of early identification and risk stratification (Chami et al., 2023; Meng et al., 2024). Timely diagnosis enables appropriate referral, optimised perinatal and postnatal management, and improved survival outcomes (Lopes et al., 2018; Zaidi & Brueckner, 2017). These needs are particularly noticeable in resource-limited settings, where delayed recognition and restricted access to

specialist care aggravate preventable complications (Chami et al., 2023). Consequently, scalable and accessible screening strategies are essential to reduce diagnostic delays and address persistent global inequities in CHD care (Chami et al., 2023).

3.2 Aetiology and Risk Factors

The aetiology of congenital heart disease is multifactorial, involving a complex interaction between genetic, epigenetic, and environmental influences that disrupt normal cardiac development during early embryogenesis (Choudhury et al., 2025; F. Zhang et al., 2024). Experimental studies in animal models have shown that disruptions in key developmental pathways can lead to structural abnormalities, although the precise mechanisms in humans remain only partially understood (Meller et al., 2020; Sun et al., 2015). This biological complexity suggests that CHD rarely arises from a single causal factor, but rather from the interaction of multiple vulnerabilities operating during critical periods of gestation (Lemieux et al., 2024).

Genetic and molecular contributors play a substantial role in CHD pathogenesis (Chhatwal et al., 2023; Choudhury et al., 2025). Numerous chromosomal abnormalities, copy-number variations, and gene mutations have been associated with congenital cardiac defects, including both syndromic and non-syndromic presentations (Meller et al., 2020). However, environmental and maternal conditions frequently modify this genetic predisposition. Maternal infections, diabetes, obesity, alcohol or drug exposure, and nutritional deficiencies, particularly insufficient folic-acid intake, have all been linked to an increased risk for abnormal cardiac development (Lemieux et al., 2024; S. Zhang et al., 2025). These exposures may influence embryogenesis through metabolic, inflammatory, or teratogenic mechanisms, increasing the likelihood of structural defects.

Epidemiological studies also highlight the importance of socioeconomic and environmental determinants (Xu et al., 2025). In a matched case–control study conducted in China, S. Zhang et al. (2025) reported that short duration of folic-acid supplementation, rural residence, and low maternal wealth index were among the strongest predictors of CHD occurrence, although findings for folic-acid supplementation are not fully consistent across populations (Wang et al., 2025). These findings highlight that CHD risk is shaped not only by biological susceptibility but also by broader structural inequities that affect maternal–child health outcomes.

Although genetic testing and foetal echocardiography have advanced substantially in high-income countries, social determinants of health remain dominant predictors of disease burden in low- and middle-income settings (Cheng et al., 2024; World Health Organization, 2023). Limited access to antenatal screening, delayed diagnosis, and restricted specialist care contribute to preventable morbidity and mortality (Cheng et al., 2024; Lopez-Barreda et al., 2025).

3.3 Diagnosis Pathways and Global Disparities

3.3.1 Prenatal Diagnosis

Congenital heart disease comprises life-threatening structural malformations of the heart and great vessels that are present at birth and frequently require intervention within the first year of life (Helman et al., 2021).

Early identification during pregnancy allows for preventive management, optimised delivery planning, and better neonatal outcomes (Reddy et al., 2022; van Velzen et al., 2016).

Foetal echocardiography remains the foundation of prenatal diagnosis, providing detailed visualisation of cardiac anatomy, haemodynamics, and structural abnormalities (Meller et al., 2020). When complex lesions are identified antenatally, referral to specialised tertiary centres allows coordinated obstetric and neonatal care, including immediate postnatal intervention, which has been associated with reduced morbidity and mortality (Reddy et al., 2022).

Despite these advances, prenatal detection remains incomplete. Even in high-income settings, detection rates for CHD rarely exceed 50–60%, reflecting important limitations in routine screening (Meller et al., 2020). Missed diagnoses may result from limited access to specialised ultrasonography, variability in operator experience, and technical challenges such as maternal obesity or excessive amniotic fluid (Hashiramoto et al., 2025). Detection also varies considerably by lesion type: severe anomalies such as hypoplastic left-heart syndrome are identified in most cases, whereas conditions such as anomalous pulmonary venous return or small septal defects are frequently overlooked (Meller et al., 2020).

The availability of foetal echocardiography is closely linked to healthcare resources and system organisation. In many low- and middle-income countries (LMICs), access to high-quality antenatal imaging and specialist expertise remains limited, leading to delayed diagnosis and restricted referral pathways (Cheng et al., 2024). These disparities highlight the persistent global inequities in maternal-child healthcare and early CHD detection, and emphasise the need for scalable, resource-appropriate diagnostic strategies (Lopez-Barreda et al., 2025).

3.3.2 Postnatal Diagnosis

When prenatal detection is missed or unavailable, the diagnosis of congenital heart disease relies primarily on postnatal clinical evaluation (Meller et al., 2020). Suspicion is typically raised during routine newborn assessment through abnormal findings on physical examination or screening procedures such as pulse oximetry, electrocardiography, or chest radiography (Kumar et al., 2025; Meller et al., 2020).

Cardiac auscultation remains a fundamental first-line diagnostic tool, allowing clinicians to identify heart murmurs suggestive of structural abnormalities (Oliveira et al., 2022; Zhao et al., 2019). Definitive diagnosis, however, is established through echocardiography, which provides direct visualisation of the cardiac structure and haemodynamics, and is widely regarded as the diagnostic gold standard for CHD confirmation (Meller et al., 2020; Reddy et al., 2022).

The accuracy of auscultation varies considerably depending on clinician expertise. Zhao et al. (2019) found that general paediatricians demonstrate moderate sensitivity but high specificity when detecting pathological murmurs, highlighting the risk of missed cases and emphasising the need for additional training or computer-assisted support. These limitations have stimulated growing interest in digital auscultation and automated decision-support tools capable of enhancing clinician performance (Alkhodari et al., 2024; Gudigar et al., 2024).

Pulse-oximetry screening has become a standard component of newborn screening programmes for detecting critical CHD (Kumar et al., 2025). By comparing pre-ductal (right hand) and post-ductal (foot) oxygen saturation values, the method can identify duct-dependent or cyanotic lesions. However, its

sensitivity remains limited for certain defects, particularly left-sided obstructive lesions such as coarctation of the aorta, which may present with normal oxygen saturation values (Kumar et al., 2025; Siefkes et al., 2024).

Recent approaches have attempted to improve detection by incorporating additional physiological parameters into machine-learning-based systems. For example, Siefkes et al. (2024) demonstrated that combining pulse oximetry with perfusion index and pulse-delay features improved sensitivity for critical CHD while maintaining specificity. These findings suggest that machine learning can enhance, rather than replace, existing screening methods.

3.3.3 Global Disparities in Detection and Management

Over the past two decades, advances in paediatric cardiac surgery and perioperative care have led to survival rates exceeding 90% in children who receive timely treatment (Cheng et al., 2024). However, this progress has been achieved predominantly in high-income countries (HICs). In low- and middle-income countries (LMICs), where approximately nine in every ten children with CHD are born, the prognosis remains considerably poorer owing to a combination of systemic, economic, and logistical barriers (Cheng et al., 2024; Xu et al., 2025).

Paediatric cardiac services in LMICs are frequently concentrated in a limited number of urban tertiary centres, leaving large rural populations without access to echocardiography or paediatric cardiology expertise (Cheng et al., 2024). The World Health Organization (2023) highlights that these inequities are further amplified by shortages of skilled healthcare professionals, insufficient neonatal screening programmes, and the absence of structured referral pathways. Even when a diagnosis is established, financial limitations, transportation barriers, and restricted surgical capacity frequently delay definitive treatment (Cheng et al., 2024).

Several studies underscore how early diagnostic gaps contribute directly to these outcomes. Meller et al. (2020) observed that intrauterine detection is strongly dependent on system-level access to antenatal imaging and operator skill, with many obstetric ultrasounds reporting only foetal heart rate without anatomical cardiac assessment. Additionally, technical factors such as maternal obesity, polyhydramnios, and suboptimal imaging equipment may hide structural defects even in experienced hands (Hashiramoto et al., 2025). Consequently, potentially correctable lesions, including ventricular septal defects, atrioventricular septal defects, and patent ductus arteriosus, may remain undiagnosed until clinical deterioration occurs (Lopez-Barreda et al., 2025).

Differences in lesion-specific detection rates further illustrate these disparities. While severe anomalies such as hypoplastic left-heart syndrome may be identified prenatally in up to 90% of cases, conditions such as total anomalous pulmonary venous return or minor septal defects are frequently missed (Meller et al., 2020). These discrepancies persist even in well-resourced healthcare systems and are magnified in regions with limited training opportunities (Bakker et al., 2019).

Beyond technical and workforce limitations, socio-economic factors also contribute substantially to CHD outcomes. A systematic review by Cheng et al. (2024) reported that poverty is associated with higher CHD-related mortality, and that greater distance from specialised centres predicts delayed presentation and worse outcomes. Cultural beliefs and limited health literacy may also contribute, as families may not perceive

asymptomatic murmurs as requiring medical attention, postponing evaluation until complications develop (World Health Organization, 2023).

Taken together, these global inequities highlight the need for cost-effective, scalable diagnostic strategies applicable in resource-limited settings. Technologies dependent on advanced imaging or high-bandwidth infrastructure are often unmanageable in these settings (Cheng et al., 2024). In contrast, approaches based on easily obtainable clinical data, including auscultation findings, oxygen saturation, and basic vital signs, offer greater practical feasibility. Artificial-intelligence-based decision-support systems, particularly those employing interpretable probabilistic reasoning such as Bayesian networks, represent a promising path for bridging the diagnostic gap. By formalising expert knowledge while accommodating uncertain or missing data, Bayesian approaches may provide transparent, low-cost diagnostic assistance to frontline clinicians in rural and underserved settings (Helman et al., 2021; Shantha Selvakumari et al., 2025).

3.4 Artificial Intelligence and Machine Learning in Healthcare

3.4.1 Overview of Artificial Intelligence and Machine Learning

Over the past decade, the application of artificial intelligence (AI) and machine learning (ML) in medical diagnostics has expanded rapidly (Topol, 2019). AI broadly refers to computational systems designed to compete with human cognitive processes such as learning, reasoning, and decision-making (Reddy et al., 2022). ML, a subset of AI, focuses on algorithms that learn patterns from data to generate predictions or classifications without relying on explicitly defined rules (Helman et al., 2021).

In the context of CHD, ML models have been applied to a range of tasks, from risk prediction and screening enhancement to diagnostic classification, using diverse modalities that include medical imaging, electrocardiography, and phonocardiography (Helman et al., 2021; Oliveira et al., 2022; Reddy et al., 2022). These approaches are intended not to replace clinicians, but rather to function as decision-support tools capable of analysing large and complex datasets to identify subtle patterns that may not be readily apparent to human observers (Davenport & Kalakota, 2019).

A scoping review by Helman et al. (2021) identified a rapidly growing body of research applying ML to both critical and non-critical CHD. The reviewed approaches ranged from traditional statistical models, such as logistic regression, to more complex methods including decision trees, random forests, support-vector machines, and deep-learning architectures. However, the review also highlighted several methodological limitations, including small sample sizes, inconsistent labelling, and limited integration into clinical workflows.

More recent studies suggest that these challenges remain largely unresolved. A review by Shantha Selvakumari et al. (2025) of computerised diagnostic methods in paediatric cardiology highlighted fragmentation across data modalities and a pressing need for standardisation, external validation, and model explainability. Although reported performance metrics are often high in controlled experimental settings, translation into routine clinical practice remains limited, in part due to restricted interpretability and class imbalance (Kelly et al., 2019). Similarly, Mayer et al. (2024) noted that many datasets used in CHD AI

research are small, single-centre collections with inconsistent annotation, reinforcing the urgency of open, standardised, and multicentre data-sharing initiatives.

3.4.2 Early machine-learning applications

Early machine-learning models for congenital heart disease relied primarily on structured clinical and demographic variables rather than raw physiological signals (Begic et al., 2021). Classical algorithms such as support-vector machines (SVMs), logistic regression, and Naïve Bayes classifiers demonstrated the feasibility of automated classification but were inherently constrained by the quality, dimensionality, and limited size of available datasets.

For example, a Naïve Bayes classifier applied to paediatric electrocardiograms successfully differentiated long-QT syndrome from normal recordings (Qu et al., 2010), illustrating both the interpretability and modest data requirements of probabilistic models.

As research progressed, attention shifted toward challenges such as class imbalance, which is common in paediatric cardiology datasets where pathological cases are relatively rare. Approaches such as cost-sensitive learning were introduced to improve the detection of minority classes. A study presented at the International Joint Conference on Neural Networks by Tavares et al. (2013) proposed weighted SVMs to enhance the prediction of paediatric heart disease in unbalanced datasets, effectively increasing sensitivity for minority classes without compromising overall accuracy. These early developments laid the groundwork for more complex models capable of capturing nonlinear relationships and integrating multiple data types.

The first attempts to automate heart-sound classification using machine learning emerged in the mid-2010s, when researchers began extracting time- and frequency-domain features from phonocardiograms (PCGs) and training SVM or random-forest models for binary classification (normal versus pathological murmurs) (Gudigar et al., 2024). Although typically limited by small sample sizes and handcrafted features, these studies demonstrated the feasibility of automated auscultation as a diagnostic adjunct.

As data availability and computational capacity improved, ML applications gradually shifted from feature-engineering approaches toward deep-learning frameworks capable of end-to-end representation learning directly from raw signals. These architectures now dominate much of the current CHD diagnostic research landscape (Alkhodari et al., 2024; Gudigar et al., 2024; Hsieh et al., 2025).

3.4.3 Auscultation and Phonocardiogram-Based Diagnostics

Digital phonocardiography, recorded as phonocardiograms (PCGs), is gaining renewed attention as a low-cost, portable diagnostic modality capable of supporting CHD detection outside tertiary centres (Gudigar et al., 2024). The availability of larger, open datasets has accelerated methodological progress.

The ZCHSound paediatric database, introduced by Jia et al. (2024), includes over 1 200 children with both normal and pathological recordings, providing benchmark data for supervised and unsupervised learning. Baseline ensemble models achieved strong F1-scores on clean subsets but degraded markedly under real-world noise conditions, emphasising the importance of robust pre-processing and denoising strategies.

A complementary dataset was described in *Computing in Cardiology* by Hassanuzzaman et al. (2023), where more than 700 paediatric phonocardiograms were used to train a deep-learning classifier that differentiated between several CHD subtypes. The study demonstrated that convolutional architectures could capture subtle frequency-domain features that may be imperceptible to human listeners, although the authors reported reduced sensitivity for low-grade murmurs and noisy environments (Hassanuzzaman et al., 2023).

At a broader level, the systematic review by Gudigar et al. (2024) analysed more than 70 studies applying ML or deep learning to PCG signals. The review concluded that while deep architectures—particularly convolutional and recurrent networks—consistently outperform classical ML methods, persistent weaknesses include noise sensitivity, class imbalance, limited external validation, and insufficient interpretability (Gudigar et al., 2024). It further advocated for standardised recording protocols and multicentre validation datasets to ensure reproducibility.

Among recent architectures, attention-based and hybrid models show particular promise. Alkhodari et al. (2024) proposed a transformer network incorporating channel-wise attention to identify congenital valvular murmurs in children. Using multi-site PCGs, the model achieved approximately 90% classification accuracy and provided interpretability through visualisation of attention maps, most notably highlighting the aortic-valve channel for “absent-murmur” predictions and mitral/tricuspid channels when murmurs were present.

Hsieh et al. (2025) advanced this concept with an integrated residual–recurrent neural network (RRNN). Validated on 500 paediatric participants and externally on the PhysioNet 2016 dataset, the model achieved approximately 90% internal and 95% external accuracy while maintaining high specificity across most lesion types. Performance decreased slightly for grade-1 and diastolic murmurs, revealing persistent limitations in data representation. By coupling deep temporal learning with a gradient-boosting layer, the RRNN improved both accuracy and partial interpretability.

Earlier work also demonstrated that simpler, computationally accessible models could achieve clinically meaningful performance. Begic et al. (2021) combined clinical and auscultation variables within a support-vector and Naïve Bayes framework, attaining 97% accuracy in differentiating innocent from pathological murmurs using only 16–19 low-cost features. Similarly, Khan et al. (2023) presented a convolutional deep-learning model for CHD recognition and reported cross-dataset generalisability.

Collectively, these studies suggest that AI-assisted auscultation can approach specialist-level interpretation and may support decentralised or community-based screening (Alkhodari et al., 2024). However, they also reveal recurring limitations—particularly data imbalance, interpretability constraints, and noise sensitivity—which motivate exploration of alternative, more transparent probabilistic frameworks such as Bayesian modelling.

3.4.4 Prenatal and Epidemiological Risk Prediction

Parallel to signal-based diagnostic approaches, several studies have explored machine-learning models for maternal and foetal risk prediction using demographic, clinical, and socioeconomic variables (Reddy et al., 2022). Unlike postnatal diagnostic systems that analyse physiological signals or imaging data, these

approaches aim to identify pregnancies at elevated risk for CHD before birth, enabling earlier referral and targeted surveillance.

In a matched case–control study, S. Zhang et al. (2025) applied five machine-learning algorithms to 47 maternal sociodemographic and clinical variables. Key predictors included rural residence, low household income, and limited folic-acid supplementation. The model's interpretable feature-importance ranking demonstrated how transparent ML approaches can inform targeted public-health interventions and resource allocation.

Although epidemiological risk models do not provide definitive diagnosis, they complement postnatal diagnostic tools by supporting population-level stratification and prioritisation of high-risk pregnancies (World Health Organization, 2023). Their interpretability, low computational requirements, and reliance on routinely collected data make them particularly suitable for low-resource settings where advanced imaging infrastructure or specialist evaluation may be unavailable.

3.4.5 Data Quality, Imbalance, and Methodological Challenges

Despite substantial advances in artificial intelligence for congenital heart disease (CHD) diagnosis, several methodological issues continue to hinder clinical implementation. The most prominent limitations across studies involve data heterogeneity, imbalanced class distributions, and insufficient standardisation and validation.

Data heterogeneity arises because most CHD datasets originate from single institutions with non-uniform acquisition protocols, equipment, and annotation practices. Variations in sampling rates, sensor types, preprocessing pipelines, and labelling standards complicate model comparison, reproducibility, and external generalisation (Kelly et al., 2019). This problem is particularly pronounced in paediatric cardiology, where data collection is constrained by ethical considerations, limited patient numbers, and logistical barriers. Reviews by Mayer et al. (2024), Gudigar et al. (2024) and Shantha Selvakumari et al. (2025) emphasise that the scarcity of multicentre datasets and consistent documentation severely limits generalisability and prevents the establishment of robust benchmarks. Even large public repositories such as ZCHSound (Jia et al., 2024) or PhysioNet provide only partial solutions, as they differ in age ranges, noise profiles, and recording environments.

A second major challenge concerns class imbalance. Most screening datasets contain far more recordings of healthy children than pathological cases, resulting in models that achieve high overall accuracy but perform poorly in detecting rare yet clinically critical conditions. Early research addressed this issue through cost-sensitive learning approaches, such as weighted support-vector machines that assign higher misclassification costs to minority classes to improve sensitivity (Tavares et al., 2013). More recent strategies, including synthetic oversampling and focal-loss optimisation, have further attempted to mitigate imbalance (Chen et al., 2024), though these techniques may introduce artificial patterns, reduce interpretability, or distort probability estimates. In contrast, Bayesian or probabilistic models can address imbalance more naturally through prior probability weighting that reflects the true epidemiological prevalence of disease while preserving transparent probabilistic reasoning (Arora et al., 2019; Pearl, 2014).

Another critical limitation is the scarcity of external validation. As noted by Gudigar et al. (2024), fewer than 15% of published CHD AI studies evaluate their models on independent datasets. Without such validation, performance metrics may reflect overfitting to local data characteristics rather than genuine predictive ability. This problem is compounded by inconsistent reporting practices—some studies emphasise accuracy, while others report sensitivity, specificity, or F1-score—making cross-study comparison difficult (Skivington et al., 2021).

Collectively, these methodological constraints explain why many promising algorithms remain confined to research settings rather than clinical adoption. Overcoming these barriers will require collaborative, multicentre data collection, standardised acquisition protocols, transparent reporting, and rigorous external validation. It will also depend on the adoption of interpretable algorithms capable of expressing uncertainty—an area where Bayesian methods hold distinct advantages. By integrating clinical priors, accommodating missing or noisy inputs, and producing probabilistic outputs that align with clinical reasoning, Bayesian frameworks can address several limitations inherent in current machine-learning approaches while enhancing credibility and real-world applicability in paediatric cardiology.

3.5 Interpretability and Bayesian Modelling

Deep-learning systems, while powerful, frequently operate as “black boxes”, where the internal mechanisms that drive predictions remain difficult to interpret to clinicians. Although these models can achieve high classification accuracy, their lack of explainability limits clinical trust, reduces adoption in routine practice, and raises concerns related to regulatory approval and ethical accountability (Topol, 2019). In high-stakes settings such as paediatric cardiology, where diagnostic decisions directly influence life-saving interventions, clinicians must understand not only *what* a model predicts but *why* a prediction is made (Helman et al., 2021; Shantha Selvakumari et al., 2025).

These concerns have driven increasing interest in interpretable and probabilistic frameworks that explicitly represent uncertainty and relationships between variables. Bayesian approaches are particularly well aligned with medical reasoning because they formalise diagnosis as a process of probabilistic inference: prior knowledge is combined with observed evidence to update the likelihood of competing hypotheses (Arora et al., 2019; Kyrimi et al., 2021; Pearl, 2014). This mirrors the way clinicians integrate symptoms, examination findings, and test results during clinical decision-making.

3.5.1 Probabilistic reasoning in medical diagnostics

Bayesian inference provides a natural mathematical structure for diagnostic reasoning. In this framework, each clinical variable, such as murmur timing, pitch, oxygen saturation, or heart rate, is represented as a node within a directed acyclic graph, while edges encode conditional dependencies that reflect physiological or causal relationships. When new evidence becomes available, posterior probabilities are updated, yielding interpretable estimates of disease likelihood rather than binary predictions (Korb & Nicholson, 2011).

Early applications of probabilistic classifiers in cardiology, including Naïve Bayes models for electrocardiographic interpretation, demonstrated that competitive performance can be achieved with

relatively small datasets. Importantly, these models retain transparency, allowing users to understand how different variables contribute to predictions. In contrast to many machine-learning approaches, Bayesian models can also accommodate missing or uncertain inputs, which is particularly relevant in real-world paediatric settings where data may be incomplete or inconsistently recorded (Arora et al., 2019; Kyrimi et al., 2021).

3.5.2 Bayesian networks as interpretable decision-support tools

Bayesian networks extend these principles by combining graphical structure with conditional probability tables to represent both domain knowledge and empirical data (Korb & Nicholson, 2011). Their graphical nature makes relationships between variables explicit and traceable: clinicians can inspect how specific combinations of findings influence the final probability of disease (Kyrimi et al., 2021). For example, the combination of a systolic murmur, specific acoustic characteristics, and reduced oxygen saturation may increase the probability of a structural cardiac defect. Such relationships can be directly inspected, interpreted, and validated against clinical knowledge, which is not possible with most black-box models.
This level of transparency contrasts with deep neural networks, whose internal feature transformations are rarely understandable to end users. The ability to explain predictions is increasingly recognised as essential for safe clinical AI deployment (Topol, 2019). In addition, Bayesian networks typically require fewer training examples and can incorporate expert knowledge through prior probabilities or structural constraints, making them particularly suitable for small or imbalanced datasets and for low- and middle-income settings where large, annotated datasets are unavailable (Cheng et al., 2024).

3.5.3 Bridging machine learning and Bayesian reasoning

Recent research has explored hybrid approaches that integrate machine-learning feature extraction with Bayesian inference. In such architectures, deep or signal-processing methods first derive informative representations from high-dimensional data (e.g., phonocardiograms), which are subsequently incorporated into probabilistic graphical models to enable interpretable reasoning. This combination allows the model to retain the predictive power of modern machine learning while preserving the transparency, uncertainty quantification, and clinical plausibility required for decision support (Arora et al., 2019; Shantha Selvakumari et al., 2025).
This framework enables clinicians to understand, verify, and trust model predictions, thereby bridging the gap between advanced computational methods and actionable clinical decision-making. In the context of congenital heart disease, where heterogeneity, missing data, and resource constraints are common, Bayesian modelling offers a particularly appropriate balance between performance, interpretability, and practicality. For these reasons, Bayesian networks form the methodological foundation of the present dissertation.

3.6 Synthesis and Methodological Motivation

Over six decades after the first successful paediatric cardiac surgeries, timely and accurate diagnosis of congenital heart disease remains a global challenge. Although high-income countries benefit from

sophisticated imaging technologies and specialist expertise, many regions continue to experience delayed detection due to limited resources, workforce shortages, and inequitable access to care. As a result, conditions that are potentially treatable may remain undiagnosed until clinical deterioration occurs.

Recent advances in artificial intelligence and machine learning have expanded the possibilities for automated and computer-assisted screening. Across modalities ranging from echocardiography to phonocardiography, increasingly sophisticated algorithms have demonstrated impressive classification performance. However, these systems remain constrained by recurring limitations, including small and heterogeneous datasets, class imbalance, lack of external validation, and limited interpretability. Consequently, despite strong technical results, translation into routine clinical practice remains limited.

In this context, probabilistic and interpretable approaches offer an important alternative. Bayesian networks combine domain knowledge with data-driven learning while explicitly representing uncertainty and conditional dependencies. Their ability to accommodate missing information, operate effectively with moderate sample sizes, and provide transparent reasoning makes them particularly well suited to paediatric cardiology and resource-limited environments.

Taken together, the global burden of congenital heart disease, the persistent diagnostic inequities between high- and low-resource settings, and the methodological limitations of current black-box machine-learning approaches highlight the need for accessible, interpretable, and clinically trustworthy decision-support tools. Approaches that rely solely on advanced imaging or large, highly curated datasets are often impractical in real-world paediatric contexts, where data may be incomplete, heterogeneous, or noisy.

In response to these challenges, this dissertation adopts a probabilistic modelling strategy based on Bayesian networks, integrating routinely collected clinical examination findings and digital auscultation variables with cardiology and echocardiography findings used to define the reference outcomes. By explicitly representing conditional dependencies between signs, symptoms, and outcomes, this framework supports transparent probabilistic reasoning and produces risk estimates that may inform referral and triage decisions.

The following chapter describes the study design, datasets, variable processing, and Bayesian modelling procedures used to develop and internally evaluate these probabilistic models.

4 Methods

4.1 Study Design and Modelling Framework

This study was designed as a retrospective observational modelling study aimed at developing and evaluating interpretable Bayesian network-based models to support paediatric congenital heart disease (CHD) screening using routinely collected clinical and auscultation data.

The methodological process included outcome definition, predictor selection, variable transformation, model construction, and performance evaluation. Particular emphasis was placed on ensuring clinical plausibility, avoiding information leakage, and maintaining transparency and interpretability throughout the modelling process.

Bayesian networks were selected as the primary modelling approach due to their ability to represent probabilistic dependencies between heterogeneous clinical variables, explicitly model uncertainty, and perform inference in the presence of incomplete data. These characteristics make them well suited to paediatric screening contexts, where clinical information is often incomplete, heterogeneous, and developmentally dependent.

Model development followed an iterative process. An initial general Bayesian network model was constructed to assess baseline feasibility and identify methodological challenges related to missing data, outcome definition, and predictor configuration. Based on these findings, subsequent models were refined through improved outcome definition, optimisation of predictor selection, and exploration of alternative modelling strategies. These included classification-oriented Bayesian network structures, such as Naïve Bayes and Tree-Augmented Naïve Bayes, as well as alternative strategies for representing and retaining incomplete predictor information.

All models were developed using the same underlying dataset and evaluated against composite reference outcomes derived from cardiology and echocardiography diagnostic findings using the rule-based procedure described in Section 4.3.

4.2 Data Sources and Study Population

The dataset used in this study was derived from paediatric subjects evaluated within the Caravanas screening programme, a large-scale cardiovascular initiative aimed at identifying cardiac abnormalities in children and adolescents. Assessments were based on clinical examination, physiological measurements, and auscultation, reflecting real-world screening conditions where patients may be seen either through routine evaluations or symptom-driven referrals (Oliveira et al., 2022).

The initial consolidated dataset included 943 subjects and 106 variables, covering demographic information, clinical findings, physiological measurements, auscultation data, and cardiology and

echocardiography diagnostic records. Each subject had been assigned a unique identifier, which allowed information from different sources to be linked, including demographic data, clinical findings, physiological measurements, and cardiology and echocardiography records. A structured data cleaning and cohort selection process was then carried out to align the dataset with the study objectives.

Foetal assessments were excluded (n = 35), as the study focuses on postnatal screening. One subject was removed due to major inconsistencies in the recorded data that could not be resolved. Subjects aged 18 years or older at the time of screening were also excluded (n = 40) to maintain a paediatric cohort. Duplicate records were identified through unique subject identifiers and removed (n = 2), ensuring that each subject was represented only once. After these steps, the final dataset consisted of 865 unique paediatric subjects.

To complement the clinical dataset, expert-reviewed auscultation variables from the PhysioNet CirCor DigiScope dataset (Oliveira et al., 2022) were incorporated. These included expert-reviewed auscultation classifications and structured annotations, which were matched to the corresponding clinical records.

Subjects were not excluded from the cleaned study cohort solely because of missing predictor data. However, model-specific complete-case subsets were subsequently defined according to the predictors required for each analysis, as described in section 4.5.

Overall, the resulting dataset represents a heterogeneous paediatric screening population, with a wide range of ages, clinical indications, and physiological profiles, consistent with real-world screening and decision-support scenarios.

4.3 Outcome Definition (Reference Standard)

The primary objective of this study was to develop Bayesian network models capable of identifying and stratifying CHD in paediatric patients. Accordingly, outcome variables were defined to reflect both the presence and severity of congenital structural heart disease, based on cardiology and echocardiography diagnostic findings.

Diagnostic information in the source dataset was recorded as unstructured text across multiple cardiology and echocardiography fields. To enable structured modelling, the available diagnostic fields from each source were consolidated into separate cardiology and echocardiography text variables. These source-specific records were then processed independently using predefined keyword-based rules informed by paediatric cardiology terminology.

Each source was classified as indicating congenital heart disease, another non-congenital cardiac abnormality, normal cardiac findings, or unavailable diagnostic information. Congenital heart disease included both simple and complex structural defects, while the Other_Cardiac category captured non-congenital conditions such as arrhythmias, acquired cardiac disease, cardiomyopathies, and other functional or valvular abnormalities.

The first outcome variable, CHD_Binary, represented the presence or absence of congenital heart disease. Participants were classified as positive when either the cardiology or echocardiography source indicated CHD. They were classified as negative when neither source indicated CHD and at least one source contained usable diagnostic information. The outcome was classified as missing only when diagnostic information was unavailable from both sources.

The second outcome variable, CHD_4Class, provided a more detailed classification according to a predefined hierarchical procedure. Participants were classified as Complex_CHD when either source indicated CHD and the corresponding diagnostic text contained terminology associated with a complex congenital defect. In the absence of complex CHD, participants were classified as Simple_CHD when either source indicated congenital heart disease. Other_Cardiac was assigned when neither source indicated CHD but at least one source indicated a non-congenital cardiac abnormality. Normal was assigned when both sources were normal or when one source was normal and the other contained no diagnostic information. The outcome was classified as missing only when diagnostic information was unavailable from both sources.

Cardiology and echocardiography therefore contributed jointly to the construction of composite reference outcomes. Although echocardiography is clinically regarded as the gold standard for confirming structural congenital heart disease, it was not assigned formal precedence over cardiology findings within the implemented classification rule.

Outcome completeness was assessed after classification. Of the 865 subjects in the final dataset, 701 (81%) had sufficient information to define the outcome variables, while 164 (19%) were excluded from supervised modelling because the outcome could not be defined.

Using both binary and multi-class outcomes allowed evaluation of different modelling strategies, supporting both simplified disease detection and more detailed diagnostic stratification depending on model performance and clinical relevance.

4.4 Predictor Selection and Variable Preprocessing

Predictor selection and preprocessing were carried out systematically to ensure clinical relevance, methodological consistency, and compatibility with discrete Bayesian network modelling. All candidate variables were reviewed in terms of clinical plausibility, interpretability, completeness, and risk of information leakage. Particular attention was given to retaining variables that reflect information available at the time of screening, while excluding those representing downstream diagnostic conclusions or administrative artefacts.

Selection was guided by the objective of modelling congenital heart disease risk using data obtainable during routine clinical assessment. Variables derived from cardiology reports, echocardiography, or other post-diagnostic procedures were excluded to avoid incorporating information not available at the point of screening. Administrative variables without clinical relevance, such as internal identifiers, were also excluded from modelling, although retained for data linkage.

Demographic and contextual variables were included to provide clinically relevant background information. Gender was retained as a predictor, while chronological age was transformed into the Munich Age Classification System (*MACS_Age_Class*), an ordinal variable representing developmental stage. This transformation allowed age-dependent physiological variables to be interpreted within an appropriate clinical context, while avoiding redundancy between continuous age and derived measures.

Anthropometric variables were standardised using age-adjusted growth references. Raw height, weight, and body mass index were replaced by *Weight_Age_Z_Class* and *Height_Age_Z_Class*, derived using World Health Organization growth standards. Continuous Z-scores were categorised into Low, Normal, and High

classes to facilitate discrete modelling and improve interpretability. This approach allowed growth abnormalities to be represented independently of age.

Physiological variables were also transformed into clinically meaningful categories. Heart rate, systolic blood pressure, and diastolic blood pressure were classified as Low, Normal, or High based on age-specific reference ranges aligned with MACS stages and paediatric guidelines. Preductal oxygen saturation was categorised into Extremely Low, Low, and Normal using clinically relevant thresholds, and body temperature was categorised using standard paediatric cut-offs. Raw continuous values were excluded after transformation to avoid redundancy and ensure compatibility with the modelling framework.

Clinical examination findings were retained and recoded where necessary to ensure consistent categorical representation. Variables describing general condition, perfusion, respiration, peripheral pulses, and auscultation findings were structured to reflect physiologically meaningful groupings. Where variables were only applicable under specific conditions, structural missingness was explicitly represented using categories such as *Not Applicable*, rather than treated as random missing data. This allowed the underlying clinical meaning to be preserved.

Some variables were further simplified into binary indicators to improve robustness and reduce sparsity. Clinical indication variables were combined into a composite indicator (*Indication_Any*), defined as "Yes" if any individual indication (e.g., murmur, cyanosis, syncope, dyspnoea, genetic syndrome, family history, chest pain, fatigue, or rhythm disturbance) was present, and "No" otherwise. Missing values were retained when all underlying variables were missing. Selected examination variables were also dichotomised into Normal versus Abnormal categories, including *Perfusion_binary*, *Peripheral_Pulses_binary*, *Pulmonary_Auscultation_binary*, and *Cardiac_Auscultation_binary*. In all cases, missing values were preserved.

Cardiac auscultation variables, including murmur presence and characteristics, were retained given their central role in CHD screening. These variables were categorised to maintain clinical interpretability while preserving uncertainty where appropriate.

Variables were excluded if they showed excessive missingness, were redundant with derived variables, lacked relevance to CHD screening, or represented post-diagnostic information that could introduce information leakage.

Structural missingness was recoded into explicit contextual categories where clinically appropriate, such as *Not Applicable* or *Unknown*. Sporadic missing values were retained as NA in the cleaned dataset. For structure learning and internal evaluation, model-specific complete-case subsets were subsequently defined according to the predictors included in each model. Following these steps, the cleaned modelling dataset contained discretised and clinically interpretable predictors suitable for model development.

A full description of all retained variables, including transformation procedures and completeness, is provided in Appendix A.

4.5 Missing Data Handling

Missing data were present across several predictor variables, reflecting the real-world clinical screening context in which not all physiological measurements or examination findings are available for every

subject. Missingness was assessed during preprocessing to understand its extent, clinical plausibility, and potential impact on model development.

Different patterns of missingness were observed. Some variables exhibited structural missingness, arising when a variable was only recorded under specific clinical conditions. For example, detailed murmur characteristics were only documented when a murmur was present. In these cases, missing values represented the absence of a clinically applicable finding rather than incomplete data collection and were therefore preserved and explicitly represented within the dataset.

Other variables, particularly physiological measurements such as blood pressure and heart rate, showed sporadic missingness consistent with incomplete data collection during screening encounters. These missing values were retained in the cleaned source dataset rather than prompting global exclusion of the corresponding subjects. Model-specific complete-case restrictions were subsequently applied according to the predictor configuration evaluated.

An initial exploratory assessment demonstrated that restricting analyses to complete cases substantially reduced the available sample size, particularly when predictors with greater missingness were included. Nevertheless, because structure learning and comparative model evaluation required fully observed values for the predictors included in each model, model-specific complete-case subsets were used. Consequently, the effective sample size varied according to the predictor configuration evaluated.

Predictor values were not imputed. Multiple imputation was not performed because several variables exhibited structural missingness, such as murmur descriptors that were only applicable when a murmur was present, and imputing these values could have introduced clinically implausible information or artificial dependencies into the learned network structures. Although Bayesian networks can perform probabilistic inference when some evidence is unobserved, the internal performance estimates reported in this study were derived from the corresponding model-specific complete-case subsets.

Outcome variables were not imputed. Subjects with missing outcome classification were excluded from supervised model training and evaluation, as reliable outcome definition was required for valid model development. Overall, the adopted strategy prioritised retention of clinically meaningful information, minimised unnecessary data exclusion, and aligned with the probabilistic framework of Bayesian network modelling.

4.6 Bayesian Network Model Development

Bayesian network models were developed to represent probabilistic relationships between clinical predictors and congenital heart disease outcomes. These models take the form of directed acyclic graphs, where nodes correspond to variables and edges represent conditional dependencies, allowing explicit modelling of relationships between predictors and outcomes.

All statistical analyses were conducted using R Version 4.6.0 (R Core Team, 2026). Bayesian network structure learning, parameter learning, and probabilistic inference were performed primarily using the *bnlearn* package (Scutari, 2010).

Network structures were learned from the data using a score-based approach. The hill-climbing algorithm was applied to identify the structure that maximised the Bayesian Dirichlet equivalent uniform (BDeu) score (Scutari, 2010). This algorithm iteratively explores candidate network structures through edge

addition, removal, and reversal, selecting the configuration that best fits the data according to the chosen scoring criterion.

In addition to general Bayesian network models, classification-oriented structures were also explored. These included Naïve Bayes and Tree-Augmented Naïve Bayes (TAN) models, which impose structural constraints designed to prioritise predictive performance while retaining probabilistic interpretability.

Model development was carried out iteratively, using different predictor configurations to assess their impact on both network structure and predictive performance. All models were derived from the same cleaned source cohort, with model-specific complete-case subsets defined according to the predictor configuration evaluated.

4.7 Model Evaluation

Model performance was evaluated in terms of both structural robustness and predictive capability, reflecting the dual role of Bayesian networks as interpretable probabilistic models and classification tools.

Structural robustness of the general Bayesian network models was assessed using bootstrap resampling with 500 bootstrap samples. A network structure was learned independently from each resampled dataset. Arc strength was defined as the proportion of bootstrap networks in which a given edge was present, while direction strength represented the proportion of those networks in which the edge appeared in the same direction. For the binary general Bayesian network models, arcs with a bootstrap strength of at least 0.70 were retained in the consensus networks presented in the Results. This approach allowed stable and reproducible conditional dependencies to be identified while reducing the influence of sampling variability.

Predictive performance was evaluated using stratified cross-validation. Ten-fold cross-validation was used for BN-V1, BN-V3, NB-V1, TAN-V1, and the four-class models. Owing to the smaller complete-case sample and model-stability considerations, five-fold cross-validation was used for BN-V2. At each iteration, one fold was used for testing and the remaining folds for model training, until all observations had been used for validation. Predictions obtained for the held-out observations were combined to estimate model performance on unseen data.

For the binary classification models, predictive performance was quantified using the area under the receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, precision, and F1-score. Performance was first assessed using the conventional probability threshold of 0.5. To evaluate the models under operating conditions more consistent with the objectives of CHD screening, model-specific referral-oriented thresholds were additionally selected to achieve a predefined sensitivity target of at least 85%. At these thresholds, the corresponding specificity, accuracy, precision, F1-score, and referral rate were calculated to characterise the trade-off between case detection and referral burden.

Performance of the four-class models was evaluated using confusion matrices, overall accuracy, and class-specific recall, allowing assessment of the models' ability to distinguish between Normal, Simple_CHD, Complex_CHD, and Other_Cardiac categories.

For models evaluated using the same predictor set and complete-case cohort, pairwise comparisons of receiver operating characteristic curves were additionally performed using the DeLong test to assess whether observed differences in AUC were statistically significant.

All evaluation procedures used the same outcome definitions and analytic datasets described in the preceding sections, ensuring consistency across modelling approaches.

5 Results

5.1 Study Population and Outcome Distribution

A total of 865 paediatric subjects met the inclusion criteria and were included in the study cohort. Key demographic and clinically relevant screening variables are summarised in Table 5.1 to characterise the study population. Variables presented include core demographic characteristics (age and sex), a central clinical screening feature (presence of murmur), and outcome distribution.

Table 5.1 – Study Population Characteristics

Characteristic	N (%)
Total subjects	865
Sex	
Male	450 (52.0%)
Female	415 (48.0%)
Age group (MACS classification)	
Neonate	6 (0.7%)
Infant	125 (14.5%)
Toddler	135 (15.6%)
Early Childhood	166 (19.2%)
Late Childhood	362 (41.8%)
Adolescent	71 (8.2%)
Clinical murmur	
Present	378 (43.7%)
Absent	471 (54.5%)
Missing	16 (1.8%)
Outcome availability	
Outcome available	701 (81.0%)
Outcome missing	164 (19.0%)
CHD_Binary (n = 701)	
CHD present	300 (42.8%)
CHD absent	401 (57.2%)
CHD_4Class (n = 701)	

Normal	363 (51.8%)
Simple_CHD	239 (34.1%)
Complex_CHD	61 (8.7%)
Other_Cardiac	38 (5.4%)

The cohort showed a balanced sex distribution, with 450 males (52.0%) and 415 females (48.0%). Participants were distributed across all paediatric developmental stages represented by the Munich Age Classification System, with the largest proportion in the Late Childhood group (41.8%), followed by Early Childhood (19.2%) and Toddler (15.6%).

Clinical murmur was present in 378 subjects (43.7%), absent in 471 (54.5%), and not recorded in 16 (1.8%). Outcome classification was available for 701 subjects (81.0%), who constituted the supervised modelling population. Among these, 300 subjects (42.8%) were classified as having congenital heart disease and 401 (57.2%) as not having congenital heart disease. Using the four-class outcome definition, the majority of subjects were classified as Normal (51.8%) or Simple_CHD (34.1%), with smaller proportions of Complex_CHD (8.7%) and Other_Cardiac (5.4%).

5.2 Missing Data Analysis

Missing data were present across several predictor variables, reflecting the real-world nature of the screening setting, where physiological measurements and examination findings were not always recorded systematically. The extent of incomplete data varied considerably between variables and had a direct impact on the number of subjects available for model development.

Outcome classification was unavailable for 164 subjects (19.0%), leaving 701 subjects (81.0%) eligible for supervised modelling. This represented the maximum sample size prior to considering predictor completeness.

Most demographic and core clinical examination variables showed high levels of completeness. In contrast, several physiological measurements—particularly blood pressure variables—contained substantially higher proportions of missing data. This pattern was consistent with variability in measurement availability during routine screening encounters.

The effect of incomplete predictor data became more apparent during exploratory complete-case analyses. Restricting modelling to subjects with complete information for all selected predictors resulted in a marked reduction in effective sample size, especially in models incorporating variables with higher rates of missingness, such as blood pressure measurements.

In addition to sporadic missing data, structural missingness was observed in some variables. Detailed auscultation characteristics, for example, were only recorded when a murmur was present. In these situations, the absence of data reflected the absence of a clinically relevant finding rather than incomplete collection.

A detailed summary of predictor completeness and missing data patterns is provided in Appendix A.

5.3 Model BN-V1: Core Clinical Predictor Model

The first Bayesian Network model (BN-V1) was developed using a core set of clinical, physiological, and demographic predictors available during routine screening, including murmur presence, second heart sound characteristics, precordial examination findings, skin colour, oxygen saturation classification, heart rate classification, anthropometric Z-score classifications, age group, sex, and referral indication.

Following restriction to complete cases across all selected predictors, the modelling dataset consisted of 572 subjects (81.6% of the 701 subjects with available outcome classification), including 240 subjects with congenital heart disease (42.0%) and 332 without (58.0%).

The consensus network structure remained stable across bootstrap resampling iterations (R = 500). The strongest association was observed between Clinical_Murmur and CHD_Binary, with an arc strength of 1.00. Moderately stable associations were also identified between CHD_Binary and clinical examination variables, including skin colour and second heart sound characteristics. Strong associations between anthropometric variables, particularly weight-for-age and height-for-age classifications, were also identified. The consensus network structure derived from bootstrap resampling is shown in Figure 5.1.

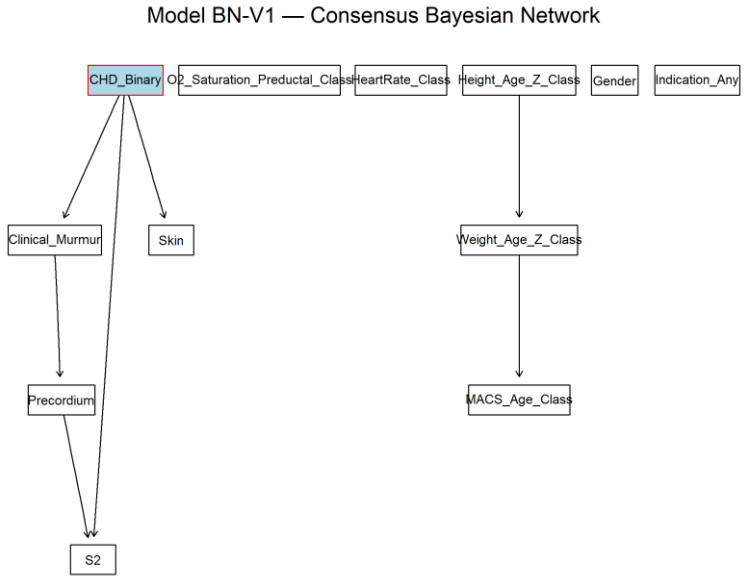


Figure 5.1 – Consensus Bayesian Network structure for Model BN-V1 derived using bootstrap resampling (R = 500) and arc strength threshold of 0.7

Predictive performance was evaluated using stratified 10-fold cross-validation. Valid probability predictions were obtained for 570 of 572 subjects (coverage: 99.6%). The model achieved an area under the receiver operating characteristic curve (AUC) of 0.772. The ROC curve is presented in Figure 5.2.

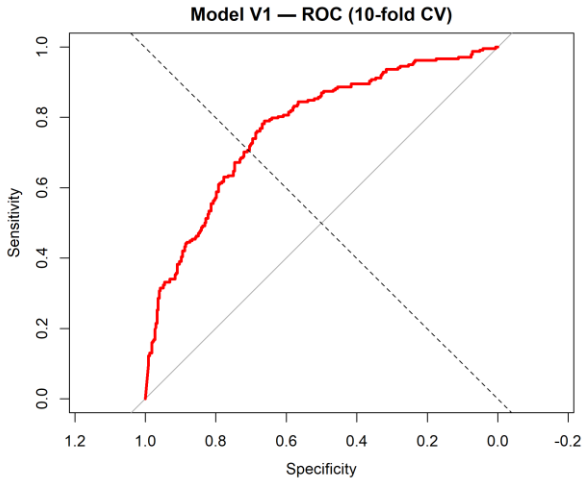


Figure 5.2 – Receiver operating characteristic curve for Model BN-V1 (AUC = 0.772)

Classification performance metrics are summarised in Table 5.2.

Table 5.2 – Predictive performance of Model BN-V1 at standard and referral-optimised probability thresholds

Metric	Threshold 0.5	Referral threshold (Sensitivity ≥ 85%)
Sample size (predicted)	570	570
AUC	0.772	—
Accuracy	70.5%	66.0%
Sensitivity	63.4%	85.3%
Specificity	75.6%	52.1%
Precision	65.1%	56.1%
F1 score	64.3%	67.7%
Referral rate	40.7%	63.5%

At the conventional probability threshold of 0.5, the model achieved an overall accuracy of 70.5%, with sensitivity of 63.4% and specificity of 75.6%.

Because the model was intended to support screening, a higher-sensitivity operating point was also evaluated. Using a referral-oriented operating point targeting sensitivity ≥85%, a probability threshold of 0.18 produced sensitivity of 85.3%, reducing the number of missed congenital heart disease cases from 87 to 35. This improvement in sensitivity was accompanied by a reduction in specificity to 52.1% and an increase in referral rate from 40.7% to 63.5%.

5.4 Model BN-V2: Expanded Predictor Model Including Blood Pressure

The second Bayesian Network model (BN-V2) extended the core predictor set by incorporating systolic and diastolic blood pressure classifications (SBP_Class and DBP_Class), in addition to the clinical and demographic variables used in Model BN-V1.

Following restriction to complete cases across all selected predictors, the modelling dataset consisted of 460 subjects (65.6% of the 701 subjects with available outcome classification), including 191 subjects with congenital heart disease (41.5%) and 269 without (58.5%).

The learned consensus Bayesian Network structure demonstrated high robustness based on bootstrap resampling (R = 500). The strongest association was observed between Clinical_Murmur and CHD_Binary (arc strength = 1.00). A strong relationship was also observed between SBP_Class and DBP_Class (arc strength = 1.00). The network also retained associations involving Skin and S2, while anthropometric variables remained clustered around age classification.

The consensus network structure derived from bootstrap resampling is shown in Figure 5.3.

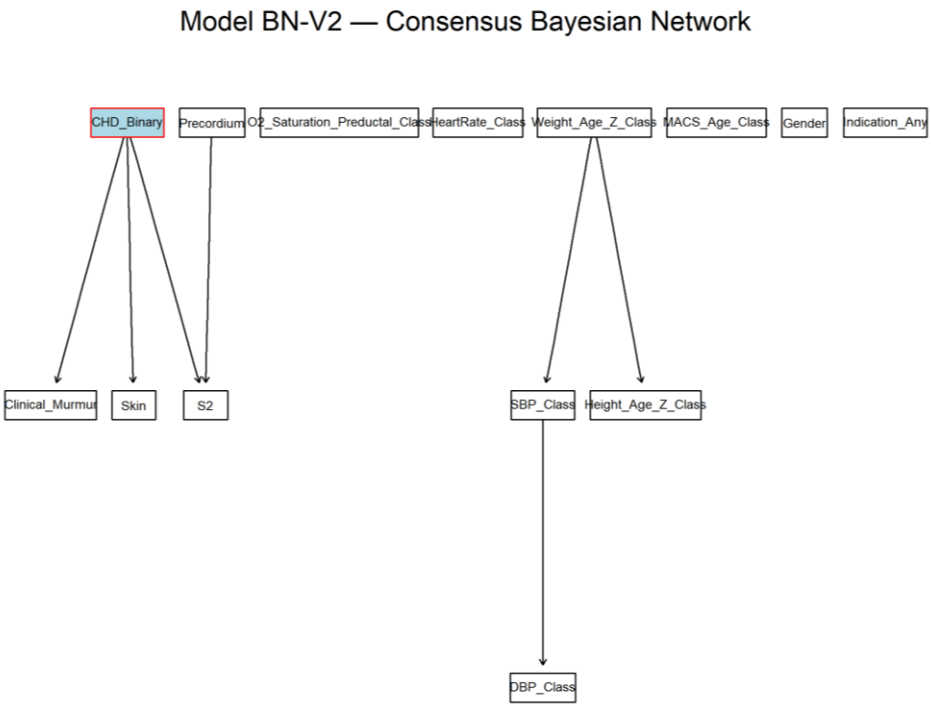


Figure 5.3 – Consensus Bayesian Network structure for Model BN-V2 derived using bootstrap resampling (R = 500) and arc strength threshold of 0.7

Predictive performance was evaluated using stratified cross-validation. Valid probability predictions were obtained for 450 of 460 subjects (coverage: 97.8%). The model achieved an area under the receiver operating characteristic curve (AUC) of 0.759. The ROC curve is shown in Figure 5.4.

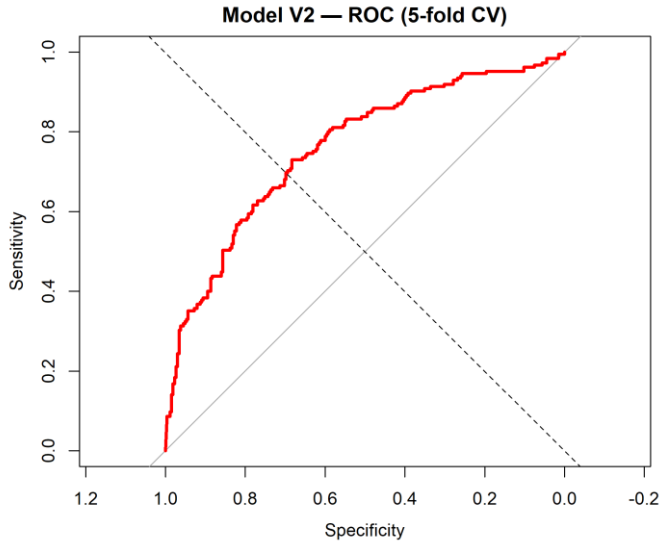


Figure 5.4 – Receiver operating characteristic curve for Model BN-V2 (AUC = 0.759)

Classification performance metrics are summarised in Table 5.3.

Table 5.3 – Predictive performance of Model BN-V2 at standard and referral-optimised probability thresholds

Metric	Threshold 0.5	Referral threshold (Sensitivity ≥ 85%)
Sample size (predicted)	450	450
AUC	0.759	—
Accuracy	70.9%	63.6%
Sensitivity	59.5%	85.4%
Specificity	78.9%	48.3%
Precision	66.3%	53.6%
F1 score	62.7%	65.8%
Referral rate	36.9%	65.6%

At the conventional probability threshold of 0.5, the model achieved an accuracy of 70.9%, with sensitivity of 59.5% and specificity of 78.9%.

At the referral-optimised threshold (sensitivity ≥ 85%), a probability threshold of 0.213 resulted in sensitivity of 85.4% and specificity of 48.3%, with an increase in referral rate from 36.9% to 65.6%. Despite the additional physiological information, overall discriminative performance remained similar to BN-V1 while substantially reducing the available sample size.

5.5 Model BN-V3: Murmur Characterisation Model

Model BN-V3 explored whether detailed murmur descriptors could improve CHD identification.

After restricting the dataset to complete cases across all selected predictors, the final modelling dataset consisted of 545 subjects, including 219 subjects with congenital heart disease (40.2%) and 326 subjects without (59.8%).

The learned consensus Bayesian Network structure demonstrated high robustness based on bootstrap resampling (R = 500). The network preserved the hierarchical relationship between murmur presence and murmur descriptors, with directed arcs from Clinical_Murmur to murmur characteristic variables. Among all predictors, Murmurs_grading showed the strongest and most consistent association with CHD_Binary (arc strength = 0.95; directional probability = 0.97). Additional associations were observed with variables including Precordium, Skin, S2, Height_Age_Z_Class, and Gender.

The consensus network structure is presented in Figure 5.5.

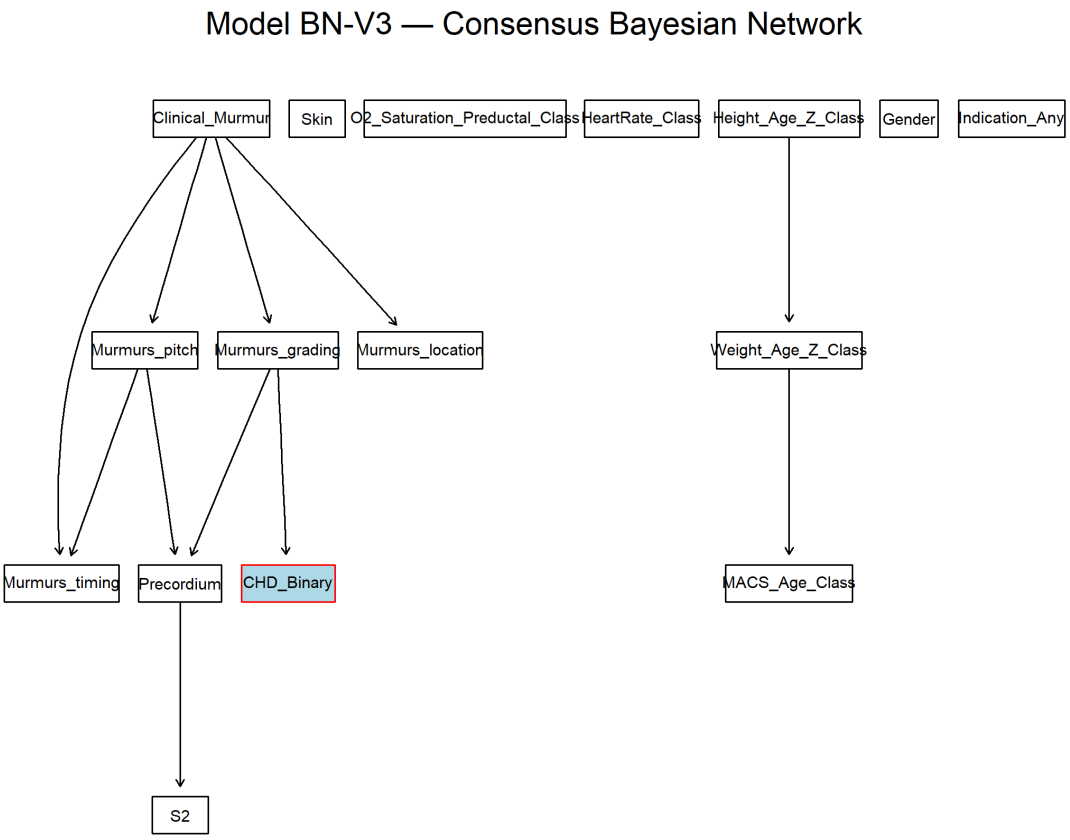


Figure 5.5 – Consensus Bayesian Network structure for Model BN-V3 derived using bootstrap resampling (R = 500) and arc strength threshold of 0.7

Predictive performance was evaluated using stratified 10-fold cross-validation. Valid probabilistic predictions were obtained for 527 of 545 subjects (coverage: 96.7%). The model achieved an area under the receiver operating characteristic curve (AUC) of 0.741. The ROC curve is presented in Figure 5.6.

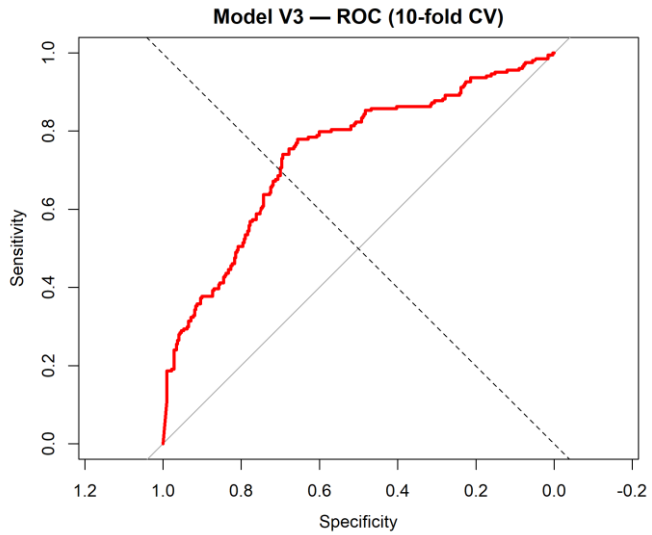


Figure 5.6 – Receiver operating characteristic curve for Model BN-V3 (AUC = 0.741)

Classification performance metrics are summarised in Table 5.4. At this threshold, the model achieved an accuracy of 69.3%, with sensitivity of 55.9% and specificity of 77.7%.

Table 5.4 – Predictive performance of Model BN-V3 at standard and referral-optimised probability thresholds

Metric	Threshold 0.5	Referral threshold (Sensitivity ≥ 85%)
Sample size (predicted)	527	527
AUC	0.741	—
Accuracy	69.3%	62.6%
Sensitivity	55.9%	85.3%
Specificity	77.7%	48.3%
Precision	61.3%	51.0%
F1 score	58.5%	63.9%
Referral rate	35.3%	65.6%

At the referral-optimised threshold (sensitivity ≥ 85%), a probability threshold of 0.205 resulted in sensitivity of 85.3% and specificity of 48.3%, with an overall accuracy of 62.6% and a referral rate of 65.6%. Although murmur grading emerged as a strong predictor within the network structure, incorporation of detailed murmur descriptors did not improve overall discriminative performance.

5.6 Analysis of Predictor Strength Across Bayesian Network Models

To further examine the relative importance of predictors across the evaluated Bayesian Network models, bootstrap arc strengths directed towards *CHD_Binary* were extracted from each model. Arc strength represents the proportion of bootstrap samples in which a given relationship appeared during structure learning and provides an indication of the stability of the association.

The strongest predictor relationships with CHD_Binary across Models BN-V1, BN-V2, and BN-V3 are presented in Figure 5.7.

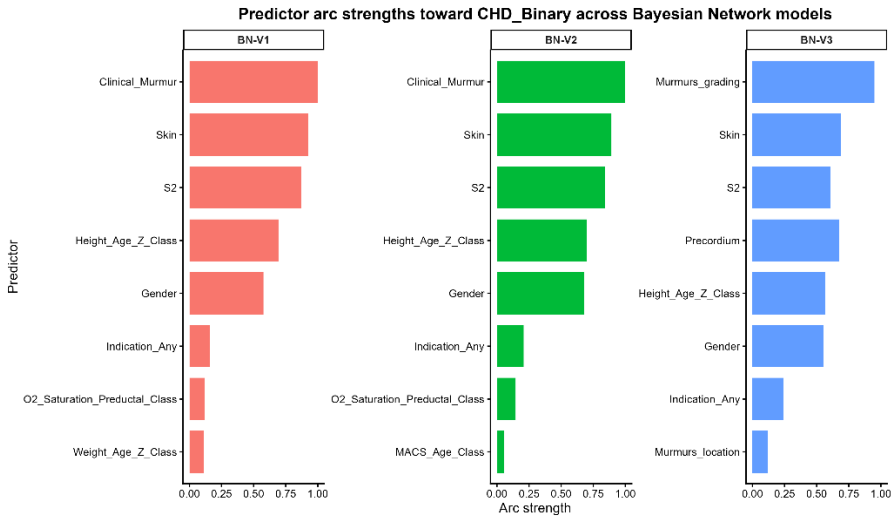


Figure 5.7 – Bootstrap arc strengths of the strongest predictors directed toward CHD_Binary across Bayesian Network models BN-V1, BN-V2, and BN-V3 with 500 bootstrap samples.

In Models BN-V1 and BN-V2, *Clinical_Murmur* consistently demonstrated the strongest association with congenital heart disease, with an arc strength of 1.00 in both models. Additional predictors with moderate arc strengths included *Skin*, *S2*, *Height_Age_Z_Class*, and *Gender*.

In Model BN-V3, which incorporated murmur descriptor variables, the strongest predictor directed toward *CHD_Binary* was *Murmurs_grading*, with an arc strength of approximately 0.95. Other variables showing moderate associations included *Precordium*, *Skin*, *S2*, *Height_Age_Z_Class*, and *Gender*.

Across all Bayesian network configurations, only a relatively small subset of predictors demonstrated consistently stable relationships with CHD, while the majority of candidate predictors showed comparatively weak or unstable relationships. Clinical murmur-related variables remained the dominant predictors across models, whereas expansion of the predictor set produced comparatively limited structural gains.

5.7 Model NB-V1: Naïve Bayes Classification Model

A Naïve Bayes classification model (NB-V1) was developed using the same predictor set as Model BN-V1 to evaluate the performance of a classification-oriented probabilistic model under the conditional independence assumption.

After restricting the dataset to complete cases across all predictors, the modelling dataset consisted of 572 subjects, including 240 with congenital heart disease (42.0%) and 332 without (58.0%), identical to the dataset used for Model BN-V1.

Predictive performance was evaluated using stratified 10-fold cross-validation, producing valid predictions for all 572 subjects. The model achieved an AUC of 0.781. The ROC curve for model NB-V1 is shown in Figure 5.8.

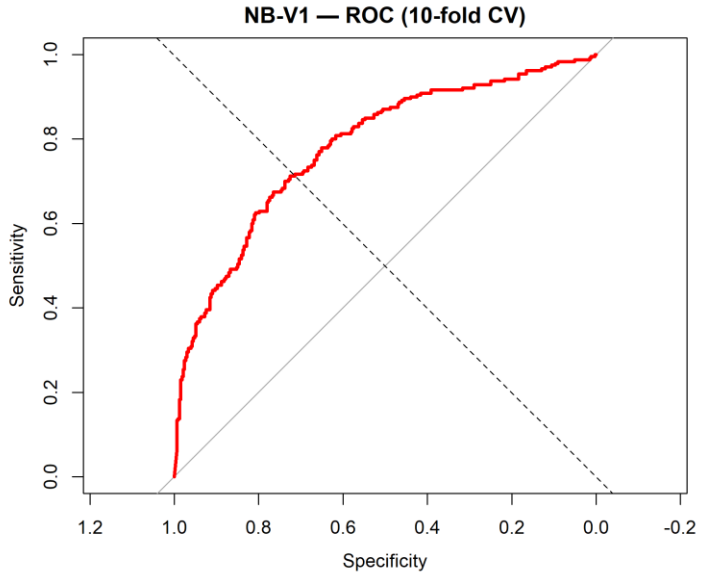


Figure 5.8 – Receiver operating characteristic curve for Model NB-V1 (AUC = 0.781)

Classification performance at the conventional probability threshold of 0.5 is summarised in Table 5.5. At this threshold, the model achieved an overall accuracy of 72.6%, with sensitivity of 62.5% and specificity of 79.8%.

Table 5.5 – Predictive performance of Model NB-V1 at standard and referral-optimised probability thresholds

Metric	**Threshold 0.5**	**Referral threshold (Sensitivity ≥ 85%)**
Sample size (predicted)	572	572
AUC	0.781	—
Accuracy	72.6%	67.5%
Sensitivity	62.5%	85.0%
Specificity	79.8%	54.8%
Precision	69.1%	57.6%
F1 score	65.6%	68.7%
Referral rate	38.0%	61.9%

At the referral-optimised threshold (sensitivity ≥ 85%), a probability threshold of 0.225 resulted in sensitivity of 85.0% and specificity of 54.8%, with a referral rate of 61.9%. Compared with BN-V1, the

Naïve Bayes approach achieved slightly improved discriminative performance despite the simplifying conditional independence assumption.

5.8 Model TAN-V1: Tree-Augmented Naïve Bayes Classification Model

A Tree-Augmented Naïve Bayes classification model (TAN-V1) was developed using the same predictor set as Models BN-V1 and NB-V1 to evaluate whether relaxing the conditional independence assumption of the Naïve Bayes classifier would improve predictive performance.

After restricting the dataset to complete cases across all predictors, the modelling dataset consisted of 572 subjects, including 240 children with congenital heart disease (42.0%) and 332 without (58.0%), corresponding to the same dataset used for Models BN-V1 and NB-V1.

The learned network structure for TAN-V1 is shown in Figure 5.9.

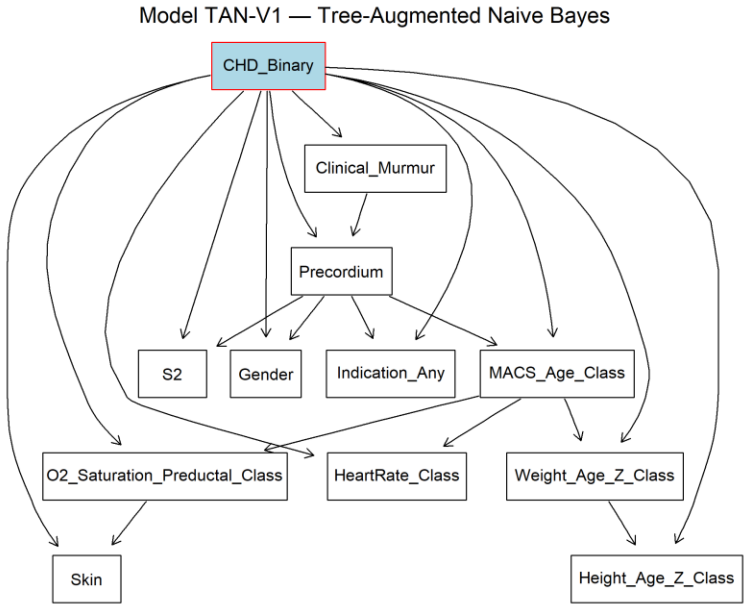


Figure 5.9 – Tree-Augmented Naïve Bayes network structure for Model TAN-V1

The learned Tree-Augmented Naïve Bayes structure maintained direct connections between CHD_Binary and all predictors, consistent with the TAN formulation. In addition, the model introduced dependencies between predictors. Notable relationships were observed among auscultatory examination variables, linking Clinical_Murmur, Precordium, and S2. MACS_Age_Class was connected to physiological and anthropometric variables, including HeartRate_Class, Weight_Age_Z_Class, and Height_Age_Z_Class. A direct relationship was also observed between Weight_Age_Z_Class and Height_Age_Z_Class.

Predictive performance was evaluated using stratified 10-fold cross-validation. Valid probability predictions were obtained for 563 of 572 subjects (coverage: 98.4%). The model achieved an area under the receiver operating characteristic curve (AUC) of 0.778. The ROC curve is shown in Figure 5.10.

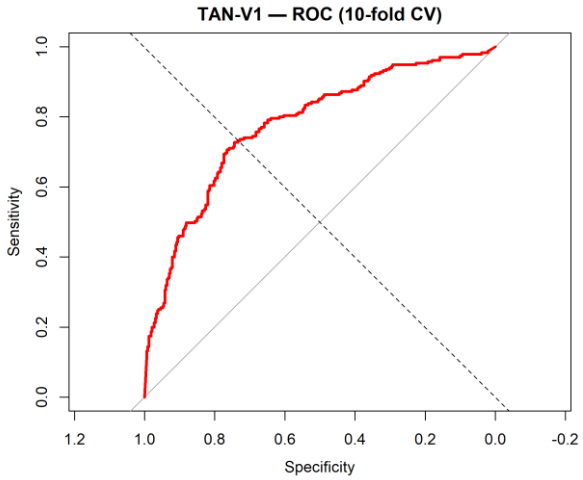


Figure 5.10 – Receiver operating characteristic curve for Model TAN-V1 (AUC = 0.778)

Classification performance at the conventional probability threshold of 0.5 is summarised in Table 5.6. At this threshold, the model achieved an accuracy of 72.5%, with sensitivity of 64.3% and specificity of 78.4%.

Table 5.6 – Predictive performance of Model TAN-V1 at standard and referral-optimised probability thresholds

Metric	Threshold 0.5	Referral threshold (Sensitivity ≥ 85%)
Sample size (predicted)	563	563
AUC	0.778	—
Accuracy	72.5%	64.8%
Sensitivity	64.3%	85.1%
Specificity	78.4%	50.3%
Precision	68.0%	55.1%
F1 score	66.1%	66.9%
Referral rate	39.4%	64.5%

At the referral-optimised threshold (sensitivity ≥ 85%), a probability threshold of 0.192 resulted in sensitivity of 85.1% and specificity of 50.3%, with a referral rate of 64.5%. Allowing dependencies between predictors produced only modest performance gains compared with the simpler Naïve Bayes classifier.

5.9 Comparative Analysis of Binary Outcome Models

The binary classification models demonstrated broadly comparable discriminative performance despite differences in predictor configuration and network structure. Comparative evaluation focused on sample size, classification performance, and behaviour under referral-oriented operating thresholds.

Model BN-V1, developed using the core clinical predictor set, achieved an AUC of 0.772 using 572 complete cases and stratified 10-fold cross-validation. Expanding the predictor set to include systolic and diastolic blood pressure classifications in BN-V2 reduced the available modelling sample to 460 subjects because of increased missingness, while overall predictive performance remained similar (AUC = 0.759). Incorporation of detailed murmur descriptors in BN-V3 also did not improve predictive performance, with the model achieving the lowest AUC among the evaluated Bayesian network configurations (AUC = 0.741). At the standard probability threshold of 0.5, all three Bayesian network models demonstrated similar sensitivity and higher specificity, reflecting a tendency toward conservative classification. BN-V1 achieved the highest sensitivity among the Bayesian network configurations (63.4%), whereas BN-V2 and BN-V3 showed slightly higher specificity.

In addition to overall discriminative performance assessed using AUC, models were compared at a predefined referral-oriented operating point targeting sensitivity ≥85%, in order to examine specificity and referral burden under a clinically conservative screening scenario. At these referral-oriented operating points, all models demonstrated the expected trade-off between sensitivity and specificity, accompanied by increased referral rates. Specificity ranged from 48.3% to 52.1% across the Bayesian Network models, while referral rates increased to approximately two-thirds of the evaluated population. Among all evaluated models, NB-V1 demonstrated the most favourable balance, maintaining sensitivity of 85.0% with the highest specificity (54.8%) and the lowest referral rate (61.9%). However, overall differences between models remained modest.

The classification-oriented probabilistic models demonstrated comparable or slightly improved performance relative to the general Bayesian network structures. NB-V1 achieved the highest AUC overall (0.781), followed closely by TAN-V1 (AUC = 0.778). TAN-V1 introduced additional dependencies between predictors, but this increased structural flexibility resulted in only modest performance differences relative to the simpler Naïve Bayes classifier.

Pairwise comparison of ROC curves using the DeLong test did not demonstrate statistically significant differences in discriminative performance between BN-V1, NB-V1, and TAN-V1 (all $p > 0.05$). Specifically, no significant differences were observed between BN-V1 and NB-V1 ($p = 0.461$), BN-V1 and TAN-V1 ($p = 0.820$), or NB-V1 and TAN-V1 ($p = 0.751$). Formal statistical comparison was limited to these models because they were evaluated using the same predictor set and complete-case cohort.

Overall, expansion of the predictor set through incorporation of blood pressure measurements or detailed murmur descriptors did not result in substantial improvement in discriminative performance. The reduction in available sample size observed in BN-V2 illustrates the practical impact of incorporating variables with higher levels of missingness into probabilistic screening models. Similarly, although murmur grading emerged as a structurally important predictor in BN-V3, incorporation of additional auscultatory detail increased model complexity without improving overall classification performance compared with the simpler core predictor model.

Similarly, the classification-oriented probabilistic models (NB-V1 and TAN-V1) achieved discriminative performance comparable to that of the general Bayesian Network models despite their simpler structural assumptions. These findings suggest that predictive performance was influenced more by the information contained within the available screening variables than by the complexity of the modelling approach. The

implications of these findings are explored further in the Discussion. A summary of the performance metrics for all evaluated binary classification models is presented in Table 5.7.

Table 5.7 – Comparative performance of Bayesian Network models for binary congenital heart disease classification

Metric	BN-V1	BN-V2	BN-V3	NB-V1	TAN-V1
Sample size (complete cases)	572	460	545	572	572
Cross-validation	10-fold	5-fold	10-fold	10-fold	10-fold
AUC	0.772	0.759	0.741	0.781	0.778
Accuracy (0.5 threshold)	70.5%	70.9%	69.3%	72.6%	72.5%
Sensitivity (0.5 threshold)	63.4%	59.5%	55.9%	62.5%	64.3%
Specificity (0.5 threshold)	75.6%	78.9%	77.7%	79.8%	78.4%
Sensitivity (≥85% target)	85.3%	85.4%	85.3%	85.0%	85.1%
Specificity (≥85% target)	52.1%	48.3%	48.3%	54.8%	50.3%
Referral rate (≥85% target)	63.5%	65.6%	65.6%	61.9%	64.5%

5.10 Models Using Four-Class Outcome (CHD_4Class)

This section presents an exploratory analysis of models developed using the four-class outcome variable (**CHD_4Class**), which distinguishes between *Normal*, *Simple congenital heart disease*, *Complex congenital heart disease*, and *Other cardiac conditions*. This formulation allowed exploratory evaluation of whether probabilistic models could distinguish between different levels of disease severity and diagnostic categories beyond simple binary CHD detection.

5.10.1 Model BN-V1-4Class: Bayesian Network with Hill-Climbing Structure Learning

A Bayesian Network model (BN-V1-4Class) was developed using the same predictor set as Model BN-V1, replacing the binary outcome with the four-class variable. The modelling dataset consisted of 572 complete cases, including 309 Normal subjects, 186 with Simple_CHD, 54 with Complex_CHD, and 23 with Other_Cardiac.

The learned network structure is shown in Figure 5.11.

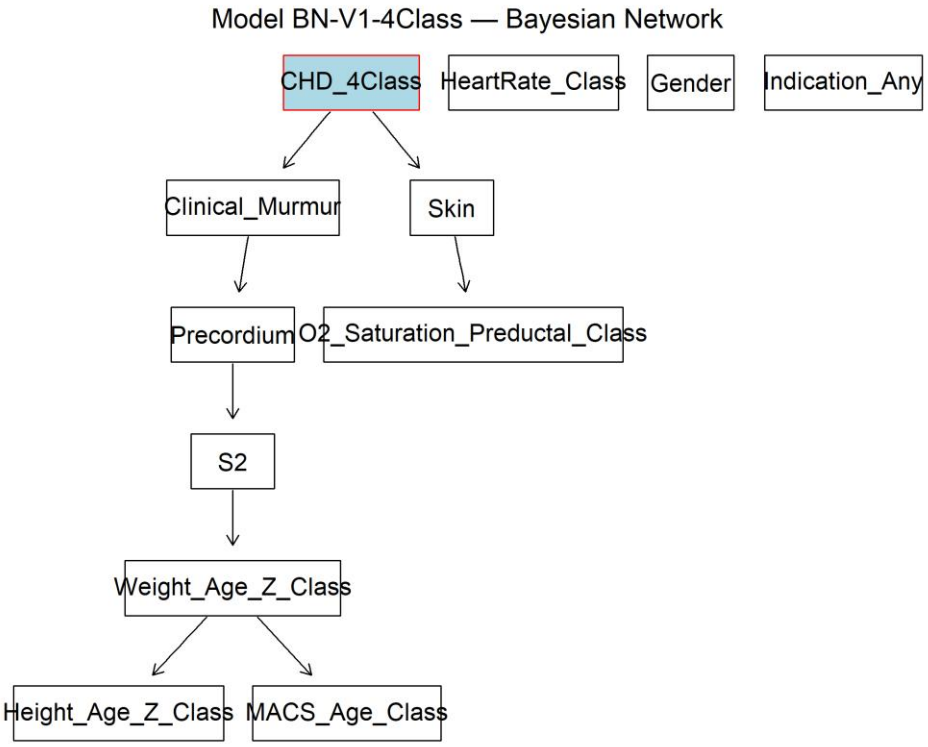


Figure 5.11 – Bayesian Network structure for Model BN-V1-4Class

The learned structure retained several relationships between clinical variables observed in the binary models, including dependencies linking Clinical_Murmur, Precordium, and S2. Associations between oxygen saturation and skin colour were also preserved. Anthropometric variables formed a coherent cluster associated with age classification, consistent with patterns observed in previous models.

Several variables, including heart rate, gender, and referral indication, showed weak or absent connections within the learned structure, suggesting a limited contribution to four-class discrimination in this dataset. Predictive performance was evaluated using stratified 10-fold cross-validation, producing valid predictions for 570 of 572 subjects (coverage: 99.6%).

Classification results are summarised in Table 5.8.

Table 5.8 – Confusion matrix for Model BN-V1-4Class

Predicted/Actual	Normal	Simple_CHD	Complex_CHD	Other_Cardiac
Normal	212	48	6	14
Simple_CHD	97	133	37	9
Complex_CHD	0	3	11	0
Other_Cardiac	0	0	0	0

Classification performance was strongest for the Normal and Simple_CHD classes, with recall values of approximately 68.6% and 72.3%, respectively. In contrast, performance for less frequent classes was

substantially lower. Recall for Complex_CHD was approximately 20%. No subjects belonging to the Other_Cardiac class were correctly classified.

5.10.2 Model NB-V1-4Class: Naïve Bayes Classification Model

A Naïve Bayes classification model (NB-V1-4Class) was developed using the same predictor set and dataset (n = 572), replacing the binary outcome with the four-class variable.

Predictive performance was evaluated using stratified 10-fold cross-validation. The resulting confusion matrix is presented in Table 5.9.

Table 5.9 – Confusion matrix for Model NB-V1-4Class

Predicted/Actual	Normal	Simple_CHD	Complex_CHD	Other_Cardiac
Normal	252	77	15	16
Simple_CHD	54	98	26	6
Complex_CHD	2	9	13	1
Other_Cardiac	1	2	0	0

Misclassification occurred predominantly between adjacent disease categories, particularly between Normal and Simple_CHD and between Simple_CHD and Complex_CHD.

Performance for less frequent classes remained poor. In particular, the Other_Cardiac class was not identified by the model, with all cases being misclassified into other categories. The model also showed limited ability to correctly distinguish Complex_CHD cases using the available predictor set.

5.10.3 Summary of Four-Class Modelling

Overall, both four-class models showed substantially lower performance than the binary classification models. Classification was more reliable for the larger and clinically broader categories, particularly Normal and Simple_CHD, whereas discrimination between less frequent classes remained limited.

Misclassification occurred predominantly between clinically related categories, especially between Simple_CHD and Complex_CHD. The Other_Cardiac category proved particularly difficult to identify, likely reflecting both the small number of available cases and the heterogeneity of conditions included within this group.

These findings suggest that the routinely collected screening variables used in this study may contain sufficient information for binary CHD detection, but provide more limited discriminatory power for finer diagnostic stratification. As a result, subsequent analyses focused primarily on the binary classification framework.

6 Discussion

6.1 Overview of the main findings

This dissertation investigated the use of interpretable probabilistic models for congenital heart disease screening in a paediatric population using clinical, physiological, and auscultation variables routinely collected during screening encounters. The principal aim was to evaluate whether Bayesian modelling approaches could support the early identification of children with congenital heart disease while maintaining clinical interpretability and applicability within real-world screening settings.

Overall, the findings indicate that routinely collected screening data contain sufficient predictive information to discriminate between children with and without congenital heart disease. Across the evaluated binary models, AUC values ranged from 0.741 to 0.781, with no statistically significant differences observed between the principal Bayesian approaches. Although the classification-oriented models (NB-V1 and TAN-V1) achieved the highest AUC values, the general Bayesian Network models offered the additional advantage of representing clinically meaningful conditional relationships between predictors, enhancing model interpretability.

A consistent finding across all binary models was the dominant role of murmur-related variables in CHD prediction. Clinical murmur emerged as the most influential predictor in BN-V1 and BN-V2, while murmur grading assumed this role in BN-V3 following the incorporation of detailed murmur descriptors. Additional clinically coherent associations were observed with second heart sound characteristics, skin colour, precordial findings, and anthropometric classifications, supporting the physiological plausibility and interpretability of the learned network structures.

Expansion of the predictor set did not improve discriminative performance. The inclusion of blood pressure variables in BN-V2 reduced the available sample size because of missing data, while the incorporation of detailed murmur descriptors in BN-V3 increased model complexity without improving classification performance. These findings suggest that a relatively small set of clinically relevant screening variables captured most of the predictive information available for binary congenital heart disease detection within the present dataset.

Comparison between the general Bayesian Network and classification-oriented Bayesian approaches highlighted the balance between predictive performance and interpretability. Although NB-V1 and TAN-V1 achieved slightly higher AUC values, these differences were not statistically significant. In contrast, the general Bayesian Network models explicitly represented clinically meaningful relationships between variables, enhancing interpretability and supporting their potential integration into clinical decision-

making, an advantage that has been recognised as a key strength of Bayesian Networks in healthcare applications (Kyrimi et al., 2021).

Evaluation of referral-oriented operating points demonstrated the potential clinical utility of these models as screening support tools. When decision thresholds were adjusted to achieve a sensitivity of approximately 85%, specificity decreased and referral rates increased across all binary models. This trade-off reflects the fundamental objective of screening, where minimising missed cases is generally prioritised over reducing false-positive referrals, recognising that some increase in false-positive referrals may be acceptable to improve early identification of affected children (Oster et al., 2025; Van Calster et al., 2026). The probabilistic nature of Bayesian models represents an important practical advantage, allowing decision thresholds to be adapted according to clinical priorities and available healthcare resources.

In contrast to the consistent performance observed for binary classification, the exploratory multi-class models demonstrated considerably greater difficulty distinguishing between specific diagnostic categories. Performance was highest for the more prevalent *Normal* and *Simple_CHD* classes, whereas classification of *Complex_CHD* and *Other_Cardiac* remained limited. These findings suggest that routinely collected screening variables are better suited to identifying children at increased risk of congenital heart disease than to distinguishing between the four diagnostic categories, which typically require more comprehensive clinical assessment and confirmatory imaging (Meng et al., 2024).

Overall, these findings support the use of Bayesian models as interpretable decision-support tools for congenital heart disease screening using routinely collected clinical data. Their greatest value lies in supporting early risk stratification and referral rather than replacing diagnostic evaluation. The following sections discuss these findings in greater depth, with particular emphasis on model performance, clinically relevant predictors, methodological considerations, and implications for future research and clinical implementation.

6.2 Interpretation of model performance for binary CHD detection

The results of this study demonstrate that Bayesian modelling approaches using routinely collected clinical, physiological, and auscultation variables can discriminate between children with and without congenital heart disease using information available during routine screening. Across all evaluated binary models, AUC values ranged from 0.741 to 0.781, with no statistically significant differences observed between the principal Bayesian approaches. These findings indicate that clinically useful risk stratification can be achieved without relying on increasingly complex model structures.

This level of performance is clinically plausible given the nature of the available predictors. Unlike echocardiography, which provides direct visualisation of cardiac structure, routine screening variables capture indirect manifestations of disease that may be subtle, age-dependent, and physiologically variable. Consequently, perfect discrimination would not be expected when relying solely on screening data, and the observed performance should therefore be interpreted in the context of a decision-support tool designed to support referral rather than establish a definitive diagnosis. This distinction is consistent with contemporary approaches to congenital heart disease screening, where the primary objective is to identify children who

require further diagnostic evaluation, while echocardiography remains the reference standard for diagnosis (Hoffman & Kaplan, 2002; Meng et al., 2024; Oster et al., 2025).

The relatively narrow range of AUC values across modelling approaches suggests that predictive performance was influenced more by the information content of the available predictors than by the specific modelling strategy. Although NB-V1 and TAN-V1 achieved slightly higher AUC values than BN-V1, these differences were small and not statistically significant. This observation is consistent with the broader clinical machine learning literature, which recognises that predictive performance depends not only on the choice of algorithm but also on the quality and clinical relevance of the available data (Arora et al., 2019; Helman et al., 2021). The comparison between BN-V1, BN-V2, and BN-V3 further supports the observation that increasing model complexity did not necessarily improve predictive performance. Neither the inclusion of systolic and diastolic blood pressure classifications in BN-V2 nor the addition of detailed murmur descriptors in BN-V3 resulted in meaningful improvements in discrimination. This suggests that the core predictor set captured most of the discriminatory information available for binary congenital heart disease detection within the present dataset (Kyrimi et al., 2021).

One possible explanation for these findings is that several of the additional predictors displayed limited variability, with most observations falling within the normal category, while their inclusion also increased missing data and reduced the effective sample size available for model training. Together, these factors likely limited the benefit of expanding the predictor set (Helman et al., 2021).

At the conventional probability threshold of 0.5, all models favoured specificity over sensitivity, reflecting a tendency to correctly classify children without congenital heart disease while missing a proportion of affected cases. Although this operating point may be appropriate for general classification tasks, it is less well suited to screening, where the clinical consequences of false negatives generally outweigh those of false-positive referrals. When decision thresholds were adjusted to achieve a referral-oriented sensitivity of approximately 85%, specificity decreased and referral rates increased across all models, illustrating the expected trade-off between maximising case detection and limiting unnecessary referrals. This flexibility represents an important advantage of probabilistic models, allowing decision thresholds to be adapted according to clinical priorities, available healthcare resources, and acceptable referral burdens. Such flexibility is consistent with current congenital heart disease screening practice, which prioritises early identification of affected children while recognising that an increase in false-positive referrals may be acceptable to improve case detection (Oster et al., 2025; Van Calster et al., 2026).

The similarity in threshold-adjusted performance across models further suggests that different Bayesian approaches converge towards comparable operating characteristics when applied to congenital heart disease screening. Consequently, model selection may depend less on small differences in discrimination and more on considerations such as interpretability, ease of implementation, and integration into clinical workflows (Kyrimi et al., 2021; Topol, 2019). Overall, these findings support the use of Bayesian models as complementary decision-support tools for early risk stratification and referral rather than as standalone diagnostic systems.

6.3 Clinical Interpretation of the Most Informative Predictors

The analysis of predictor relationships across the Bayesian Network models provides insight into the clinical variables that contributed most strongly to congenital heart disease detection in this paediatric screening population. Across all models, a consistent pattern emerged in which a relatively small number of clinically relevant predictors accounted for most of the discriminatory information, whereas many routinely collected variables contributed little to overall model performance. These findings reinforce the importance of clinically informed predictor selection when developing screening-oriented decision-support models.

6.3.1 Auscultation and Clinical Examination Predictors

Across all evaluated models, auscultation-related variables emerged as the dominant predictors of congenital heart disease. In BN-V1 and BN-V2, *Clinical_Murmur* demonstrated the strongest and most stable association with *CHD_Binary*, while *Murmur_Grading* became the principal predictor in BN-V3 following the inclusion of detailed murmur descriptors. These findings reinforce the central role of cardiac auscultation in paediatric screening and are consistent with previous work demonstrating that murmur characteristics contain the greatest discriminatory information for identifying structural heart disease (Begic et al., 2021; Gudigar et al., 2024). Similarly, Oliveira et al. (2022), using the CirCor DigiScope dataset, showed that auscultatory features extracted from heart sounds were the primary drivers of automated murmur detection and congenital heart disease screening.

Although murmur grading emerged as the strongest predictor in BN-V3, incorporating additional murmur descriptors such as timing, pitch, and location did not improve overall model performance. This suggests that murmur intensity captured most of the clinically relevant information required for screening within the present dataset, whereas more detailed murmur characteristics are likely to contribute greater value during diagnostic evaluation than initial risk stratification. These findings support the use of structured murmur assessment in screening programmes while also illustrating that increasing predictor detail does not necessarily translate into improved discrimination, consistent with previous observations that the quality of auscultatory information is often more important than increasing the complexity of extracted features (Gudigar et al., 2024).

Beyond murmur-related variables, second heart sound characteristics, skin colour, and precordial findings also demonstrated stable associations with congenital heart disease across the Bayesian Network models. These relationships are physiologically plausible, reflecting well-recognised manifestations of structural cardiac abnormalities, and their consistent identification across bootstrap resampling supports the clinical validity of the learned network structures. Together, these findings indicate that the models relied primarily on variables obtained through routine cardiovascular examination, reinforcing both the interpretability of the Bayesian Networks and their potential value as clinically coherent decision-support tools for paediatric congenital heart disease screening (Kyrimi et al., 2021).

6.3.2 Secondary Predictors and Predictor Selection

Beyond the clinical examination findings, anthropometric variables formed a consistent cluster across all Bayesian Network models, particularly weight-for-age, height-for-age, and age classification. Because these variables were derived from age-standardised WHO Z-scores, part of their association with age was expected and reflects the methodology used to construct the predictors. Nevertheless, growth impairment

is a recognised consequence of chronic or haemodynamically significant congenital heart disease (Meng et al., 2024), suggesting that anthropometric measures may provide complementary information on disease burden despite contributing less to discrimination than direct clinical examination findings.

In contrast, several routinely collected variables, including heart rate classification, gender, and referral indication, demonstrated weak or inconsistent associations with congenital heart disease and contributed little to overall model performance. These findings reinforce the importance of clinically informed predictor selection in screening-oriented models, where inclusion of variables with limited discriminatory value may increase model complexity without providing meaningful improvements in predictive performance. Similar conclusions have been reported in the Bayesian Network literature, which emphasises that clinically useful models depend on the selection of relevant and informative predictors rather than maximising the number of included variables (Kyrimi et al., 2021).

6.4 Why Expanded Predictor Sets Did Not Improve Performance

One of the principal findings of this dissertation was that expanding the predictor set did not improve binary CHD classification. Neither the inclusion of blood pressure variables (BN-V2) nor detailed murmur descriptors (BN-V3) resulted in improved discrimination compared with BN-V1. These findings suggest that, within the present dataset, increasing the number of predictors offered little benefit once the most informative clinical variables had already been included.

The inclusion of systolic and diastolic blood pressure classifications in BN-V2 was motivated by their potential physiological relevance to CHD. However, these variables did not improve model discrimination and substantially reduced the effective sample size because of their higher proportion of missing data. In addition, blood pressure abnormalities are not expected across all forms of congenital heart disease and may therefore provide limited discriminatory information in heterogeneous screening populations. Their clinical value is likely greatest in specific lesions, particularly coarctation of the aorta, where blood pressure differences constitute an important diagnostic finding, although the contribution of this variable will depend on the distribution of disease subtypes within the screened population (Gurvitz et al., 2026). Although systolic and diastolic blood pressure demonstrated the expected physiological dependency within the learned network structure, this information did not provide meaningful additional discriminatory value beyond the core clinical examination findings.

The inclusion of detailed murmur descriptor variables in BN-V3 was intended to capture more granular auscultatory information and improve classification performance. However, despite increasing model complexity, BN-V3 did not outperform BN-V1. As discussed in Section 6.3.1, murmur grading emerged as the most informative descriptor, suggesting that murmur intensity captured most of the clinically relevant information required for screening within the present dataset. In contrast, additional descriptors such as timing, pitch, and anatomical location contributed little independent discriminatory information. Although these characteristics are recognised as important for comprehensive murmur characterisation and may support more detailed diagnostic assessment, their inclusion did not improve binary screening performance in this study (Gudigar et al., 2024; Oliveira et al., 2022). This suggests that, for initial congenital heart

disease screening, murmur presence and intensity provide most of the discriminatory information, whereas more detailed murmur characterisation is likely to be of greater value during diagnostic evaluation.

Taken together, these findings demonstrate that expanding the predictor set beyond a core group of clinically relevant variables did not improve predictive performance. Instead, the additional variables introduced greater complexity while offering little incremental discriminatory value. Model BN-V1, which relied on a smaller set of robust screening variables, achieved performance comparable to or better than the more complex model configurations.

These findings reinforce the importance of prioritising clinically relevant and consistently available predictors over simply increasing the number of variables included in a model. From a practical perspective, models based on a focused set of high-value screening variables are likely to be easier to implement in routine clinical practice, where time, data completeness, and resource availability are often constrained. Consequently, these findings support the development of streamlined, interpretable Bayesian decision-support tools tailored to the realities of paediatric congenital heart disease screening, where practical implementation is as important as maximising predictive performance (Van Calster et al., 2026).

6.5 Comparison Between General Bayesian Networks, Naïve Bayes, and Tree-Augmented Naïve Bayes Models

The comparative analysis of modelling approaches in this study provides important insight into the trade-offs between predictive performance, structural flexibility, and interpretability in probabilistic models for congenital heart disease screening. Three Bayesian modelling approaches were evaluated: general Bayesian Networks (BN-V1, BN-V2, and BN-V3), the Naïve Bayes classifier (NB-V1), and the Tree-Augmented Naïve Bayes model (TAN-V1).

Across all evaluated models, discriminative performance was broadly similar, with AUC values ranging from 0.741 to 0.781. NB-V1 achieved the highest AUC, followed closely by TAN-V1 and BN-V1, although the absolute differences between models were modest. Formal comparison using the DeLong test confirmed that these differences were not statistically significant, suggesting that predictive performance was determined primarily by the information content of the available screening variables rather than by the modelling strategy itself.

The Naïve Bayes classifier operates under the assumption of conditional independence between predictors given the outcome. Despite this simplifying assumption, NB-V1 achieved strong performance, indicating that the dominant predictive signals within the dataset, particularly those related to murmur findings and clinical examination variables, were sufficiently informative to support effective classification without requiring extensive modelling of inter-variable dependencies. This observation is consistent with previous findings in clinical machine learning, where Naïve Bayes frequently demonstrates competitive performance despite its theoretical simplifications, especially in settings dominated by a small number of strong predictors (Helman et al., 2021; Kyrimi et al., 2021).

The Tree-Augmented Naïve Bayes model extends the Naïve Bayes framework by permitting limited dependencies between predictors while preserving a classification-oriented structure. In this study, TAN-V1 introduced clinically plausible relationships between variables, including associations among

auscultatory findings and between age classification and physiological variables. However, this additional structural flexibility did not produce meaningful improvement in discriminative performance compared with NB-V1. This suggests that modelling limited dependencies between predictors provided little additional discriminatory information beyond that already captured by the core screening variables.

In contrast, general Bayesian Networks permit unrestricted modelling of conditional dependencies between variables, providing a more flexible and expressive representation of the underlying clinical data structure. The general Bayesian Network models learned clinically meaningful conditional dependencies between predictors, providing an interpretable representation of the underlying data structure beyond purely classification-oriented objectives, consistent with the principal strengths of Bayesian Networks described by Kyrimi et al. (2021). However, the increased flexibility of general Bayesian Networks did not translate into superior predictive performance compared with the simpler classification-oriented approaches. This may reflect the fact that the most informative variables in the dataset were directly associated with the outcome, reducing the incremental value of modelling more complex inter-variable relationships. It may also reflect limitations related to sample size, missing data, and discretisation of variables, all of which can constrain the capacity of more flexible models to exploit subtle dependencies reliably (Van Calster et al., 2026).

The findings of this study therefore suggest that, for binary congenital heart disease classification using routine screening variables, simpler classification-oriented Bayesian models such as Naïve Bayes and TAN may achieve comparable or slightly superior predictive performance relative to more flexible Bayesian Network structures. Nevertheless, this modest performance advantage should be considered alongside the greater interpretability offered by general Bayesian Networks through their explicit representation of conditional dependencies (Kyrimi et al., 2021; Van Calster et al., 2026).

From a clinical perspective, model selection should depend on the intended application. Classification-oriented models may be preferable when predictive performance and implementation efficiency are the primary objectives, whereas general Bayesian Networks offer greater value when model interpretability and representation of clinically meaningful relationships are priorities (Kyrimi et al., 2021; Topol, 2019).

Overall, these findings demonstrate that no single Bayesian approach was universally superior for congenital heart disease screening. Instead, model selection should reflect the desired balance between predictive performance, interpretability, and clinical applicability.

6.6 Threshold Optimisation and Clinical Utility in Screening Contexts

Given the screening-oriented aim of this study, it was important to evaluate model performance under operating thresholds aligned with the priorities of paediatric CHD screening. Conventional classification models are frequently assessed using a default probability threshold of 0.5; however, this threshold does not necessarily reflect the clinical objectives of screening, where minimising missed cases is generally prioritised over maximising overall classification accuracy (Van Calster et al., 2026).

At the standard threshold of 0.5, all evaluated binary models demonstrated a similar performance profile characterised by moderate sensitivity and higher specificity. This operating point favours correct identification of children without congenital heart disease but results in a substantial proportion of false-

negative cases. In the context of CHD screening, such a trade-off may be clinically undesirable, as delayed recognition of significant cardiac abnormalities can lead to important adverse consequences, particularly in children with more severe forms of the disease (Oster et al., 2025).

To evaluate model performance under conditions more consistent with screening objectives, thresholds were adjusted to achieve a predefined sensitivity target of approximately 85%. Achieving this operating point required lowering the classification threshold, reducing false-negative classifications at the expense of lower specificity and increased referral rates. This trade-off is consistent with the objectives of screening, where identifying affected children is generally prioritised over minimising false-positive referrals (Oster et al., 2025).

Among the evaluated models, NB-V1 demonstrated the most favourable balance at the predefined high-sensitivity operating point, maintaining sensitivity close to 85% while achieving the highest specificity and the lowest referral rate. However, the differences between models remained relatively modest, suggesting that the overall referral implications were broadly similar across modelling approaches when aligned with screening-oriented objectives.

The increase in referral burden represents an important consideration for clinical implementation. Although reducing false-negative cases is desirable, higher referral rates increase healthcare resource utilisation and may contribute to unnecessary specialist assessment and parental anxiety. Consequently, the optimal operating threshold is likely to depend on the intended screening environment, balancing early detection against available resources and acceptable referral rates (Van Calster et al., 2026). This flexibility represents one of the principal advantages of probabilistic Bayesian models (Kyrimi et al., 2021).

Another important strength of Bayesian models is that they generate probabilistic outputs rather than fixed class assignments. This enables thresholds to be adapted according to clinical priorities and supports risk-based decision-making instead of rigid binary classification. In practice, such models could facilitate tiered referral strategies by distinguishing children at high, intermediate, and low risk of congenital heart disease (Topol, 2019).

Overall, the findings suggest that the clinical utility of Bayesian models lies not only in their discriminative performance but also in their ability to support flexible, risk-based decision-making through threshold optimisation. Even without perfect discrimination, models capable of reducing missed cases while maintaining acceptable referral rates may provide meaningful support for congenital heart disease screening and referral.

6.7 Interpretation of the Multi-Class Modelling Results

The exploratory multi-class analysis provides important insight into the limitations of probabilistic modelling based on routine screening variables for more detailed congenital heart disease classification. Although the binary models demonstrated consistent performance in distinguishing children with and without congenital heart disease, this performance did not extend to the four-class setting. Both BN-V1-4Class and NB-V1-4Class showed limited ability to differentiate between specific diagnostic categories, particularly the less frequent *Complex_CHD* and *Other_Cardiac* classes.

6.7.1 Challenges of Multi-Class Classification Using Routine Screening Variables

One of the clearest findings of this study is that predictive performance achieved in binary congenital heart disease detection did not translate to multi-class disease classification. This difference is clinically plausible because distinguishing between the presence and absence of congenital heart disease represents a fundamentally simpler task than differentiating between multiple diagnostic categories and levels of disease complexity with overlapping clinical characteristics. While binary classification requires the identification of general indicators of cardiovascular abnormality, multi-class classification demands more specific information capable of distinguishing between different anatomical defects and disease severities.

The reduced performance observed in the four-class models suggests that the routine clinical, physiological, and auscultation variables available in this study provided sufficient information to identify children at increased risk of congenital heart disease but were less capable of distinguishing between individual diagnostic categories. This is likely because many congenital heart disease subtypes share similar screening-level manifestations, meaning that the available predictors capture broad indicators of cardiovascular abnormality rather than subtype-specific characteristics. Similar observations have been reported in previous work using the CirCor DigiScope dataset, where heart sound recordings were used to identify abnormal murmurs and support referral, while definitive diagnosis remained dependent on specialist clinical assessment and echocardiography (Gudigar et al., 2024; Oliveira et al., 2022).

These findings reinforce the distinction between screening and diagnosis in congenital heart disease. The primary objective of screening is to identify children who require further investigation rather than to establish a definitive diagnosis or classify specific disease subtypes, which generally require comprehensive clinical assessment and imaging (Meng et al., 2024). Consequently, the lower performance of the multi-class models should not be interpreted as a limitation of the Bayesian modelling approach itself, but rather as reflecting the limited diagnostic information available from routine screening variables alone.

6.7.2 Class Imbalance and Clinical Implications

The performance of the multi-class models was further influenced by the unequal distribution of outcome categories within the analytic dataset. Most participants were classified as either Normal or Simple_CHD, whereas Complex_CHD and Other_Cardiac were comparatively underrepresented. This imbalance reduced the number of observations available to learn subtype-specific patterns and likely contributed to the lower recall and higher misclassification rates observed for the less frequent categories. Class imbalance is a recognised challenge in clinical machine learning, particularly in multi-class classification problems where rare classes contribute relatively little to model training (Helman et al., 2021; Kyrimi et al., 2021).

Nevertheless, class imbalance alone is unlikely to explain the observed reduction in performance. Even with a more balanced dataset, routine screening variables are unlikely to provide the anatomical and physiological detail required for reliable differentiation between congenital heart disease subtypes. Definitive classification typically depends on complementary investigations, particularly echocardiography, which remains the reference standard for characterising cardiac anatomy and disease severity (Meng et al., 2024; Oster et al., 2025).

These findings therefore reinforce the intended role of probabilistic models within congenital heart disease screening. Rather than replacing diagnostic assessment, Bayesian models are best viewed as decision-support tools that identify children at increased risk and support timely referral for specialist evaluation. Their greatest clinical value lies in facilitating early risk stratification using routinely collected screening information, while recognising that definitive diagnosis and subtype classification require additional clinical and imaging assessment.

6.8 Strengths of the Study

This study presents several methodological and clinical strengths that support the validity, relevance, and potential applicability of the findings within the context of paediatric congenital heart disease screening.

One of the principal strengths lies in the use of clinically grounded predictor selection aligned with real-world screening practice. All variables included in the modelling process were derived from information routinely obtainable during standard clinical assessment, including demographic characteristics, physiological measurements, and auscultation findings. By deliberately excluding variables representing downstream diagnostic information, such as echocardiographic findings, the study avoided information leakage and ensured that model performance reflected realistic screening conditions rather than diagnostic confirmation. This strengthens the clinical relevance of the findings and enhances the potential applicability of the proposed models as screening support tools (Van Calster et al., 2026).

Another important strength is the use of interpretable probabilistic modelling through Bayesian Networks. Unlike many machine learning approaches that operate as “black-box” classifiers, Bayesian models provide explicit graphical representation of conditional dependencies between variables (Kyrimi et al., 2021; Rudin, 2019). This allowed visualisation and interpretation of clinically meaningful relationships identified within the dataset, including the dominant role of murmur-related variables and the clustering of anthropometric and age-related features. Such interpretability is particularly valuable in healthcare settings, where transparency and explainability are essential for clinical trust, adoption, and integration into decision-making processes (Topol, 2019).

A further strength of the study is its comprehensive modelling and validation strategy. Multiple Bayesian approaches, including general Bayesian Networks, Naïve Bayes, and Tree-Augmented Naïve Bayes, were evaluated, allowing comparison of predictive performance and interpretability across different probabilistic modelling paradigms (Van Calster et al., 2026). Internal validation combined stratified cross-validation with bootstrap resampling to assess the stability of the learned network structures, while formal comparison of ROC curves using the DeLong test provided additional statistical context for interpreting differences in model performance.

Another strength lies in the emphasis on clinical applicability. Model performance was evaluated using referral-oriented operating thresholds in addition to conventional classification thresholds, providing a more realistic assessment of performance within screening settings. Furthermore, the probabilistic framework was able to accommodate incomplete predictor information during inference, reflecting the uncertainty and variable data availability commonly encountered in routine clinical practice (Kyrimi et al., 2021). Predictor engineering was also based on established paediatric reference standards, ensuring that model inputs remained clinically meaningful and readily interpretable.

Finally, the study utilised a heterogeneous paediatric screening cohort spanning a broad range of ages, clinical presentations, and referral indications. Although external validation was not performed, this diversity enhances the ecological validity of the findings and supports the potential applicability of the proposed modelling approach to similar screening populations.

Taken together, these strengths support the credibility of the study and reinforce its contribution to the development of clinically interpretable probabilistic models for congenital heart disease screening. The combination of clinically grounded predictor selection, transparent modelling approaches, systematic comparative evaluation, and screening-oriented analysis provides a strong foundation for future research and potential clinical application.

6.9 Limitations

Several limitations should be considered when interpreting the findings and their potential applicability to clinical practice. These relate primarily to the characteristics of the available dataset, methodological decisions required during model development, and the absence of external validation.

One of the principal limitations is that all models were developed and evaluated using a single dataset derived from a specific paediatric cardiovascular screening programme. Although the dataset reflects routine clinical practice and includes a heterogeneous range of ages and referral indications, it may not be representative of broader paediatric populations or other healthcare settings. Furthermore, the prevalence of congenital heart disease within the analytic cohort was substantially higher than would be expected in population-based screening, reflecting enrichment with referred or clinically suspected cases. Although discrimination metrics such as AUC, sensitivity, and specificity are relatively independent of disease prevalence, predictive values and referral implications are not. In addition, the absence of external validation limits assessment of the generalisability and transportability of the proposed models, and independent validation in other screening populations is therefore required before clinical implementation (Van Calster et al., 2026).

Outcome construction represents an additional limitation. The source dataset did not contain a single predefined reference outcome; therefore, CHD_Binary and CHD_4Class were derived from unstructured cardiology and echocardiography diagnostic records using predefined keyword-based rules and a hierarchical combination procedure. CHD was considered present when either diagnostic source indicated congenital heart disease, while outcome status was classified as missing only when both sources lacked diagnostic information. Although this approach maximised the use of the available diagnostic records and enabled consistent outcome construction, echocardiographic confirmation was not uniformly available, and discordant source information was resolved through the predefined classification hierarchy rather than independent clinical adjudication. Some degree of outcome misclassification may therefore remain.

Missing data represented a further methodological challenge. Although Bayesian Networks can accommodate incomplete predictor information during probabilistic inference (Kyrimi et al., 2021), complete-case subsets were required during structure learning and comparative model evaluation, resulting in reduced sample sizes for some analyses, particularly BN-V2. Multiple imputation was not performed because several variables exhibited structural missingness (for example, murmur descriptors were only recorded when a murmur was present), and imputing these values could have introduced artificial

relationships into the learned network structures. In addition, several routinely collected predictors displayed limited variability, with most observations falling within normal clinical ranges. This restricted variability may have reduced the opportunity for more complex Bayesian Network structures to exploit additional conditional dependencies, potentially contributing to the similar performance observed across the different Bayesian modelling approaches (Helman et al., 2021).

Another methodological consideration relates to the discretisation of continuous variables. Continuous physiological and anthropometric measurements were categorised using established paediatric reference standards to maximise clinical interpretability and facilitate discrete Bayesian Network modelling. While this approach aligns with routine clinical decision-making and produces models that are more readily interpretable, it inevitably reduces information contained within the original continuous measurements and may have limited overall discriminative performance (Van Calster et al., 2026). Future studies could explore continuous or hybrid Bayesian Network approaches to determine whether preserving continuous information improves predictive performance while maintaining sufficient clinical interpretability.

The performance of the multi-class models was further limited by the combination of class imbalance and the restricted scope of the available predictor set (Helman et al., 2021). The relatively small number of Complex_CHD and Other_Cardiac cases reduced the ability of the models to learn reliable subtype-specific patterns, while routine clinical examination and auscultation variables alone provide limited anatomical and physiological detail for distinguishing between the four diagnostic categories. The incorporation of richer data sources, including quantitative phonocardiographic features, longitudinal clinical information, or imaging-derived variables, together with larger and more balanced datasets, may improve future multi-class modelling (Gudigar et al., 2024).

Despite these limitations, the study demonstrates the feasibility of developing clinically interpretable Bayesian models using routinely collected screening variables. The findings provide a foundation for future work focused on external validation, evaluation of alternative modelling strategies, and incorporation of additional data sources to further improve the performance and clinical applicability of Bayesian decision-support tools for congenital heart disease screening.

7 Conclusion

This dissertation investigated the development and evaluation of interpretable probabilistic models for congenital heart disease (CHD) screening in paediatric populations using routinely collected clinical, physiological, and auscultation variables. The primary aim was to determine whether Bayesian modelling approaches could support early identification of children with congenital heart disease while maintaining the interpretability and clinical applicability required for real-world screening.

The findings demonstrate that Bayesian models can provide clinically meaningful risk stratification for binary congenital heart disease detection using routine screening data alone. Across the evaluated modelling approaches, including general Bayesian Networks, Naïve Bayes, and Tree-Augmented Naïve Bayes models, predictive performance was consistently moderate, with no statistically significant differences observed between the principal binary models. These findings indicate that clinically useful screening performance can be achieved without relying on increasingly complex model structures and suggest that the predictive value of the models was influenced more by the quality and clinical relevance of the available screening variables than by the modelling strategy itself.

A key contribution of this dissertation is the demonstration that a relatively small set of routinely collected clinical examination and auscultation variables captured most of the discriminatory information required for binary CHD detection. Murmur-related variables consistently emerged as the dominant predictors, while expansion of the predictor set through the inclusion of blood pressure variables and detailed murmur descriptors did not improve model performance. These findings highlight the importance of clinically informed predictor selection and support the development of streamlined screening models that prioritise robust and consistently available clinical information over increasing model complexity.

Beyond predictive performance, this study demonstrates the value of Bayesian modelling as an interpretable framework for clinical decision support. Unlike many machine learning approaches, Bayesian Networks explicitly represent clinically meaningful relationships between variables while producing probabilistic estimates that can be adapted to different screening objectives. This flexibility allows decision thresholds to be adjusted according to clinical priorities, healthcare resources, and acceptable referral burdens, supporting their use as tools for early risk stratification and referral rather than as standalone diagnostic systems.

The exploratory multi-class analysis further demonstrated that routine screening variables alone are insufficient for reliable classification of the diagnostic categories represented in the four class outcome.

This finding reinforces the distinction between screening and diagnosis, highlighting that while routinely collected clinical information may be adequate for identifying children who require further investigation, definitive disease classification continues to depend on specialist assessment and imaging modalities such as echocardiography.

Overall, this dissertation demonstrates that interpretable Bayesian modelling provides a feasible and clinically relevant approach for supporting congenital heart disease screening using routinely collected clinical data. Rather than replacing specialist assessment or echocardiography, these models have the potential to support earlier identification of children at increased risk and prioritise referral for those most likely to benefit from further investigation. This may be particularly valuable in healthcare systems where access to paediatric cardiology services and diagnostic imaging is limited, allowing available specialist resources to be directed towards children at greatest risk while reducing unnecessary referrals. Although further external validation is required before clinical implementation, this work provides a foundation for the development of transparent, evidence-based decision-support tools that have the potential to improve the efficiency, equity, and accessibility of congenital heart disease screening by supporting more informed referral decisions, particularly in resource-constrained settings.

8 Implications for Future Research

The findings of this study highlight several important directions for future research aimed at improving the performance, generalisability, and clinical applicability of probabilistic models for congenital heart disease screening in paediatric populations.

A primary priority for future work is external validation of the developed models using independent datasets from different clinical environments and populations. Although the models demonstrated stable internal performance through cross-validation, validation in external cohorts is essential to assess robustness, reproducibility, and transportability across healthcare settings. Evaluation in population-based screening cohorts with congenital heart disease prevalence distributions closer to routine clinical practice would also allow assessment of calibration, predictive values, referral implications, and the need for recalibration of probabilistic outputs to different healthcare settings and disease prevalences.

Another important direction involves integration of additional data sources capable of improving predictive discrimination. This study intentionally focused on variables obtainable during routine screening encounters in order to preserve clinical practicality and applicability in resource-limited contexts. However, future models could incorporate quantitative acoustic features extracted from digital auscultation recordings, including frequency-domain, temporal, or spectral characteristics derived from phonocardiogram signals. These features may capture subtle physiological patterns not identifiable through conventional clinical auscultation alone and may improve both binary detection and more refined disease stratification.

Future work should also investigate alternative strategies for handling missing data. Although Bayesian Networks naturally accommodate incomplete observations during inference, structure learning relied on complete-case subsets, reducing the effective sample size for some analyses. Comparative evaluation of complete-case analysis, missingness-preserving approaches, and targeted imputation strategies may help identify the optimal balance between information retention, bias reduction, and model interpretability.

Given the limited performance observed in the multi-class models, future work may benefit from focusing on clinically actionable risk stratification rather than detailed subtype classification. Categorising children into low-, intermediate-, and high-risk groups may better reflect real-world screening pathways, where the objective is to guide referral decisions rather than establish definitive diagnoses. Larger and more balanced

datasets, together with richer diagnostic information such as quantitative phonocardiographic or imaging-derived features, may nevertheless improve future subtype classification.

Future studies should also move beyond retrospective model development towards prospective implementation and clinical evaluation. Assessing the impact of model-assisted screening on referral patterns, diagnostic yield, time to diagnosis, healthcare resource utilisation, and patient outcomes would provide important evidence regarding real-world clinical utility. In parallel, research focusing on clinician interaction with probabilistic decision-support systems will be essential to ensure usability, interpretability, and trust in clinical practice. User-centred implementation strategies, including integration into electronic health systems and incorporation of clinician feedback, may ultimately determine the success of such tools in routine care environments.

Finally, future research should continue to explore the broader role of interpretable artificial intelligence methods in paediatric medicine. One of the key strengths of Bayesian approaches lies in their ability to combine probabilistic prediction with transparent representation of clinically meaningful relationships between variables. Further methodological work aimed at improving interpretability, uncertainty estimation, and causal understanding within probabilistic frameworks may enhance both clinician acceptance and clinical impact.

Overall, future research should aim not only to improve predictive performance, but also to strengthen the interpretability, robustness, integration, and practical clinical value of probabilistic models for paediatric congenital heart disease screening.

9 References

Alkhodari, M., Hadjileontiadis, L. J., & Khandoker, A. H. (2024). Identification of Congenital Valvular Murmurs in Young Patients Using Deep Learning-Based Attention Transformers and Phonocardiograms. *IEEE Journal of Biomedical and Health Informatics*, *28*(4), 1803–1814. https://doi.org/10.1109/JBHI.2024.3357506

Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., & Druzdzel, M. J. (2019). Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine. *Value in Health*, *22*(4), 439–445. https://doi.org/10.1016/j.jval.2019.01.006

Bakker, M. K., Bergman, J. E. H., Krikov, S., Amar, E., Cocchi, G., Cragan, J., de Walle, H. E. K., Gatt, M., Groisman, B., Liu, S., Nembhard, W. N., Pierini, A., Rissmann, A., Chidambarathanu, S., Sipek Jr, A., Szabova, E., Tagliabue, G., Tucker, D., Mastroiacovo, P., & Botto, L. D. (2019). Prenatal diagnosis and prevalence of critical congenital heart defects: an international retrospective cohort study. *BMJ Open*, *9*(7), e028139. https://doi.org/10.1136/bmjopen-2018-028139

Begic, E., Gurbeta Pokvic, L., Begic, Z., Begic, N., Dedic, M., Mrsic, D., Jamakovic, M., Vila, N., & Badnjevic, A. (2021). From Heart Murmur to Echocardiography - Congenital Heart Defects Diagnostics Using Machine-Learning Algorithms. *Psychiatria Danubina*, *33*(Suppl 13), 236–246.

Chami, J., Strange, G., Baker, D., Cordina, R., Grigg, L., Celermajer, D. S., & Nicholson, C. (2023). Algorithmic complexity stratification for congenital heart disease patients. *International Journal of Cardiology Congenital Heart Disease*, *11*, 100430. https://doi.org/10.1016/j.ijcchd.2022.100430

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6), 137. https://doi.org/10.1007/s10462-024-10759-6

Cheng, S. P. S., Heo, K., Joos, E., Vervoort, D., & Joharifard, S. (2024). Barriers to Accessing Congenital Heart Surgery in Low- and Middle-Income Countries: A Systematic Review. *World Journal for Pediatric and Congenital Heart Surgery*, *15*(1), 94–103. https://doi.org/10.1177/21501351231204328

Chhatwal, K., Smith, J. J., Bola, H., Zahid, A., Venkatakrishnan, A., & Brand, T. (2023). Uncovering the Genetic Basis of Congenital Heart Disease: Recent Advancements and Implications for Clinical Management. *CJC Pediatric and Congenital Heart Disease*, *2*(6), 464–480. https://doi.org/10.1016/j.cjcpc.2023.10.008

Choudhury, T. Z., Gilbert, B. L., & Garg, V. (2025). Genetic and Environmental Contributors To Congenital Heart Disease. *Current Treatment Options in Cardiovascular Medicine*, *27*(1), 36. https://doi.org/10.1007/s11936-025-01091-5

Clifford, G. D., Liu, C., Moody, B., Springer, D., Silva, I., Li, Q., & Mark, R. G. (2016). Classification of Normal/Abnormal Heart Sound Recordings: the PhysioNet/Computing in Cardiology Challenge 2016. *Computing in Cardiology Conference (CinC)*, 609–612.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Gudigar, A., Raghavendra, U., Maithri, M., Samanth, J., Inamdar, M. A., Vidhya, V., Vicnesh, J., Prabhu, M. A., Tan, R.-S., Yeong, C. H., Molinari, F., & Acharya, U. R. (2024). Automated System for the Detection of Heart Anomalies Using Phonocardiograms: A Systematic Review. *IEEE Access*, *12*, 138399–138428. https://doi.org/10.1109/ACCESS.2024.3465511

Gurvitz, M., Krieger, E. V., Fuller, S., Davis, L. L., Kittleson, M. M., Aboulhosn, J. A., Bradley, E. A., Buber, J., Daniels, C. J., Dimopoulos, K., Egbe, A., Geoffrion, T. R., John, A., Khairy, P., Kim, Y. Y., Kreutzer, J., Lewis, M. J., Menachem, J. N., Moore, J. P., … Wald, R. M. (2026). 2025 ACC/AHA/HRS/ISACHD/SCAI Guideline for the Management of Adults With Congenital Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*, *153*(8). https://doi.org/10.1161/CIR.0000000000001402

Hashiramoto, S., Kaneko, M., Takita, H., Yamashita, Y., Matsuoka, R., & Sekizawa, A. (2025). Factors affecting the accuracy of fetal cardiac ultrasound screening in the first trimester of pregnancy. *Journal of Medical Ultrasonics*, *52*(1), 131–138. https://doi.org/10.1007/s10396-024-01505-0

Hassanuzzaman, M., Hasan, N. A., Al Mamun, M. A., Ahmed, K. I., Khandoker, A. H., & Mostafa, R. (2023). A Deep Learning Model for Recognizing Pediatric Congenital Heart Diseases Using Phonocardiogram Signals. *Computing in Cardiology*, (50), 1–4. https://doi.org/10.22489/CinC.2023.146

Helman, S. M., Herrup, E. A., Christopher, A. B., & Al-Zaiti, S. S. (2021). The role of machine learning applications in diagnosing and assessing critical and non-critical CHD: a scoping review. *Cardiology in the Young*, *31*(11), 1770–1780. https://doi.org/10.1017/S1047951121004212

Hoffman, J. I. E., & Kaplan, S. (2002). The incidence of congenital heart disease. *Journal of the American College of Cardiology*, *39*(12), 1890–1900. https://doi.org/10.1016/S0735-1097(02)01886-7

Hsieh, Y.-T., Liu, H.-K., Wu, J.-R., Yan, T.-Y., Huang, Y.-J., Guo, M., & Yang, M.-C. (2025). Development and validation of an integrated residual-recurrent neural network model for automated heart murmur detection in pediatric populations. *Scientific Reports*, *15*(1), 19155. https://doi.org/10.1038/s41598-025-04746-2

Jia, W., Wang, Y., Chen, R., Ye, J., Li, D., Yin, F., Yu, J., Chen, J., Shu, Q., & Xu, W. (2024). ZCHSound: Open-Source ZJU Paediatric Heart Sound Database With Congenital Heart Disease. *IEEE Transactions on Biomedical Engineering*, *71*(8), 2278–2286. https://doi.org/10.1109/TBME.2023.3348800

Kaur, I., & Ahmad, T. (2024). A cluster-based ensemble approach for congenital heart disease prediction. *Computer Methods and Programs in Biomedicine*, *243*, 107922. https://doi.org/10.1016/j.cmpb.2023.107922

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, *17*(1), 195. https://doi.org/10.1186/s12916-019-1426-2

Khan, K., Ahmad, T., & Uddin, I. (2023). Cardiac Deep Learning Model for Congenital Heart Disease Recognition. *2023 15th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 293–297. https://doi.org/10.1109/SKIMA59232.2023.10387299

Korb, K. B., & Nicholson, A. E. (2011). *Bayesian Artificial Intelligence* (2nd Edition). CRC Press.

Kumar, H., Abbas, A., & Ewer, A. K. (2025). Newborn Pulse-Oximetry Screening. *Clinics in Perinatology*, *52*(3), 539–554. https://doi.org/10.1016/j.clp.2025.06.011

Kyrimi, E., McLachlan, S., Dube, K., Neves, M. R., Fahmi, A., & Fenton, N. (2021). A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. *Artificial Intelligence in Medicine*, *117*, 102108. https://doi.org/10.1016/j.artmed.2021.102108

Lemieux, A., Khalilipalandi, S., Lauzon-Schnittka, J., Taillefer, V., Tousignant, A., Perreault, L., Rego, K., Dubois, M., Watelle, L., Roy, L.-O., & Dallaire, F. (2024). Meta-Analysis of Risk Factors for Congenital Heart Disease: Part 2, Maternal Medication, Reproductive Technologies, and Familial and Fetal Factors. *Canadian Journal of Cardiology*, *40*(12), 2496–2511. https://doi.org/10.1016/j.cjca.2024.09.011

Lopes, S. A. V. do A., Guimarães, I. C. B., Costa, S. F. de O., Acosta, A. X., Sandes, K. A., & Mendes, C. M. C. (2018). Mortality for Critical Congenital Heart Diseases and Associated Risk Factors in Newborns. A Cohort Study. *Arquivos Brasileiros de Cardiologia*. https://doi.org/10.5935/abc.20180175

Lopez-Barreda, R., Salas, W., Pavlova, M., Groot, W., & Ibla, J. C. (2025). Access to healthcare for children with CHDs in middle and low-income countries: a systematic literature review. *Cardiology in the Young*, *35*(3), 482–489. https://doi.org/10.1017/S1047951124036485

Lucas, P. J. F., van der Gaag, L. C., & Abu-Hanna, A. (2004). Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, *30*(3), 201–214. https://doi.org/10.1016/j.artmed.2003.11.001

Mangione, S., & Nieman, L. Z. (1997). Cardiac Auscultatory Skills of Internal Medicine and Family Practice Trainees. *JAMA*, *278*(9), 717. https://doi.org/10.1001/jama.1997.03550090041030

Mayer, P. J., Poyraz, R. A., Hölter, T., Thun, S., & Vorisek, C. N. (2024). Data for AI in Congenital Heart Defects: Systematic Review. *Studies in Health Technology and Informatics*, (316), 820–821. https://doi.org/10.3233/SHTI240537

Meller, C. H., Otaño, L., Marantz, P., Grinenco, S., Aiello, H., Córdoba, A., & Sáenz-Tejeira, M. M. (2020). Congenital heart disease, prenatal diagnosis and management. *Archivos Argentinos de Pediatria*, *118*(2). https://doi.org/10.5546/aap.2020.eng.e149

Meng, X., Song, M., Zhang, K., Lu, W., Li, Y., Zhang, C., & Zhang, Y. (2024). Congenital heart disease: types, pathophysiology, diagnosis, and treatment options. *MedComm*, *5*(7). https://doi.org/10.1002/mco2.631

Oliveira, J., Renna, F., Costa, P. D., Nogueira, M., Oliveira, C., Ferreira, C., Jorge, A., Mattos, S., Hatem, T., Tavares, T., Elola, A., Rad, A. B., Sameni, R., Clifford, G. D., & Coimbra, M. T. (2022). The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification. *IEEE Journal of Biomedical and Health Informatics*, *26*(6), 2524–2535. https://doi.org/10.1109/JBHI.2021.3137048

Oster, M. E., Pinto, N. M., Pramanik, A. K., Markowsky, A., Schwartz, B. N., Kemper, A. R., Hom, L. A., Martin, G. R., Berger, S., Scott Snyder, C., Bertanyi Armsby, L., Gabriel Cabrera, A., Hsu, D. T., Mery, C. M., Johnson, J., Sachdeva, R., Villafane, J., Renno, M., Das, N., … Wade, K. (2025). Newborn Screening for Critical Congenital Heart Disease: A New Algorithm and Other Updated Recommendations: Clinical Report. *Pediatrics*, *155*(1). https://doi.org/10.1542/peds.2024-069667

Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Qu, L., Vetter, V. L., Bird, G. L., Qiu, H., & White, P. S. (2010). A Naïve Bayes classifier for differential diagnosis of Long QT Syndrome in children. *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 433–437. https://doi.org/10.1109/BIBM.2010.5706605

R Core Team. (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://doi.org/10.32614/R.manuals

Reddy, C. D., Van den Eynde, J., & Kutty, S. (2022). Artificial intelligence in perinatal diagnosis and management of congenital heart disease. *Seminars in Perinatology*, *46*(4), 151588. https://doi.org/10.1016/j.semperi.2022.151588

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Scutari, M. (2010). Learning Bayesian Networks with the **bnlearn** *R* Package. *Journal of Statistical Software*, *35*(3). https://doi.org/10.18637/jss.v035.i03

Shantha Selvakumari, R., Punithakumar, K., Selvathi, D., Chrisilla, S., Sri Abinaya, R., & Ray, N. (2025). Computerized Diagnosis Methods for Pediatric Cardiology: A Review. *IEEE Access*, *13*, 39192–39213. https://doi.org/10.1109/ACCESS.2025.3545829

Siefkes, H., Oliveira, L. C., Koppel, R., Hogan, W., Garg, M., Manalo, E., Cresalia, N., Lai, Z., Tancredi, D., Lakshminrusimha, S., & Chuah, C. (2024). Machine Learning–Based Critical Congenital Heart Disease Screening Using Dual-Site Pulse Oximetry Measurements. *Journal of the American Heart Association*, *13*(12). https://doi.org/10.1161/JAHA.123.033786

Skivington, K., Matthews, L., Simpson, S. A., Craig, P., Baird, J., Blazeby, J. M., Boyd, K. A., Craig, N., French, D. P., McIntosh, E., Petticrew, M., Rycroft-Malone, J., White, M., & Moore, L. (2021). A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ*, n2061. https://doi.org/10.1136/bmj.n2061

Sun, R., Liu, M., Lu, L., Zheng, Y., & Zhang, P. (2015). Congenital Heart Disease: Causes, Diagnosis, Symptoms, and Treatments. *Cell Biochemistry and Biophysics*, *72*(3), 857–860. https://doi.org/10.1007/s12013-015-0551-6

Tavares, T. R., Oliveira, A. L. I., Cabral, G. G., Mattos, S. S., & Grigorio, R. (2013). Preprocessing unbalanced data using weighted support vector machines for prediction of heart disease in children. *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2013.6706947

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Van Calster, B., van Smeden, M., van Amsterdam, W., Coemans, M., Wynants, L., & Steyerberg, E. W. (2026). The Enemies of Reliable and Useful Clinical Prediction Models: A Review of Statistical and Scientific Challenges. *Annual Review of Statistics and Its Application*, *13*(1), 465–492. https://doi.org/10.1146/annurev-statistics-042324-123749

van der Linde, D., Konings, E. E. M., Slager, M. A., Witsenburg, M., Helbing, W. A., Takkenberg, J. J. M., & Roos-Hesselink, J. W. (2011). Birth Prevalence of Congenital Heart Disease Worldwide. *Journal of the American College of Cardiology*, *58*(21), 2241–2247. https://doi.org/10.1016/j.jacc.2011.08.025

van Velzen, C., Clur, S., Rijlaarsdam, M., Bax, C., Pajkrt, E., Heymans, M., Bekker, M., Hruda, J., de Groot, C., Blom, N., & Haak, M. (2016). Prenatal detection of congenital heart disease—results of a national screening programme. *BJOG: An International Journal of Obstetrics & Gynaecology*, *123*(3), 400–407. https://doi.org/10.1111/1471-0528.13274

Wang, J., Zhang, S., Sun, L., Wang, L., & Wu, Q. (2025). Association between maternal folic acid and/or multivitamin supplementation time and fetal congenital heart disease: based on the China birth cohort study. *International Journal of Medical Sciences*, *22*(2), 260–268. https://doi.org/10.7150/ijms.102843

World Health Organization. (2023). *Congenital disorders - Fact Sheet.* https://www.who.int/news-room/fact-sheets/detail/birth-defects/

Xu, J., Li, Q., Deng, L., Xiong, J., Cheng, Z., & Ye, C. (2025). Global, regional, and national epidemiology of congenital heart disease in children from 1990 to 2021. *Frontiers in Cardiovascular Medicine*, *12*. https://doi.org/10.3389/fcvm.2025.1522644

Zaidi, S., & Brueckner, M. (2017). Genetics and Genomics of Congenital Heart Disease. *Circulation Research*, *120*(6), 923–940. https://doi.org/10.1161/CIRCRESAHA.116.309140

Zhang, F., Qi, L., Zhao, M., Han, S., Zhang, H., & Wang, G. (2024). Global research landscape on the genetics of congenital heart disease: A bibliometric and visualized analysis via VOSviewer and CiteSpace. *Medicine*, *103*(43), e40261. https://doi.org/10.1097/MD.0000000000040261

Zhang, S., Kang, C., Cui, J., Xue, H., Zhao, S., Chen, Y., Lu, H., Ye, L., Wang, D., Chen, F., Zhao, Y., Pei, L., & Qu, P. (2025). Development of machine learning-based models to predict congenital heart disease: A matched case-control study. *International Journal of Medical Informatics*, *195*, 105741. https://doi.org/10.1016/j.ijmedinf.2024.105741

Zhao, Q., Niu, C., Liu, F., Wu, L., Ma, X., & Huang, G. (2019). Accuracy of cardiac auscultation in detection of neonatal congenital heart disease by general paediatricians. *Cardiology in the Young*, *29*(5), 679–683. https://doi.org/10.1017/S1047951119000799

10 Appendices

Appendix A – Variable Definitions and Preprocessing Summary

This appendix summarises the outcome and predictor variables retained following preprocessing of the original dataset, including their definitions, categorisation procedures, modelling rationale, and inclusion across the probabilistic models developed in this study. Predictor inclusion varied according to the modelling strategy described in Chapter 5, and not all variables were used simultaneously across all models.

A.1 – Outcome Variables

Two outcome variables were used in this study. CHD_Binary served as the primary outcome for all binary classification models and distinguished between the presence and absence of congenital heart disease. CHD_4Class was developed for the exploratory multi-class analysis and distinguished between Normal, Simple_CHD, Complex_CHD, and Other_Cardiac findings. Both outcomes were derived by processing cardiology and echocardiography diagnostic records separately and subsequently combining the source-specific classifications according to the predefined hierarchical rules described in Section 4.3. Neither diagnostic source was assigned formal precedence within the implemented classification procedure.

Table A.1 – Outcome Variables Used in Model Development

Outcome Variable	Type	Categories	Models
CHD_Binary	Binary	CHD; No CHD	BN-V1, BN-V2, BN-V3, NB-V1, TAN-V1
CHD_4Class	Categorical	Normal; Simple_CHD; Complex_CHD; Other_Cardiac	BN-V1-4Class, NB-V1-4Class

CHD_Binary

CHD_Binary was developed as the primary outcome variable to reflect the principal objective of congenital heart disease screening: identifying children who may require further diagnostic assessment. The original datasets did not contain a predefined binary outcome variable. Diagnostic information recorded across multiple cardiology and echocardiography fields was therefore consolidated and classified separately for each source using predefined keyword-based rules.

Participants were classified as having congenital heart disease when either the cardiology or echocardiography source indicated CHD. They were classified as not having CHD when neither source indicated congenital heart disease and at least one source contained usable diagnostic information. The outcome was classified as missing only when diagnostic information was unavailable from both sources. Participants with a missing outcome after this procedure (n = 164) were excluded from supervised modelling. CHD_Binary served as the primary endpoint for all binary classification models.

CHD_4Class

CHD_4Class was derived from the same source-specific cardiology and echocardiography classifications used to construct *CHD_Binary* but retained greater diagnostic detail. Participants were classified as *Complex_CHD* when either source indicated CHD and contained terminology associated with a complex

congenital defect. In the absence of complex CHD, *Simple_CHD* was assigned when either source indicated congenital heart disease. *Other_Cardiac* was assigned when neither source indicated CHD but at least one source indicated a non-congenital cardiac abnormality. Normal was assigned when both sources were normal or when one source was normal and the other contained no diagnostic information. The outcome was classified as missing only when diagnostic information was unavailable from both sources.

The four-class outcome was developed as an exploratory extension of the primary binary outcome to investigate whether probabilistic models could distinguish between clinically relevant diagnostic categories using only routinely collected screening variables. Unlike *CHD_Binary*, which reflects the screening decision of whether congenital heart disease is present, *CHD_4Class* was intended to evaluate the feasibility of differentiating disease complexity while recognising that definitive subtype classification generally requires imaging-based assessment.

A.2 – Predictor Variables Included in each Model

Table A.2 summarises the predictor variables included in each probabilistic model developed in this study. The binary classification models shared a common set of core demographic, physiological, anthropometric, and clinical examination variables. Additional predictors were incorporated into specific Bayesian Network models to investigate whether expanding the predictor set improved predictive performance. The exploratory four-class models used the same predictor set as BN-V1 and NB-V1, differing only in the outcome variable presented in Section A.1.

Table A.2 – Predictor Variables Included in Each Model

Variable	BN-V1	BN-V2	BN-V3	NB-V1	TAN-V1	BN-V1-4Class	NB-V1-4Class
Gender	✓	✓	✓	✓	✓	✓	✓
MACS Age Class	✓	✓	✓	✓	✓	✓	✓
Indication_Any	✓	✓	✓	✓	✓	✓	✓
O2 Saturation Preductal Class	✓	✓	✓	✓	✓	✓	✓
Heart Rate Class	✓	✓	✓	✓	✓	✓	✓
Weight Age Z Class	✓	✓	✓	✓	✓	✓	✓
Height Age Z Class	✓	✓	✓	✓	✓	✓	✓
Clinical Murmur	✓	✓	✓	✓	✓	✓	✓
S2	✓	✓	✓	✓	✓	✓	✓
Precordium	✓	✓	✓	✓	✓	✓	✓
Skin	✓	✓	✓	✓	✓	✓	✓
SBP Class		✓					
DBP Class		✓					
Murmurs timing			✓				
Murmurs pitch			✓				
Murmurs grading			✓				

Murmurs location	✓

A.3 – Predictor Variable Description

The following sections describe the predictor variables used during model development, including their definition, preprocessing, categorisation, clinical relevance, and modelling role. Variables are presented in the same order as Table A.2.

A.3.1 Demographic Variables

Gender

- Definition: Biological sex recorded during screening.
- Preprocessing: Standardised to two categories (Male/Female); no further transformation was required.
- Categories and Distribution:
    - Male – n = 450 (52.02%)
    - Female – n = 415 (47.98%)
- Missingness after preprocessing: 0/865 (0.00%)
- Modelling rationale: Included as a demographic predictor because sex-related differences have been reported in the prevalence and presentation of congenital heart disease. However, it demonstrated limited predictive contribution in the Bayesian Network models.

MACS Age Class

- Definition: Age classification representing the participant's developmental stage according to the Munich Age Classification System (MACS).
- Preprocessing: Derived from the recorded age (days, months, and years) by assigning each participant to the corresponding MACS age group. The original continuous age variables were retained in the dataset but were not used directly for modelling.
- Categories and Distribution:
    - Neonate (≤28 days) – n = 6 (0.69%)
    - Infant (>28 days to <1 year) – n = 125 (14.45%)
    - Toddler (1 to <3 years) – n = 135 (15.61%)
    - Early Childhood (3 to <6 years) – n = 166 (19.19%)
    - Late Childhood (6 to <12 years) – n = 362 (41.85%)
    - Adolescent (12 to <18 years) – n = 71 (8.21%)
- Missingness after preprocessing: 0/865 (0.00%)
- Modelling rationale: Included because age is a fundamental determinant of physiological parameters, growth patterns, and the clinical presentation of congenital heart disease. Categorisation using the MACS age groups also enabled age-specific classification of

physiological variables (e.g., heart rate) and standardisation of anthropometric measurements, while improving model interpretability.

A.3.2 Clinical History

Indication_Any

- Definition: Binary variable indicating whether at least one clinical indication for cardiology referral or screening was recorded.
- Preprocessing: Derived by combining the multiple clinical indication fields available in the original dataset into a single binary variable. Participants were classified according to the presence or absence of any recorded referral indication, regardless of the specific indication.
- Categories and Distribution:
    - Yes – n = 302 (34.91%)
    - No – n = 563 (65.09%)
- Missingness after preprocessing: 0/865 (0.00%)
- Modelling rationale: Included to capture whether a participant presented with a recognised indication for cardiology assessment while reducing the dimensionality of the original dataset. Collapsing multiple indication variables into a single binary predictor improved model simplicity and avoided sparse representations arising from numerous low-frequency indication categories.

A.3.3 Physiological Variables

O2 Saturation Preductal Class

- Definition: Preductal oxygen saturation measured by pulse oximetry during the screening examination, categorised into clinically meaningful groups.
- Preprocessing: Continuous preductal oxygen saturation measurements were categorised into three clinically meaningful groups. Although oxygen saturation thresholds do not vary substantially with age, exploratory analysis identified a small number of markedly low oxygen saturation values (as low as 40%). These observations may represent either severe hypoxaemia associated with critical congenital heart disease or measurement artefacts acquired during field screening. Rather than combining these observations with the broader low-saturation group, they were retained as a separate "Extremely Low" category to preserve potentially informative variation while limiting the influence of extreme values on model learning. Missing values were retained as NA.
- Categories and Distribution:
    - Normal (95-100%) – n = 640 (73.99%)
    - Low (85-94%) – n = 155 (17.92%)
    - Extremely Low (<85%) – n = 42 (4.86%)
- Missingness after preprocessing: 28/865 (3.24%)
- Modelling rationale: Included as an objective physiological indicator of impaired oxygenation that may reflect cyanotic or haemodynamically significant forms of congenital heart disease.

Heart Rate Class

- Definition: Age-adjusted heart rate classification based on the participant's recorded heart rate during the screening examination.
- Preprocessing: Continuous heart rate measurements were categorised into Low (Bradycardia), Normal, and High (Tachycardia) using age-specific reference ranges based on the Munich Age Classification System (MACS). Classification thresholds were adapted from the Pediatric Advanced Life Support (PALS) guidelines published by the American Heart Association (AHA). The original continuous heart rate values were retained in the dataset but were not used directly for modelling.
- Categories and Distribution:
    - Low (Bradycardia) – n = 94 (10.87%)
    - Normal – n = 609 (70.40%)
    - High (Tachycardia) – n = 70 (8.09%)

  Age-specific classification thresholds are presented in Table A.3.
- Missingness after preprocessing: 92/865 (10.64%)
- Modelling rationale: Included as a routinely collected physiological variable because heart rate reflects cardiovascular status and may be altered in children with congenital heart disease. Categorisation using age-specific reference ranges standardised interpretation across paediatric developmental stages while producing a discrete predictor suitable for Bayesian Network modelling. Missing values were retained as *NA*, allowing the probabilistic models to accommodate uncertainty during inference.

Table A.3 – Age-Specific Heart Rate Thresholds Used to Derive Heart Rate Class

MACS Age Group	Low (Bradycardia)	Normal Heart Rate	High (Tachycardia)
Neonate (≤28 days)	< 100	100 – 205	> 205
Infant (>28 days to <1 year)	< 100	100 – 180	> 180
Toddler (1 to <3 years)	< 98	98 – 140	> 140
Early Childhood (3 to <6 years)	< 80	80 – 120	> 120
Late Childhood (6 to <12 years)	< 75	75 – 118	> 118
Adolescent (12 to <18 years)	< 60	60 – 100	> 100

SBP Class

- Definition: Age-adjusted systolic blood pressure classification derived from the participant's recorded systolic blood pressure during the screening examination.
- Preprocessing: Continuous systolic blood pressure measurements were categorised into Low, Normal, and High using age-specific reference ranges based on MACS. Classification thresholds were adapted from the PALS guidelines published by AHA. The original continuous systolic blood pressure values were retained in the dataset but were not used directly for modelling.

- Categories and Distribution:
    - Low – n = 409 (47.28%)
    - Normal – n = 267 (30.87%)
    - High – n = 44 (5.09%)

  Age-specific classification thresholds are presented in Table A.4.
- Missingness after preprocessing: 145/865 (16.76%)
- Modelling rationale: Included to investigate whether age-adjusted systolic blood pressure provided additional predictive information for congenital heart disease detection. Categorisation using age-specific reference ranges standardised interpretation across paediatric developmental stages while producing a discrete predictor suitable for Bayesian Network modelling. Missing values were retained as *NA*, allowing the probabilistic models to accommodate uncertainty during inference.

Table A.4 – Age-Specific Systolic Blood Pressure Thresholds Used to Derive SBP Class

MACS Age Group	Low	Normal (mmHg)	High
Neonate (≤28 days)	< 67	67 – 84	> 84
Infant (>28 days to <1 year)	< 72	72 – 104	> 104
Toddler (1 to <3 years)	< 86	86 – 106	> 106
Early Childhood (3 to <6 years)	< 89	89 – 112	> 112
Late Childhood (6 to <12 years)	< 97	97 – 115	> 115
Adolescent (12 to <18 years)	< 110	110 – 131	> 131

DBP Class

- Definition: Age-adjusted diastolic blood pressure classification derived from the participant's recorded diastolic blood pressure during the screening examination.
- Preprocessing: Continuous diastolic blood pressure measurements were categorised into Low, Normal, and High using age-specific reference ranges based on MACS. Classification thresholds were adapted from the PALS guidelines published by AHA. The original continuous diastolic blood pressure values were retained in the dataset but were not used directly for modelling.
- Categories and Distribution:
    - Low – n = 252 (29.13%)
    - Normal – n = 429 (49.60%)
    - High – n = 39 (4.51%)

  Age-specific classification thresholds are presented in Table A.5.
- Missingness after preprocessing: 145/865 (16.76%)
- Modelling rationale: Included alongside SBP Class to preserve potentially discordant systolic and diastolic blood pressure abnormalities, which may provide complementary physiological information. Categorisation using age-specific reference ranges standardised interpretation across paediatric developmental stages while producing a discrete predictor suitable for Bayesian

Network modelling. Missing values were retained as *NA*, allowing the probabilistic models to accommodate uncertainty during inference.

Table A.5 – Age-Specific Diastolic Blood Pressure Thresholds Used to Derive DBP Class

MACS Age Group	Low	Normal (mmHg)	High
Neonate (≤28 days)	< 35	35 – 53	> 53
Infant (>28 days to <1 year)	< 37	37 – 56	> 56
Toddler (1 to <3 years)	< 42	42 – 63	> 63
Early Childhood (3 to <6 years)	< 46	46 – 72	> 72
Late Childhood (6 to <12 years)	< 57	57 – 76	> 76
Adolescent (12 to <18 years)	< 64	64 – 83	> 83

A.3.4 Anthropometric Variables

Weight Age Z Class

- Definition: Age and sex standardised classification of body weight derived from WHO growth references.
- Preprocessing: Raw weight measurements were transformed into weight-for-age Z-scores using World Health Organization (WHO) growth references. For children aged 0–5 years, Z-scores were calculated using the WHO 2006 Child Growth Standards implemented through the *anthro* R package. For children aged 5–19 years, WHO 2007 LMS reference tables were used. The resulting Z-scores were categorised into Low, Normal, and High using thresholds of ±2 standard deviations from the WHO reference median.
- Categories and Distribution:
    - Low ($Z < -2$) – n = 173 (20.00%)
    - Normal ($-2 \leq Z \leq 2$) – n = 460 (53.18%)
    - High ($Z > 2$) – n = 199 (23.01%)
- Missingness after preprocessing: 33/865 (3.82%).
- Modelling rationale: Included to provide an age- and sex-standardised measure of growth that is directly comparable across the wide paediatric age range included in this study. This derived variable replaced the raw weight measurement, which is inherently dependent on age and sex and therefore unsuitable for direct comparison across paediatric populations. Body mass index (BMI) was not included because it is not validated for children younger than two years of age and was considered redundant following the derivation of the WHO age- and sex-standardised anthropometric classifications.

Height Age Z Class

- Definition: Age and sex standardised classification of body height derived from WHO growth references.

- Preprocessing: Raw height measurements were transformed into height-for-age Z-scores using World Health Organization (WHO) growth references. For children aged 0–5 years, Z-scores were calculated using the WHO 2006 Child Growth Standards implemented through the *anthro* R package. For children aged 5–19 years, WHO 2007 LMS reference tables were used. The resulting Z-scores were categorised into Low, Normal, and High using thresholds of ±2 standard deviations from the WHO reference median.
- Categories and Distribution:
    - Low ($Z < -2$) – n = 132 (15.26%)
    - Normal ($-2 \leq Z \leq 2$) – n = 554 (64.05%)
    - High ($Z > 2$) – n = 135 (15.61%)
- Missingness after preprocessing: 44/865 (5.09%)
- Modelling rationale: Included to provide an age- and sex-standardised measure of linear growth that is directly comparable across the paediatric age range. This derived variable replaced the raw height measurement, ensuring that anthropometric status was assessed independently of age and sex.

A.3.5 Physical Examination Variables

Clinical Murmur

- Definition: Presence or absence of a cardiac murmur detected by the field clinician during auscultation at the time of screening.
- Preprocessing: The original variable (Murmurs) was renamed Clinical_Murmur and standardised as a binary categorical variable representing the presence (Present) or absence (Absent) of a clinically detected murmur. Missing values were retained as *NA*.
- Categories and Distribution:
    - Present – n = 378 (43.70%)
    - Absent – n = 471 (54.45%)
- Missingness after preprocessing: 16/865 (1.85%)
- Modelling rationale: Included as a core clinical predictor because it represents the clinician's auscultatory assessment during routine screening, prior to echocardiographic confirmation. As a readily available clinical finding, murmur presence reflects real-world screening practice and provides important information for congenital heart disease risk assessment while remaining suitable for Bayesian Network modelling.

S2

- Definition: Clinical assessment of the second heart sound (S2) recorded during cardiac auscultation at the time of screening.
- Preprocessing: Used directly as a categorical variable representing whether the second heart sound was assessed as Normal or Abnormal. No further transformation was applied. Missing values were retained as *NA*.

- Categories and Distribution:
    - Normal – n = 833 (96.30%)
    - Abnormal – n = 20 (2.31%)
- Missingness after preprocessing: 12/865 (1.39%)
- Modelling rationale: Included because abnormalities of the second heart sound may reflect structural or haemodynamic cardiac abnormalities associated with congenital heart disease. As part of the routine cardiac examination, S2 provided complementary auscultatory information beyond the presence of a murmur while remaining readily available in clinical screening settings.

Precordium

- Definition: Clinical assessment of the character of the precordial impulse recorded during the physical examination.
- Preprocessing: Used directly as a categorical variable representing the clinician's assessment of the precordium. No further transformation was applied. Missing values were retained as *NA*.
- Categories and Distribution:
    - Normal – n = 503 (58.15%)
    - Quiet – n = 337 (38.96%)
    - Hyperdynamic – n = 11 (1.27%)
- Missingness after preprocessing: 14/865 (1.62%)
- Modelling rationale: Included because abnormalities of the precordial impulse may reflect underlying alterations in cardiac function or haemodynamics associated with congenital heart disease. As part of the routine cardiovascular examination, this variable provided complementary clinical information alongside auscultation findings while remaining readily available in screening settings.

Skin

- Definition: Clinical assessment of skin appearance recorded during the physical examination, reflecting observable signs of peripheral perfusion and oxygenation.
- Preprocessing: Used directly as a categorical variable representing the clinician's assessment of skin appearance. No further transformation was applied. Missing values were retained as *NA*.
- Categories and Distribution:
    - Normal – n = 827 (95.61%)
    - Cyanotic – n = 14 (1.62%)
    - Hypoperfused – n = 9 (1.04%)
- Missingness after preprocessing: 15/865 (1.73%), carried forward from the original variable.
- Modelling rationale: Included because skin appearance provides readily observable clinical indicators of cardiopulmonary status during routine screening. Findings such as cyanosis and hypoperfusion may reflect impaired oxygenation or circulatory compromise associated with congenital heart disease, providing complementary information to the cardiovascular examination while remaining easily assessable in routine clinical practice.

A.3.6 Murmur Descriptor Variables

Murmur Timing

- Definition: Clinical assessment of the phase of the cardiac cycle during which a murmur was detected during auscultation.
- Preprocessing: Used as a categorical variable describing murmur timing. Because this variable was only applicable when a murmur was present, structural missingness was explicitly recoded into contextual categories. Participants without a clinically detected murmur were classified as Not Applicable, participants with missing Clinical_Murmur status were classified as Unknown, and participants with a documented murmur but no recorded timing were classified as Missing. No imputation was performed.
- Categories and Distribution:
    - Systolic – n = 371 (42.89%)
    - Diastolic – n = 3 (0.35%)
    - Continuous – n = 3 (0.35%)
    - Not Applicable – n = 471 (54.45%)
    - Unknown – n = 16 (1.85%)
- Missingness after preprocessing: 1/865 (0.12%) represented a true missing value (murmur present but timing not recorded). In addition, 471 observations (54.45%) were classified as Not Applicable because no murmur was detected, and 16 observations (1.85%) were classified as Unknown because the parent variable (Clinical_Murmur) was missing.
- Modelling rationale: Included to investigate whether detailed auscultatory characteristics improved congenital heart disease classification beyond murmur presence alone. Murmur timing is a fundamental descriptor used during cardiac auscultation and provides clinically relevant information for distinguishing physiological from pathological murmurs. Explicit encoding of contextual categories preserved the hierarchical relationship with Clinical_Murmur while preventing structural absence from being conflated with data uncertainty.

Murmur Pitch

- Definition: Clinical assessment of the tonal quality of a cardiac murmur detected during auscultation.
- Preprocessing: Used as a categorical variable describing murmur pitch. As this variable was only applicable when a murmur was present, structural missingness was explicitly recoded into contextual categories. Participants without a clinically detected murmur were classified as Not Applicable, participants with missing Clinical_Murmur status were classified as Unknown, and participants with a documented murmur but no recorded murmur pitch were classified as Missing. No imputation was performed.
- Categories and Distribution:

  - Harsh – n = 193 (22.31%)
  - Smooth – n = 176 (20.35%)
  - Not Applicable – n = 471 (54.45%)
  - Unknown – n = 16 (1.85%)
- Missingness after preprocessing: 9/865 (1.04%) represented true missing values (murmur present but pitch not recorded). In addition, 471 observations (54.5%) were classified as Not Applicable because no murmur was detected, and 16 observations (1.85%) were classified as Unknown because the parent variable (Clinical_Murmur) was missing.
- Modelling rationale: Included to investigate whether the tonal quality of a murmur provided additional discriminatory information beyond murmur presence alone. Murmur pitch is a recognised auscultatory characteristic that may assist in distinguishing physiological from pathological murmurs. Explicit encoding of contextual categories preserved the hierarchical relationship with Clinical_Murmur while preventing structural absence from being conflated with data uncertainty.

Murmur Grading

- Definition: Clinical assessment of murmur intensity recorded during auscultation using the Levine grading scale.
- Preprocessing: Used as an ordinal categorical variable representing murmur intensity. As this variable was only applicable when a murmur was present, structural missingness was explicitly recoded into contextual categories. Participants without a clinically detected murmur were classified as Not Applicable, participants with missing Clinical_Murmur status were classified as Unknown, and participants with a documented murmur but no recorded murmur grade were classified as Missing. No imputation was performed.
- Categories and Distribution:
  - Grade I/VI – n = 153 (17.69%)
  - Grade II/VI – n = 142 (16.42%)
  - Grade III/VI – n = 35 (4.05%)
  - Grade IV/VI – n = 11 (1.27%)
  - Grade V/VI – n = 2 (0.23%)
  - Not Applicable – n = 471 (54.45%)
  - Unknown – n = 16 (1.85%)
- Missingness after preprocessing: Thirty-five observations (4.05%) represented true missing values (murmur present but grade not recorded). In addition, 471 observations (54.5%) were classified as Not Applicable because no murmur was detected, and 16 observations (1.85%) were classified as Unknown because the parent variable (Clinical_Murmur) was missing.
- Modelling rationale: Included to investigate whether murmur intensity provided additional discriminatory information beyond murmur presence alone. Murmur grading is a standard component of cardiac auscultation and reflects the intensity of turbulent blood flow, making it a clinically relevant descriptor for congenital heart disease screening. Explicit

encoding of contextual categories preserved the hierarchical relationship with Clinical_Murmur while preventing structural absence from being conflated with data uncertainty.

Murmur Location

- Definition: Clinical assessment of the primary anatomical location where a cardiac murmur was best heard during auscultation.
- Preprocessing: Used as a categorical variable describing the site of maximal murmur intensity. As this variable was only applicable when a murmur was present, structural missingness was explicitly recoded into contextual categories. Participants without a clinically detected murmur were classified as Not Applicable, participants with missing Clinical_Murmur status were classified as Unknown, and participants with a documented murmur but no recorded location were classified as Missing. No imputation was performed.
- Categories and Distribution:
    - Left Lower Sternal Border (LLSB) – n = 42 (4.86%)
    - Left Upper Sternal Border (LUSB) – n = 33 (3.82%)
    - Erb's Point (EP) – n = 29 (3.35%)
    - Apex – n = 4 (0.46%)
    - Right Upper Sternal Border (RUSB) – n = 1 (0.12%)
    - Left Sternal Border (unspecified) – n = 2 (0.23%)
    - EP and LLSB – n = 6 (0.69%)
    - LLSB and Apex – n = 1 (0.12%)
    - LUSB and LLSB – n = 4 (0.46%)
    - Not Applicable – n = 471 (54.45%)
    - Unknown – n = 16 (1.85%)
- Missingness after preprocessing: 256/865 (29.60%) represented true missing values (murmur present but location not recorded). In addition, 471 observations (54.50%) were classified as Not Applicable because no murmur was detected, and 16 observations (1.85%) were classified as Unknown because the parent variable (Clinical_Murmur) was missing.
- Modelling rationale: Included to investigate whether the anatomical location of a murmur provided additional discriminatory information beyond murmur presence alone. The site of maximal murmur intensity is a standard component of cardiac auscultation and may provide clinically relevant information regarding the underlying cardiac lesion. Explicit encoding of contextual categories preserved the hierarchical relationship with Clinical_Murmur while preventing structural absence from being conflated with data uncertainty.