IMPERIAL

MEng Individual Project

Imperial College London

Department of Computing

Deep Learning Pipeline for Automated Diagnosis of Fetal Congenital Cardiac Anomalies

Author:
Ishita Mahatme

Supervisor:
Dr. Bernhard Kainz

Second Marker:
Dr. Ben Glocker

June 17, 2024

Abstract

Congenital heart defects (CHDs) are the most prevalent birth anomalies that significantly impact infant health, with adverse outcomes such as disability and even death. However, prenatal detection in the UK remains low at 54%. We propose a deep-learning pipeline for robust image analysis from ultrasound data to enhance neonatal prognosis. We demonstrate the anatomical significance of the spine and introduce a novel blob-based method for its localisation. Our pipeline uses a plane detection model to extract cardiac views and performs semantic segmentation to predict the spine point. The spine serves as a landmark for automated extraction of established features like the cardiac axis and a novel metric - the angle between the aorta and duct vessels. We correlate these features with congenital heart diseases using logistic regression to effectively detect anomalies. Our pipeline achieves a notable accuracy of 90% in diagnosing Tetralogy of Fallot and 60% for Transposition of Great Arteries on held-out clinical data, surpassing the current clinical detection rate. These results emphasise the value of cardiac biomarkers and the efficacy of deep learning models for prompt detection of CHDs, leading to improved fetal outcomes.

Acknowledgements

I would like to thank my supervisor, Dr. Bernhard Kainz, for his invaluable guidance and unwavering support throughout this project. He directed me to the right resources and facilitated effective communication, both of which were instrumental for conducting my research. I am also grateful to my second marker, Dr. Ben Glocker, for his insightful input during the interim stage of the project. Under their mentorship, I have gained a wealth of knowledge and experience in the field of Deep Learning and Machine Learning for Imaging.

Special thanks to Dr. Thomas Day, Dr. Samuel Budd, and Mr. Robert Wright of the iFIND group for their expertise and contributions in various aspects of this project which have been crucial to its development.

I would also like to acknowledge my Personal Tutor, Dr. Maria Valera Espina, for her helpful advice throughout my four years at Imperial College London.

Furthermore, my heartfelt gratitude goes to my family for always believing in me and motivating me, especially my father, Girish, and my mother, Dr. Priya, who being a paediatrician has inspired my passion for this research.

Lastly, I extend my sincere appreciation to my friends, some of whom I have known since my first day at university. Their company has made this journey memorable.

Contents

Chapter 1

Introduction

1.1 Motivation

Congenital heart diseases represent the various anomalies present at birth that affect normal cardiac structures. They are a significant challenge in prenatal care as they are the most prevalent anomaly and a leading cause of perinatal mortality. In 2021, CHDs accounted for 4.9 deaths per 10,000 live births in the UK, with 3.2 occurring within the first month of life [1]. Timely diagnosis can enable prompt intervention, improving neonate prognosis and helping parents make informed decisions about their child's treatment [2].

The interpretation of a routine fetal ultrasound taken in the second trimester of pregnancy should allow clinicians to detect up to 80% CHDs prenatally [3]. Despite this, diagnosis occurs in only 54% of affected fetuses, contributing to 12% of all infant deaths [4]. This low detection rate could be due to technical challenges posed by variability in fetal position and morphology, which may obscure the accurate detection of congenital anomalies.

Advancements in Artificial Intelligence (AI), particularly Deep Learning (DL), are revolutionising the field of medical imaging by enabling thorough analysis of ultrasound scans. DL models can automatically learn relevant features and patterns from diverse medical images, significantly aiding radiologists in detecting and diagnosing various anomalies. DL-based early disease detection has led to superior patient care by optimising clinical processes and treatment pathways [5, 6]. However, despite these advancements, there has been limited application of deep learning for diagnosing congenital heart diseases, likely due to its lower prevalence and greater heterogeneity compared to other medical conditions.

In a developing fetus, the spine is one of the earliest structures and can serve as a landmark for extracting key measurements. One such metric, the cardiac axis, has been established to correlate with CHDs [7, 8, 9]. However, these studies rely on manual techniques for angle measurement, which are prone to intra and inter-observer variability.

This variability underscores the need for standardised methods to compute ultrasonographic markers and encourages the development of automated tools for ultrasound analysis. While there have been few deep-learning pipelines that have outperformed clinicians in CHD detection [10], they primarily focused on using ultrasound images as input to predict disease status, neglecting the potential of utilising parameters that can be extracted from these images. In our study, we aim to address this diagnostic gap, by proposing a fully automated DL-based approach to analyse biomarkers such as the cardiac axis from ultrasound scans for the early diagnosis of CHDs.

Fetal ultrasound scans typically capture images of the entire body, but when studying the heart, we want to focus specifically on views that depict cardiac structures. Extracting these views, however, necessitates a certain level of expertise from the sonographer and is a time-consuming process. We will streamline this by using a convolutional neural network to identify standard cardiac views from ultrasound scans automatically.

To compute features from these identified planes, we will evaluate machine learning techniques to identify the spine since we hypothesise that the location of the spine together with the identification of cardiac structures hold enough information for robust detection of disease. Semantic segmentation, a technique that delineates anatomical structures through pixel-level classification, shows promise in this regard. However, we only have access to limited annotated spine points. We further hypothesise that by converting these points into blobs and using them as ground truth masks to train a supervised semantic segmentation model, we can achieve approximate spine point localisation. Given that congenital defects are characterised by structural deformities, we will use spine and fetal structure labels to measure their relative sizes and use their centroids as reference points to study spatial configuration (Section 3.2).

After selecting relevant planes and extracted features, we can perform clinically valuable tasks like disease diagnosis. This can be framed as a binary classification problem, where the fetus is either diagnosed with CHD or classified as healthy. Since input ultrasound images may contain multiple cardiac views, we may detect multiple features per plane. In such cases, we will aggregate the features to provide a more holistic and robust estimate to the classifier. In our study, alongside the cardiac axis, we introduce a novel feature, the vascular angle (Section 3.2.3), and explore its efficacy in disease classification.

The ultimate goal of this project is to deploy an end-to-end deep-learning pipeline for the early detection of congenital heart diseases like hypoplastic left heart syndrome, transposition of the great arteries and tetralogy of Fallot. We hypothesise that integrating robust image analysis methods will enable the development of a comprehensive machine-learning solution for streamlined disease diagnosis. We will build this pipeline in a modular fashion, allowing for the flexibility to incorporate and interchange various detection, segmentation, and classification models. Figure 1.1 shows the flowchart of our proposed approach.

1.2 Clinical Collaboration

This project aims to further the advancements in fetal cardiac analysis using deep learning in collaboration with esteemed institutions such as St. Thomas Hospital London and King's College London, through the iFind [11] research group. We hope to deploy a pipeline for automatic analysis of ultrasound scans of fetal echocardiograms. We believe this will reduce the need for highly experienced sonographers, whose expertise demands significant time and resources for training, and facilitate early detection of potential congenital heart diseases. Prompt detection and intervention for CHDs has proven to dramatically reduce fetal mortality rates. Thus, the potential deployment of our model into the clinic provides a tangible motivation for this project and fuels our passion for this research.

1.3 Contributions

- **Development of an Automated Deep-Learning Pipeline:** We built and evaluated a comprehensive deep-learning pipeline for the early diagnosis of congenital

heart diseases, translating the AI concepts of multi-class classification, semantic segmentation, and binary classification into clinically valuable processes (Figure 3.1).

- **Improvement in Detection Rates:** Our pipeline significantly surpassed the current UK TOF detection rate of around 76.6%, achieving a remarkable 90.91% sensitivity and 88.89% specificity on real clinical data in diagnosing Tetralogy of Fallot across two disease subgroups (Section 4.3.2).

- **Introduction of Novel Features:** We utilised the less-explored 3VT and RVOT planes for TGA diagnosis and introduced a new feature, the vascular angle (Section 3.2), to study the spatial alignment of the aorta and duct. We also measured the sizes of the duct and aorta, naming it the *DA* ratio. Together, these features obtained 80.61% accuracy for TGA classification (Section 4.2.3). To our knowledge, both these features have not been previously correlated with CHDs.

- **Unique Spine Localisation Method:** We established a new method for localising the spine in ultrasound scans by converting spine points to blobs and then taking their centroid (Section 3.2.1). This automated the measurement of valuable biometric parameters like the cardiac axis, which uses the spine as a landmark.

- **Collaboration with iFind Group:** We containerized the code for our pipeline (Section 3.6.3) and provided it to the iFind Group as a scalable and efficient tool. The modular nature of this tool enables further refinement for various clinical applications in fetal cardiac analysis from ultrasound data.

1.4 Challenges

- **Data Limitations:** Deep learning algorithms use datasets which often contain thousands or even ten thousands of data points. We explore disease classification, facing the challenge of limited or no labelled data specific to the targeted fetal cardiac conditions. Annotation and labelling is a time-consuming and costly process, requiring the expertise of medical professionals. Hence, we have adopted transfer learning to overcome this constraint.

- **Data Privacy Concerns:** Our research involves medical image analysis of fetal ultrasound data acquired from a fetal cardiology clinic. Privacy is a major concern when dealing with sensitive patient data and the dataset we have used has been anonymised to protect the identity of the volunteer. We have addressed ethical concerns in Chapter 6.

- **Multiple Data Formats:** The medical images provided to us were of a wide range of formats including DICOM, PNG, NPZ, and NIfTI, with some files containing metadata that had to be incorporated to interpret its contents. We were able to handle this efficiently using SimpleITK, an open-source image analysis toolkit, which provides a comprehensive set of image readers and writers compatible with a wide array of medical image formats.

- **Clinical Adoption:** The deep learning algorithms used to develop models that will be integrated into clinics must be accurate and reliable enough to aid in making informed decisions. Thorough evaluation of the algorithms are essential, along with a means to provide reasoning behind the diagnosis [5].

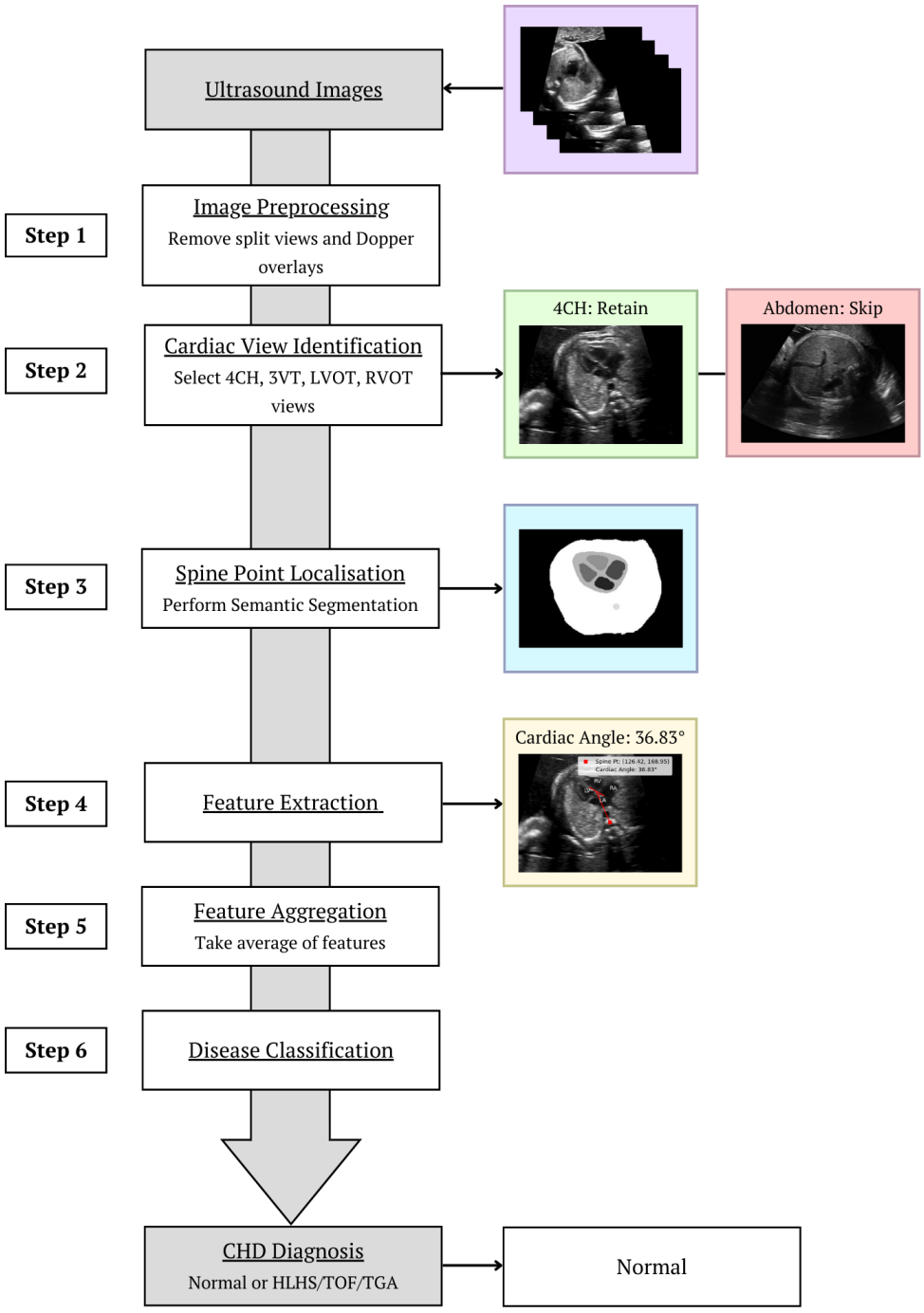


Figure 1.1: Flowchart of the Proposed Approach from Input Ultrasound Images to CHD Diagnosis. Steps 2, 3, and 4 are performed for each input image. The features extracted from all retained images are then aggregated to detect the disease.

Chapter 2

Background

Artificial intelligence is an umbrella term encompassing a broad array of technologies used to develop machines with human cognitive abilities. In medical settings, these technologies have been leveraged for diagnostic purposes with evidence of outperforming healthcare professionals.

This chapter highlights some of the requisite medical knowledge to understand the context of our research including the interpretation of the dataset, following which, we explain some of the fundamental concepts and methods that are revolutionising AI in healthcare.

2.1 Fetal Echocardiography

2.1.1 Medical Imaging

Medical imaging is defined to be the methods used in the visualisation of various tissues and organs of the human body. X-ray, ultrasound, MRI and CT scanners are common imaging modalities used to capture these visuals in the form of images or videos which can then be interpreted by trained physicians to diagnose underlying conditions, detect abnormalities and refer the patient to appropriate treatment. [12]

2.1.2 Ultrasound

An ultrasound (US), also referred to as a sonogram, is a medical test that generates real-time pictures or videos of the interior of the body using high-frequency sound waves. During pregnancy, a fetal ultrasound provides a means of monitoring the baby's health by analysing delicate structures such as the heart. It is often performed by a skilled sonographer using a transducer or probe which, when held in different positions, can capture different anatomical cross-sections of the heart.

In England, the NHS offers at least two ultrasound scans during gestation, which take place between 11 and 14 weeks and 18 and 21 weeks. As part of their Fetal Anomaly Screening Programme (FASP) [13], the second scan examines the fetus for 11 physical conditions including congenital heart diseases.

2.1.3 Congenital Heart Disease

Congenital heart disease represents the various anomalies present at birth that affect normal cardiac structures. The term "congenital" implies that the condition exists from birth and is either hereditary or acquired during development in the uterus. 1 in 100 babies

in the UK are born with a congenital heart disease, making it one of the most prevalent congenital defects [14].

- **Hypoplastic Left Heart Syndrome (HLHS):** HLHS is a complex congenital condition where the left side of the heart is underdeveloped, including the left ventricle, aorta, and mitral valve, significantly affecting the normal pumping of blood around the body. This condition causes lower-than-normal oxygen-saturation levels in neonates which, when untreated, can result in fatal outcomes within the first hours or days of life.[15]

- **Transposition of the Great Arteries (TGA):** As the name implies, TGA is a heart condition characterised by the abnormal switch in positions of the two primary arteries, the aorta and pulmonary artery, responsible for carrying blood away from the heart. This disrupts the normal blood flow, causing the circulation of oxygen-poor blood through the body. Thus, it poses a life-threatening risk, with a mortality rate of 50% if left untreated within the first month of infancy [16]. In the UK, prenatal diagnosis takes place in around 84.9% of affected infants.[17]

- **Tetralogy of Fallot (TOF):** TOF is a disease described collectively by four cardiac alterations. The first is pulmonary stenosis, a constriction of the pulmonary valve that restricts blood flow from the right ventricle to the lungs. This causes the right ventricle to work harder, causing right ventricular hypertrophy, a thickening of its muscle. Also, there is a large hole, referred to as the ventricular septal defect (VSD), in the wall separating the ventricles. Finally, an overriding aorta, abnormally positioned, which allows some deoxygenated blood from the right ventricle to mix with oxygenated blood from the left ventricle before reaching the body. While TOF is detected antenatally in 76.6% of cases [18], it can lead to cyanosis (a bluish appearance of the baby) during infancy due to insufficient oxygenation of the blood.[19]

2.1.4 Standard Views

Four-Chamber View

The Four-Chamber (4CH) view (Figure 2.1a) is the most common screening examination, obtained through a transverse scan of the thorax and helps visualise the four chambers of the prenatal heart through which blood circulates as follows[20]:

1. Right Atrium: It receives a mixture of oxygenated and deoxygenated blood.

2. Left Atrium: It receives oxygenated blood from the right atrium through a shunt (foramen ovale). The two atria are observed to be of the same size in a normal fetal heart.

3. Right Ventricle: It receives less-oxygenated blood from the right atrium and sends it to the lungs to be oxygenated, with the majority being shunted through the ductus arteriosus.

4. Left Ventricle: It receives oxygenated blood from the left atrium and passes it to the aorta. In a healthy fetus, the ventricles are roughly of equal size.

Three Vessels and Trachea View

The mediastinum is the central region within the thorax that lies between the cavities enclosing the lungs and holds vital structures such as the heart and its vessels, the trachea and a network of nerves.

The Three Vessels and Trachea (3VT) view (Figure 2.1d) is obtained through a transverse scan of the upper mediastinum, providing a simultaneous view of the spatial depiction of the arch formed between the aorta and duct along with their proximity to the trachea. It also allows for the identification of the superior vena cava, thus completing the depiction of all the major vessels in a developing fetus [21].

In a normal scan, the three vessels are arranged in a straight line in decreasing order of their diameter from left to right:

1. Pulmonary Trunk (PT): It is the most anterior vessel, carrying blood deficit of oxygen from the heart to the lungs. However, in fetuses, due to incomplete lung functionality, blood in the pulmonary artery is shunted into the aorta via the Ductus Arteriosus (Duct).
2. Aorta (Ao): It is an artery that lies in the centre and supplies the body with oxygenated blood.
3. Superior Vena Cava (SVC): It is a vein that is present posterior to the aorta and brings oxygen-poor blood from upper parts of the body into the heart.

Left Ventricular Outflow Tract View

The Left Ventricular Outflow Tract (LVOT) (Figure 2.1b) is a muscular channel in the left ventricle that transports blood outward and into the aorta. This view is acquired by angling the transducer anteriorly towards the fetal right shoulder from the 4CH plane. The aorta is visible in a normal heart, with its anterior wall extending from the interventricular septum.

Right Ventricular Outflow Tract View

The Right Ventricular Outflow Tract (RVOT) (Figure 2.1c) is the pathway through which blood flows out of the right ventricle and into the pulmonary trunk. In this view, the pulmonary artery, depicted to the left of the aorta, bifurcates into the Ductus Arteriosus and the right pulmonary artery. The SVC is located to the right of the aorta.

Figure 2.1 illustrates how these views contribute to understanding fetal circulation.

2.1.5 Cardiac Angle

The cardiac axis is one of the basic morphological features of the fetal heart which describes its rotational inclination within the thoracic cavity. [23] The thorax is divided into two equal halves: the left (L) and the right (R), by a conceptual line extending from the posterior spine to the anterior sternum (spinosternal line). Typically, the normal fetal heart is situated predominantly on the left side, and its apex is oriented to the left at an angle of 45°± 20° with respect to the longitudinal axis of the chest cavity. This angle is referred to as the "cardiac angle."[24]

2.2 Deep Learning

Machine learning is a sub-field of AI that allows computers to make data-guided predictions. These traditional algorithms, however, rely on structured data and human-guided pre-processing. This limitation paved the way for deep learning, a branch of machine learning, which unlocks the potential to harness unstructured data and automate feature extraction, opening the doors to exciting applications across various industries.

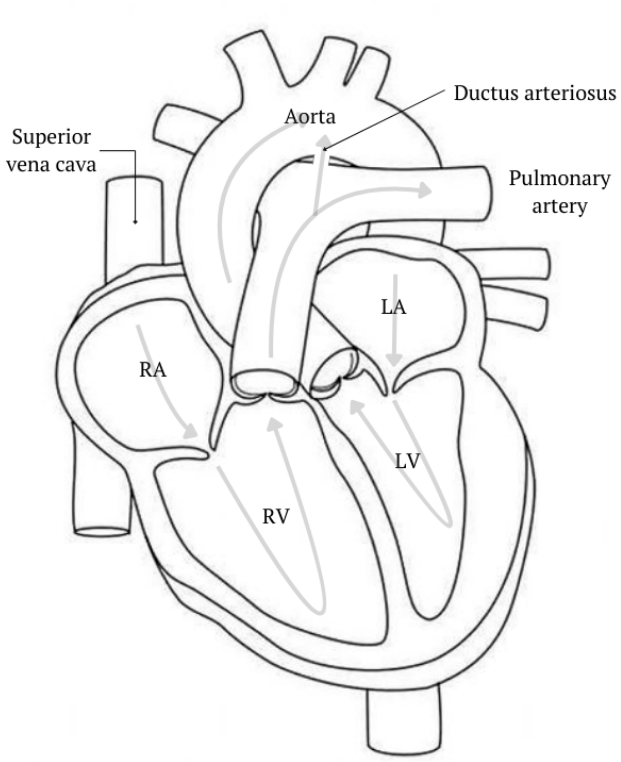
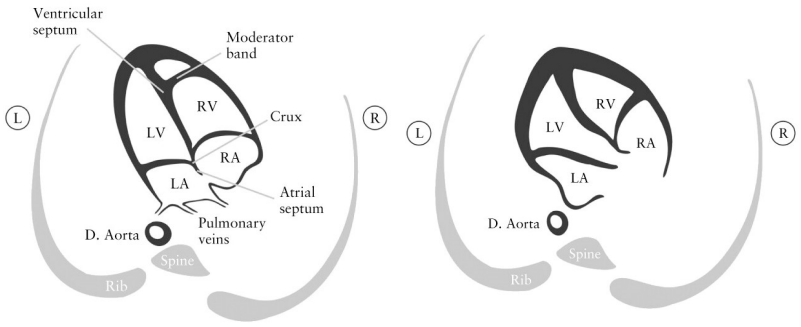
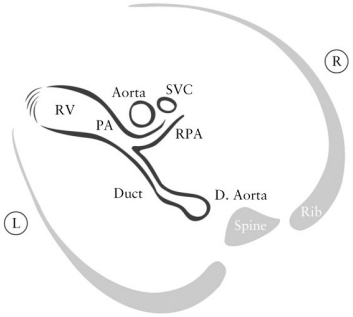


Figure 2.1: Labelled diagram of the fetal circulation showing the anatomical structures described under the four cardiac views: 4-CH, 3VT, LVOT and RVOT illustrated below. LA: left atrium; LV: left ventricle; RA: right atrium; RV: right ventricle;

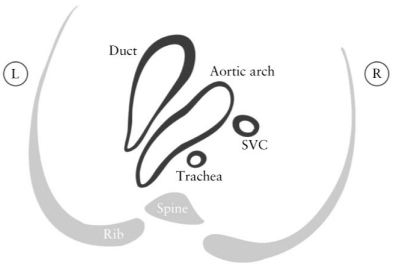


(a) 4CH View

(b) LVOT View



(c) RVOT View



(d) 3VT View

Figure 2.2: Labelled diagrams of standard views in a fetal echocardiogram. The 4CH view is obtained by an axial scan across the fetal chest. Tilting the probe from this view towards the fetal head sequentially reveals the other views: LVOT, RVOT, and 3VT. L: left; R: right; LA: left atrium; LV: left ventricle; RA: right atrium; RV: right ventricle; D. Aorta: descending aorta; PA: pulmonary artery; RPA: right pulmonary artery; SVC: superior vena cava; Source [22]

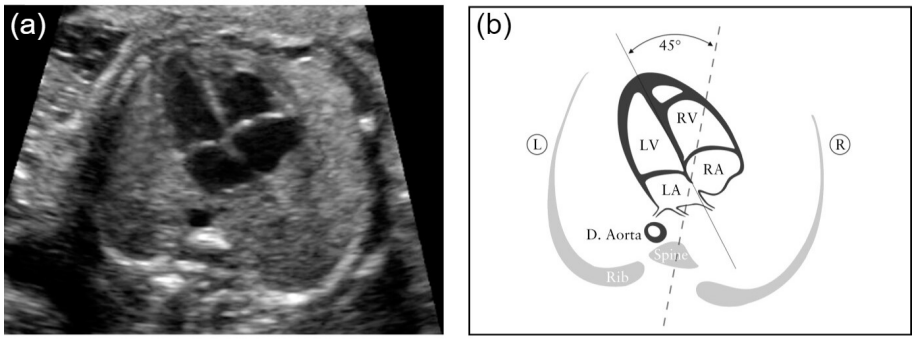


Figure 2.3: (a) Grayscale 4-CH view ultrasound image (b) Cardiac position and axis where LA: left atrium; LV: left ventricle; RA: right atrium; RV: right ventricle; dAo: descending aorta; Source [24]

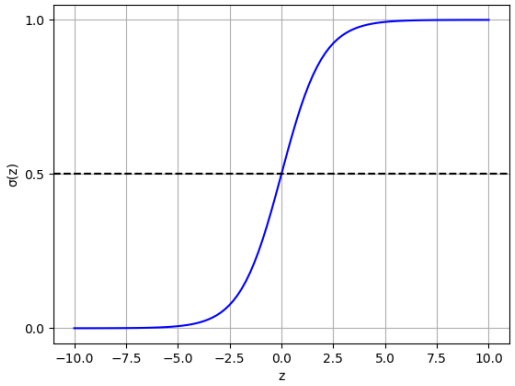


Figure 2.4: Sigmoid Function

2.2.1 Logistic Regression

Supervised learning is a sub-paradigm of machine learning which trains a model using labelled inputs along with their desired outputs. Logistic Regression is a supervised machine learning algorithm that studies relationships between independent variables (features) and a dependent variable (outcome). It uses the logistic (sigmoid) function to perform classification with range [0, 1], which is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where z denotes the sum of the products between features and model parameters:

$$z = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

For binary logistic regression, the output is dichotomous in nature i.e. it can have only two possible values - 1 (positive) or 0 (negative). Setting the threshold to 0.5, the output label is determined as:

$$\text{Output} = \begin{cases} 1 & \text{if } \sigma(z) \geq 0.5 \\ 0 & \text{if } \sigma(z) < 0.5 \end{cases}$$

2.2.2 Semantic Segmentation

Semantic Segmentation is a computer vision task that aims at performing per-pixel classification, grouping pixels belonging to the same class. This has largely benefited medical

image analysis by helping identify structures of anatomical interest. In this study, we employ this technique to generate label maps of 4-CH and 3VT US scans, which are then utilised for subsequent analysis.

2.2.3 Neural Networks

A neural network is a layered connection of neurons or nodes that mimics the intricacies of the human brain, forming the core building block of deep learning. The smallest individual unit of this network, the neuron, has the following components: [25]

- Input: These are the set of values or features from the dataset that the model receives for training purposes. They are often passed as numerical representations of images and other types of data in the form of pixel values.
- Weights: These are values associated with each feature to convey the influence that feature has on the final output.
- Transfer function: This is also known as the summation function and is responsible for introducing scalar multiplication between the inputs and their corresponding weights and computing their sum.
- Activation function: This performs a non-linear transformation of the weighted sum to enable the network to learn complex patterns.
- Bias: This shifts the result of the activation function by a constant amount.

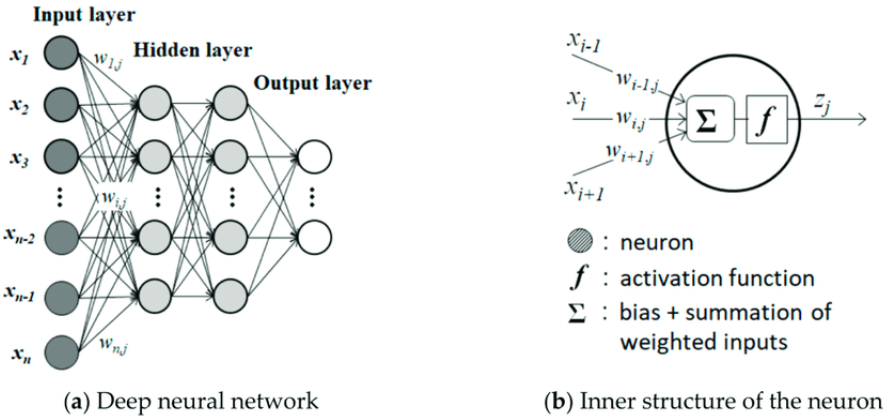
A layer is an aggregated collection of neurons and multiple layers can be arranged sequentially to create a multi-layer neural network. In deep learning, neural networks comprise three types of layers:

1. Input Layer: This is the first (conceptual) layer of the network and simply passes the raw data that we feed into the network to the latter layers without performing any computation.
2. Hidden Layer(s): This layer is present between the input and output layers and is often variable in number depending on the complexity of the task. Earlier hidden layers usually learn simpler features from the data whilst latter layers are able to identify more intricate patterns thus leading to hierarchical feature extraction.
3. Output Layer: This is the final layer of the network which returns the result or output of the task.

A wide variety of neural network architectures can be produced by experimenting with different numbers of layers, neurons, and types of activation functions used, providing it with the capability of approximating any desired function.

2.2.4 Convolutional Neural Networks

Convolutional neural network (CNN), a powerful tool in computer vision, is a widely used deep learning algorithm. Within healthcare, while medical imaging provides a means of early detection and diagnosis of diseases, it has shortcomings due to human error and lack of experienced technicians, which are overcome using CNNs. It has especially gained popularity due to its cutting-edge performance, at par with radiologists [27, 28]. These models are capable of directly learning relevant features from grid-like data and are comprised of three types of layers:



(a) Deep neural network (b) Inner structure of the neuron

Figure 2.5: Source [26]

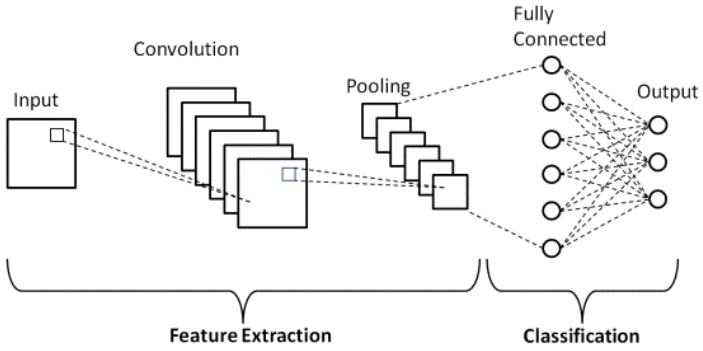


Figure 2.6: Basic CNN architecture; Source [29]

1. Convolutional Layer: This layer uses a filter or a kernel to extract features from the input image by performing a convolution operation. This operation produces a feature or activation map.
2. Pooling Layer: This layer reduces the computational cost by down-sampling the feature maps produced by the convolutional layer. This operation decreases the spatial dimensions of the convolved maps. An activation function is applied to its output to make it non-linear. The function essentially assigns an importance to each neuron, which tells the following layers whether that particular neuron is activated or not.
3. Fully Connected Layer: This layer operates on a flattened input where multidimensional data is converted into a 1-dimensional array and each neuron in this array is connected to all neurons of the next layer.

Their architecture consists of many convolutional layers, each followed by a pooling layer and together they perform feature extraction. Finally, there are numerous fully connected (FC) layers which classify the output.

A kernel is a small array of weights which slides over the input image and performs a dot product with the sub-region of the image it is currently covering and finally produces a single output pixel by computing the sum of the element-wise multiplication. It has several tunable hyperparameters, namely:[30]

- Kernel Size: This refers to the dimensions of the filter applied to the image. It plays an important role in determining the size of the region in the input image that

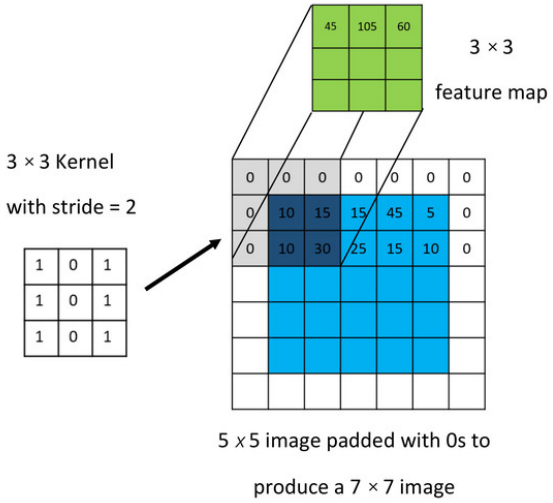


Figure 2.7: Convolving an input image padded with 0s using 3x3 kernel with stride = 2; Source [29]

corresponds to an output feature, a term coined as the receptive field of a neuron. Image characteristics can influence the size of the kernel, as sharp images often require smaller kernels to detect edges while blurry images need larger kernel sizes.

- Stride: This indicates the distance in pixels by which the window moves after a convolution or pooling task is performed.
- Padding: This is the process of adding pixels to the boundary of an image to preserve spatial dimension and retain information in the image after filters are applied.

These models are more computationally efficient than traditional neural networks as the use of convolutional layers and kernels enables parameter sharing which reduces the number of parameters required to be trained.

2.2.5 VGG-16

VGG-16 model, introduced by A. Zisserman and K. Simonyan of the Visual Geometry Group, is one such milestone in CNN architecture. It derives its name from the number of constituent layers (13 convolutional + 3 fully connected) and is an object detection and classification algorithm. The VGG's novel contribution was the idea of grouping layers into "blocks" to offer a more organised and modular approach to developing deeper networks.

2.2.6 U-Net

The U-Net is a prevalent deep learning model, specifically developed for BioMedical Image Segmentation. Its name is due to its distinctive U-shaped architecture which comprises a contracting and an expansive path interconnected via a bottleneck layer. The contracting path is made up of encoder layers of increasing depth that perform convolutional operations on the input to diminish its spatial resolution while capturing progressively abstract representations of the data. The bottleneck layer then applies a single convolution to the output of the last encoder layer and passes the resulting feature map to the expansive path. The expansive path works at expanding or increasing the spatial resolution of the feature map and reducing the number of channels through the use of upsampling layers. These layers are assisted by skip connections arising from the encoder that help identify and enhance the features within the image. The output is a semantic segmentation map.

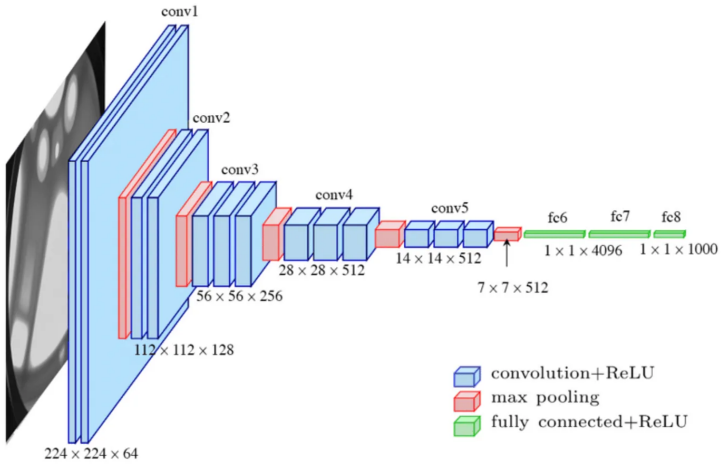


Figure 2.8: VGG-16 Architecture; Source [31]

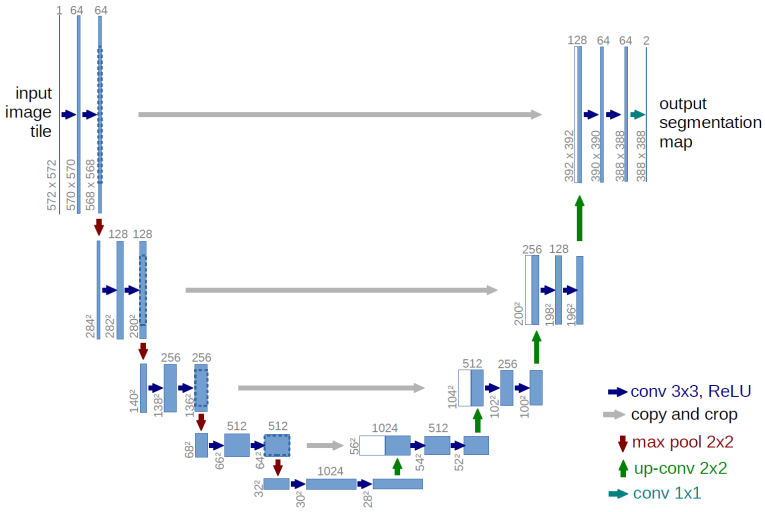


Figure 2.9: Original U-Net architecture; Source [32]

2.2.7 nnU-Net

In the biomedical domain, remarkable diversity is observed across image datasets due to variations in dimensionality, modalities and image sizes. To tackle the challenging problem of manually adapting to each task, nnU-Net was proposed as a semantic segmentation method that automatically sets-up a U-Net-based segmentation pipeline tailored to the intricacies of the provided dataset. The name nnU-Net or "no new U-Net" was coined as it achieved state-of-the-art performance across various tasks without the introduction of novel architectures, loss functions or training strategies. Instead, it replaced cumbersome manual tuning or empirical methods with systematic and efficient approaches.

nnU-Net offers comprehensive pre-processing, which includes z-score intensity normalisation, image and annotation resampling and diverse data-augmentation techniques such as random rotations, scaling, Gaussian noise and blurring to increase the size of the dataset. It is trained using fivefold cross-validation as it was designed with the assumption of absence of test data. Additionally, connected-component based post-processing is applied to eliminate isolated noise or artefacts while maintaining the segmentation accuracy. Lastly, nnU-Net trains an ensemble of 2D, 3D and 3D-Cascade U-Net configurations, to determine the best configuration that can then used to produce predictions for the test data.

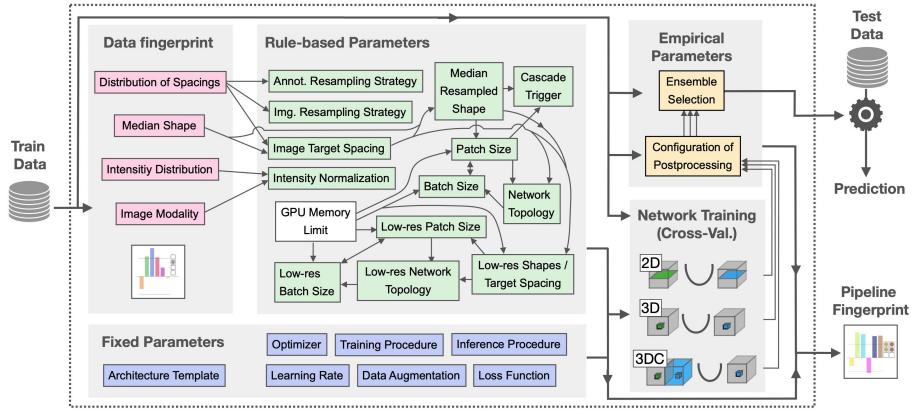


Figure 2.10: Proposed self-configuring pipeline of nnU-Net; Source [33]

2.3 Related Work

2.3.1 Traditional Approaches to Fetal Cardiac Analysis

Several studies published before 2020 have demonstrated the use of manual and semi-automatic approaches to find the correlation of the fetal cardiac axis with congenital diseases. One such noteworthy tool, FINE [34], is a semi-automatic software that leverages healthcare expertise to identify key anatomical landmarks. From this information, it constructs a virtual 3D map and calculates the cardiac axis for fetal echocardiography views. This axis has also been established as a reliable diagnostic indicator by [9, 35, 36] who have manually extracted this measure from the 4CH view.

However, these approaches rely heavily on the domain expertise of the sonographer and are prone to inconsistencies in image interpretation, which can lead to inter-operator variability. The widespread use of the 4CH view has likely been due to its ease of acquisition and has primarily benefitted the detection of defects like HLHS that are characterized by variations in the cardiac chambers. There has been limited exploration of vascular abnormalities for the assessment of conditions like TGA.

We aim to address this gap with a two-fold approach. Firstly, we introduce an automated pipeline to streamline cardiac axis measurement. Secondly, we leverage the understudied 3VT and RVOT views to delve deeper into vascular anatomy. We introduce a novel feature, the vascular angle, and investigate its potential as an indicator for TGA diagnosis.

2.3.2 Deep Learning Pipeline for Predictive Analysis of Risk Factors in Congenital Heart Disease

Pachiyannan et al. [37] implemented a Cardiac Deep Learning Model (CDLM) to determine the risk of prenatal mortality caused by CHD. Indicators such as maternal health history, which offers insights into environmental exposures during pregnancy as well as gestational age which reflects the development stage of various cardiac structures, were factored in to curate treatment recommendations tailored to each patient. The system analyses MRI images and performs segmentation on them to isolate the cardiac structures. From the segmented regions, it measures morphometric features which are used to identify potential causes of CHD-related complications, including death. The study employed a comprehensive array of datasets, including open-source clinical data, CHD datasets, and cardiac MRI scans to develop and evaluate the CDLM against existing predictive models. The system outperformed other frameworks with high accuracy in detecting CHD (over

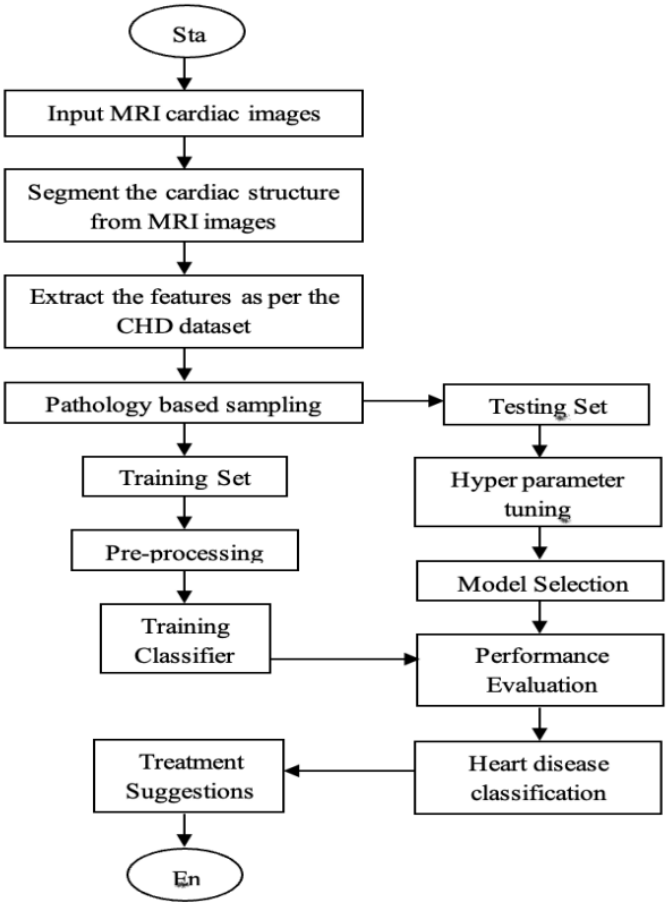


Figure 2.11: Workflow of proposed CDLM

90% sensitivity and specificity). The authors propose leveraging the CDLM's capabilities to develop more advanced models to flag high-risk cases and improve the survival of infants with CHD.

One pitfall we notice however is the reliance on MRI images. While powerful, these are significantly more expensive and resource-intensive compared to ultrasound, limiting cost-effectiveness and widespread adoption of the CDLM. Our study, in contrast, analyses available ultrasound data. While Athalye et al. [10] have presented a deep learning model that outperformed clinicians through view classification of ultrasound data, we will focus on analysing biomarkers, to establish their effectiveness in the early diagnosis of CHDs.

2.3.3 Automated Identification and Visualisation of Fetal Landmarks in Freehand Ultrasound

Baumgartner et al. [38] introduced a novel automated tool to assist operators in the complex tasks of identifying and interpreting anatomical structures during freehand fetal ultrasound examinations. Their proposed deep neural network architecture, SonoNet (Sonography Network), which builds upon the VGG16 model performs instantaneous frame detection and retrospective retrieval of fetal standard scan planes. The method uses confidence maps to localise the corresponding anatomical structures via bounding boxes under weak supervision. The reliance on only image-level labels for training, poses a significant advantage as it reduces the need and effort for labor-intensive manual annotations by sonographers.

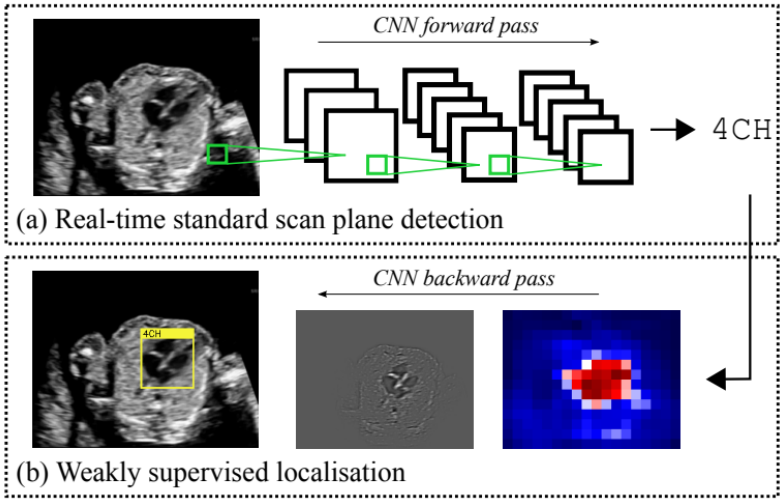


Figure 2.12: SonoNet Overview: (a) 2D fetal ultrasound video is processed in an automated fashion to identify and isolate fetal standard plane (b) if a standard view is detected (4 chamber view (4CH) in this case), a backward pass is performed through the network to localise the structure via bounding box; Source [38]

While SonoNet demonstrated exceptional performance for most fetal standard planes, it struggled to localise heart anatomies with its retrieval accuracy dropping to an average of 0.82. A contributing factor to this was the high similarity between the 3VT and RVOT views, which led SonoNet to misclassify them. Fortunately, since both planes depict the main vessels of the fetal heart, differentiating between them is not crucial for vessel biometrics calculations. Thus, in our study, we use SonoNet and SonoNext, a model that uses the same underlying architecture but has been tested in a real clinical setting.

2.3.4 Prior Collaboration with St. Thomas Hospital for CHD Detection using Segmentation

To promote the interpretability of automated diagnosis, Budd et. al [39] developed a novel extension of the Atlas-ISTN framework for neonatal detection of hypo-plastic left heart syndrome (HLHS). Unlike previous classification efforts which processed videos or numerous images, the model utilised a single 4-Chamber View (4CH) scan thus reducing computational costs. Further, the classifier used the area of each anatomical segment as a proportion of the whole, achieving state-of-the-art performance, comparable to manual annotations by experienced sonographers.

While this provides a promising approach for diagnosing HLHS, a subsequent study by Jakubowski [40] improved upon the segmentation of 4CH dataset by a 4.06% margin and also evaluated performance on the 3VT dataset. The research involved a comprehensive analysis of nine adaptations of the baseline U-Net architecture using the nnUnet framework. To address lack of substantial quantities of annotated image data, artificial images were fed into the model which improved segmentation accuracy and demonstrated exceptional accuracy in live clinical settings. Our study will further these efforts by exploring the potential of extracting features such as precise angles and anotomical measurements from the outflow tract views, LVOT and RVOT, in addition to 4CH and 3VT.

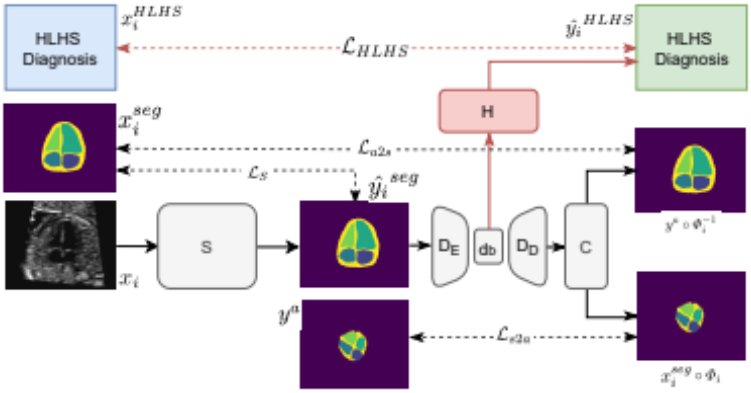


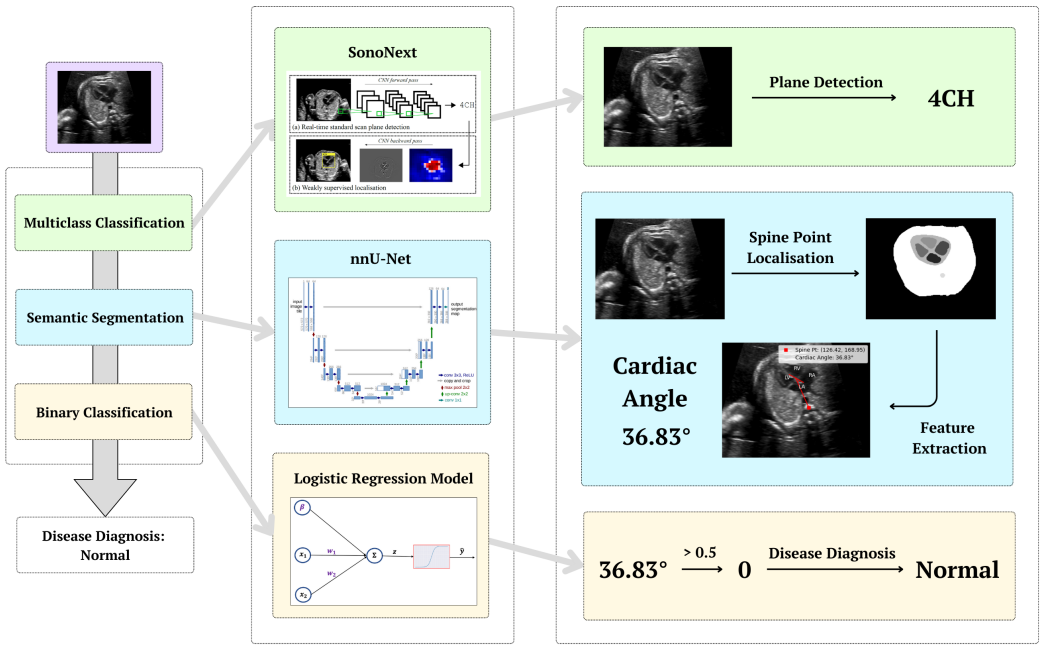
Figure 2.13: Overview of modified Atlas-ISTN with a disease prediction branch (H), Segmentation network (S), Atlas to image mapping module (D) and Transformation computation module (C); Source [39]

Chapter 3

Methods

In this chapter, we describe the design and implementation of the automated machine learning pipeline developed in this project. The proposed pipeline processes frames from ultrasound videos to detect standard planes, compute biometric measurements and predict the disease status of the fetus.

3.1 Pipeline Overview



(a) DL Tasks **(b) DL Models** **(c) Applications in the Pipeline**

Figure 3.1: Overview of proposed pipeline with constituent deep learning tasks, models and their applications in the pipeline. Here, we show a single pass of a 2D frame through the following stages of the pipeline for HLHS disease classification: (1) Plane Detection: A multi-class classification model, SonoNext, is used to identify fetal standard cardiac planes (Output: 4 chamber view (4CH)) (2) Spine Point Localisation: A 2-D U-Net architecture model, nnU-Net, is used to perform semantic segmentation on the cardiac scan for spine point localisation, and features such as the cardiac angle are geometrically extracted (Output: 36.83°) (3) Disease Diagnosis: A logistic regression model is used to perform binary disease classification from the input biometric features (Output: 0 or NORMAL); Model Architecture Sources: SonoNet [38], nnU-Net [33], Logistic Regression [41]

3.2 Cardiac Biometrics

Congenital heart diseases are often associated with deformities in the development and structure of fetal anatomy. To study these deviations from their normal arrangement, we aim to identify biometric parameters that can serve as features to predict disease status. We use segmentation masks to extract metrics that have either been medically shown to detect certain congenital heart diseases (CHDs) or those that we hypothesise could assist in their diagnosis. These masks allow us to delineate the boundaries of various structures within the fetal heart, thus facilitating the measurement of various biometric parameters.

3.2.1 Spine Point Localisation

Spine development is one of the earliest processes during pregnancy, occurring in the initial stages of gestation. A well-developed spine provides a stable and consistent reference structure within the fetal anatomy and is used by clinicians to study the arrangement of cardiac structures. Hence, we incorporate the spine label into the dataset.

The annotated data provided to us contains coordinates of the spine point for most images. We convert this spine point into a circle of radius 5 pixels using the Open Source Computer Vision Library (cv2) [42]. This blob is then passed as an additional label to our model. The transformation from a point into a blob is performed to account for potential variability in the location or annotation accuracy of the spine. Further, given that ground truth markings of structures are typically represented as regions rather than points, using a circular blob ensures consistency in data representation.

In instances where multiple spine blobs are predicted, we determine the final spine point by computing the centroid of the largest blob. We use this approach to ensure that only the most significant and likely accurate prediction is used for further analysis.

3.2.2 Cardiac Angle

As described in Section 2.1.5, the cardiac angle is defined as the angle between a line passing through the spine and sternum and another line passing along the interventricular septum. In our study, we draw one line from the spine to the centroid of the whole heart mask as a close approximation of the anteroposterior spinosternal line and another along the interventricular septum, by connecting the midpoints of the centroids of the atria and the ventricle masks. We record the angle between these two lines as the cardiac angle.

This ultrasonographic marker has been utilised in previous studies to investigate potential correlations with congenital heart diseases as well as non-cardiac and chromosomal defects [43, 9, 44]. While these research papers published promising results, the cardiac angle was measured manually, introducing the possibility of human error and bias. By automating the measurement process through deep learning, we anticipate improvements in diagnostic accuracy and reliability.

3.2.3 Vascular Angle

We introduce a new angle measurement for detecting heart defects, particularly those involving abnormalities of the two main vessels: the aorta and the pulmonary artery (or the ductus arteriosus in our study). CHDs such as TGA and TOF (Section 2.1.3) are characterised by deviations from normal arrangements of either or both vessels. Arunamata et al. [45] evaluated a DNN for detection of d-TGA (identified by the absence of crossing over of great arteries) and Arnaout et al. [46] proposed a deep learning pipeline

which uses cardiac axis to detect TOF. But, we hypothesise that the computation of our new feature, the angle between the great arteries, can also assist in the diagnosis of these diseases. Our study explores two different methods of calculating this from segmentation masks, as explained below.

Via Skeletons:

- Selecting Masks: We identify the aorta and duct by selecting the largest connected regions in each mask.
- Skeletonization: We apply the skeletonize function available in the skimage.morphology module of scikit-image to these masks. This function reduces the binary masks to their skeletal representations, which are 1-pixel-wide outlines that retain the original shape.
- Endpoint Selection: To obtain straight lines from the skeletons, we calculate the endpoints of these simplified shapes using a kernel. If more than two endpoints are found, we choose the pair farthest apart based on the Euclidean distance (straight-line distance).
- Angle Calculation: Finally, we record the angle between the lines drawn by connecting the selected endpoints as the vascular angle.

Via Centroids: We extract the centroids of the largest connected components for the aorta, duct, and spine. We then calculate the vascular angle, using the centroids of the aorta and duct as points, with the spine centroid serving as the vertex. This approach is chosen to address any issues that may be encountered with the previous approach, specifically the calculation of incorrect inverted or obtuse angles due to inaccurate masks. By using the spine point as a reference in our calculations and avoiding the use of skeletonization, we aim to achieve more accurate and reliable angle measurements.

3.2.4 DA Ratio and Distance between Centroids

In cases of TGA, a distinguishing finding is the observation of only two vessels in the 3VT view, rather than the usual three seen in healthy cases. To incorporate this clinical observation into our research and enhance disease diagnosis, we include additional features based on the arrangement and size of these vessels. We calculate the ratio, a measurement we abbreviate as *DA* ratio, between the areas, determined by the number of pixels, of the largest connected duct and aorta masks. We also measure the distance between the centroids of these two structures.

3.3 Model Architectures and Implementation

3.3.1 SonoNet and SonoNext

The first step in the pipeline is to detect the standard planes from the input frames. For this task, we use SonoNet [38] and SonoNext [47], two AI tools that can automatically detect 13 standard planes in ultrasound imaging. While SonoNet's validation primarily stemmed from retrospective data analysis, SonoNext was tested in a prospective randomized controlled trial. It achieved a sensitivity of 88.9% and specificity of 98%, providing robust evidence of its clinical contributions. Given our focus on detecting congenital heart diseases, we target the identification of the four fetal cardiac views: 4CH, 3VT, LVOT and RVOT. SonoNet and SonoNext are used as filters to extract these cardiac planes from the input image or video streams.

In the original paper, Baumgartner et al. propose 4 distinct model architectures (Figure 3.2) and conclude that SonoNet-32 achieves real-time execution and superior performance in detection and localisation. Based on this, we choose SonoNet-32 for our work and use its PyTorch implementation found at [48]. SonoNext, also based on the SonoNet-32 architecture, is provided to us in the ONNX format. Both models are pre-trained and are used for inference within our pipeline.

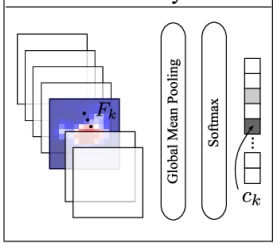
	Feature Extractors	Adaptation Layers	Classification Layer
SonoNet-64	2x[3x3x64/1], MP, 2x[3x3x128/1], MP, 3x[3x3x256/1], MP, 3x[3x3x512/1], MP, 3x[3x3x512/1]	1x[1x1x256/1], 1x[1x1xK]	
SonoNet-32	2x[3x3x32/1], MP, 2x[3x3x64/1], MP, 3x[3x3x128/1], MP, 3x[3x3x256/1], MP, 3x[3x3x256/1]	1x[1x1x128/1], 1x[1x1xK]	
SonoNet-16	2x[3x3x16/1], MP, 2x[3x3x32/1], MP, 3x[3x3x64/1], MP, 3x[3x3x128/1], MP, 3x[3x3x128/1]	1x[1x1x64/1], 1x[1x1xK]	
SmallNet	1x[7x7x32/2], MP, 1x[5x5x64/2], MP, 2x[3x3x128/1]	1x[1x1x64/1], 1x[1x1xK]	

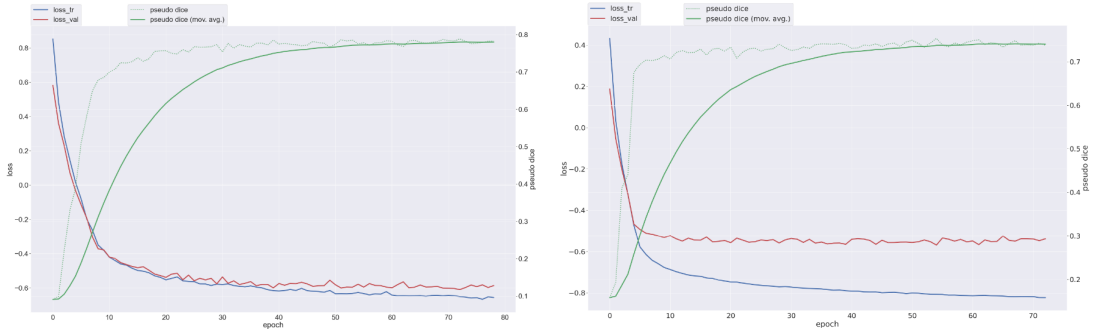
Figure 3.2: Architecture of three versions of SonoNet and SmallNet. SonoNet-64 takes its initial 13 layers from the VGG-16 architecture and has 64 kernels in its first convolutional layer. SonoNet-32 and SonoNet-16 are derived from SonoNet-64 architecture by halving and quartering the number of kernels respectively. SmallNet is a simpler architecture inspired by AlexNet. Source [38]

3.3.2 nnU-Net

We use the nnU-Net model defined in a previous project [40] for the semantic segmentation task with the hyperparameters shown in Table 3.1. Most of the changes we made were in the steps concerning the curation of the datasets as it involves extending the current model to accommodate the spine point and other new labels. We also updated some of the outdated code to version 2.5 by referring to the original nnU-Net GitHub repository [49]. The model in the paper was trained for a constant 150 number of epochs. We introduced early stopping as we noticed convergence in the accuracy around 75 epochs itself. After the training process, two models were retained: the best performing which was selected based on the highest-scoring Dice coefficient and the final model. We perform evaluation on the test set using the best-performing model.

Parameter	**Value**
Initial learning rate	0.001
Learning rate scheduler	PolyLRScheduler
Weight decay	3×10^{-5}
Optimizer	SGD
Momentum	0.99
Nesterov	True
Oversample foreground percent	0.33
Number of iterations per epoch	500
Number of validation iterations	50
Number of epochs	150
Early stopping patience	5
Batch size 4CH	16
Batch size 3VT	36

Table 3.1: Hyperparameter Values for nnUNet Architecture



(a) 4-CH (b) 3VT

Figure 3.3: Training Curves for nnUNet Model

3.3.3 Logistic Regression

For the final step in our disease diagnosis pipeline, we employ a Logistic Regression model from scikit-learn's linear_model library. Logistic regression is a well-established machine-learning technique for binary classification tasks. In our case, it aims to predict the presence or absence of a specific fetal CHD based on the extracted biometric features. The model takes a vector of features as input along with their ground truth disease labels. The output is a probability score between 0 and 1 for each data point to which we apply a threshold of 0.50. A score greater than 0.5 indicates the presence of disease (positive class, labelled as 1), while a score lesser than 0.5 indicates a lower likelihood (normal class, labelled as 0).

3.3.4 VGG-16

To assess the importance of biometric markers in CHD detection, we compare the performance of the above logistic regression model with a VGG-16 model imported from the PyTorch library's torchvision.models subpackage. It is pre-trained on the large-scale ImageNet [50] dataset, and we fine-tune it on our ultrasound image dataset for disease classification. We adapt the first convolutional layer to accept single-channel greyscale ultrasound images and replace the final layers with a classifier suitable for binary classification (Table 3.2). We apply the sigmoid activation function to the output layer to generate binary class labels, again using a threshold of 0.5.

3.4 Datasets

3.4.1 Spine Point Localisation

The nnU-Net uses supervised learning and requires ground truth labels during training. For this purpose, we utilise a comprehensive dataset of labelled ultrasound images of the 4-CH and 3VT view. An experienced sonographer manually extracted those images which provide optimal 4-CH and 3VT views from videos of ultrasound scans performed at 18–24 weeks of pregnancy that were recorded using the Toshiba Aplio i700, i800 and Philips EPIQ V7 G devices. These images were further annotated using LabelBox [51] platform by a fetal cardiologist and three experienced sonographers.

4-CH: In the 4-CH dataset, we use ground truth masks for the left atrium, right atrium, left ventricle, right ventricle, whole heart, spine, and thorax. There were a total of 1895 US images which were distributed into the training (1137 images), validation (379 images) and test (379 images) datasets. These scans varied widely in size, with dimensions ranging

Parameter	Value
Input channel adaptation	nn.Conv2d(1, 64, kernel_size=3, padding=1)
Classifier replacement	nn.Sequential(nn.Conv2d(512, 4096, kernel_size=7), nn.ReLU(inplace=True), nn.Dropout(), nn.Conv2d(4096, 4096, kernel_size=1), nn.ReLU(inplace=True), nn.Dropout(), nn.Conv2d(4096, 1, kernel_size=1))
Output activation	Sigmoid
Loss function	nn.BCELoss()
Optimizer	optim.Adam(model.parameters(), lr=0.0001)
Batch size	32

Table 3.2: Hyperparameter Values for VGG-16 Model

from 224x288 pixels to 1024x1280 pixels. Due to this, the earlier plotting of the spine as a fixed radius blob of 5 units was modified to utilise a threshold-based scaling approach. The provided spine points were first converted from world coordinates to voxel space using an affine transformation matrix associated with the segmentation map.

3VT: In the 3VT dataset, we focus on the labels for the aorta, ductus arteriosus, SVC, thorax and spine point. Our dataset consisted of a total of 892 US images and was further split into training (651 images, 605 with spine), validation (79 images, 70 with spine) and test (162 images, 150 with spine) set.

While only some images contained the ground truth spine point, we use those images which do not contain spine points as well, since we believe that the distribution of spine-positive and negative images across our datasets will prepare our model for real-world application by simulating diverse clinical scenarios.

3.4.2 Disease Detection

The logistic regression and VGG-16 models used for disease diagnosis require ground truth disease labels during training. We had disease status for CHDs such as HLHS and TGA and normal cases in the 4CH, RVOT and 3VT view. The data was derived from anonymised ultrasound scans of patients. The initial characters of the patient IDs indicated whether the fetus was healthy (NORM or FN short for Fetal Normal) or was diagnosed with HLHS (HLH) or TGA. The images were in PNG format and had dimensions of 288 x 224 pixels. While splitting patient images into training and test sets, we ensured no overlap of images from the patient in both sets.

HLHS Classification: We used two datasets for this task in 4-CH and RVOT views.

- **4-CH dataset:** This contained a total of 962 ultrasound images obtained from 8 patients. The main objective of this dataset was to predict the spine point, which would then be used to compute the cardiac angle. This was done using the best-performing model nnU-Net pre-trained on the 4-CH dataset. Notably, we found that out of 85 images belonging to one patient from the normal case, 84 lacked the spine point, hence we excluded this patient (FN006) from the dataset. This also

helped reduce the imbalance between the NORM and HLHS classes. The remaining 7 patients were split into the training and test datasets, as summarised in Table 3.3.

- **RVOT dataset:** This consisted of 1087 ultrasound images derived from 13 patients. Initially, images of only good quality which offer a clear visualisation of anatomical structures typically seen in an RVOT view were sampled. However, this resulted in an imbalanced dataset (1024 HLHS and 358 normal cases). To address this, medium and lower-quality scans were included to augment the normal dataset. Figure A.1 shows that vascular angles computed for a patient in three randomly selected images were consistent, irrespective of scan quality, supporting our decision to include these scans in the dataset. The final dataset was created by considering the number of scans per patient where the vascular angle could be computed and is summarised in Table 3.4. Vascular angles were computed using skeletonization.

Split	Patient ID	No. of Images
Train	NORM012	180
	NORM019	138
	HLH004	130
	HLH049	131
	HLH031	19
Test	NORM028	160
	HLH048	136

Table 3.3: HLHS Classification 4CH Dataset
Total HLHS images: 416
Total NORM images: 478
Total Train images: 598 (66.89%)
Total Test images: 296 (33.11%)

Split	Patient ID	No. of Images
Train	FN006	130
	FN008	251
	HLH004	2
	HLH016	34
	HLH029	170
	HLH031	43
	HLH033	8
	HLH039	6
	HLH049	140
	HLH060	13
Test	NORM012	94
	NORM019	14
	HLH048	155

Table 3.4: HLHS Classification: RVOT Dataset
Total HLHS images: 598
Total NORM images: 489
Total Train images: 797 (73.32%)
Total Test images: 290 (26.68%)

TGA Classification: The dataset used for this task contained a total of 939 ultrasound images captured in the 3VT view from 10 patients. To measure the biometric features, we first segmented the images using the best-performing nnU-Net model pre-trained on 3VT images. We constructed balanced datasets to mitigate misclassification resulting from imbalanced datasets, by ensuring an equal number of images (400 for training and 200 for testing) from each class for which vascular angles could be computed. The dataset was then split into a training set and test set as summarised in Table 3.5. Vascular angles were computed using centroids and additional features such as *DA* ratio and distance between centroids were measured.

3.4.3 Pipeline Evaluation

Finally, we utilised a specialised dataset obtained through a tertiary health screening program to evaluate our machine learning pipeline. This screening was conducted after referrals from initial screenings and involved focused cardiac ultrasound examinations at King's College Hospital and Evelina London Children's Hospital. This sensitive dataset has been

Split	Patient ID	No. of Images
Train	FN008	148
	NORM012	270
	NORM019	2
	TGA002	200
	TGA009	49
	TGA027	155
Test	FN006	34
	NORM028	15
	TGA033	52
	TGA037	14

Table 3.5: TGA Classification Dataset (3VT plane)
Total TGA images: 470
Total NORM images: 469
Total Train images: 824 (87.75%)
Total Test images: 115 (12.25%)

ethically approved as part of the iFind3 study. To protect patient privacy, folder names were formatted as "patientID_diseaseID," where "patientID" was an anonymised identifier and "diseaseID" was a code for the diagnosed CHD. Table 3.6 summarises the disease IDs used to test our pipeline.

Preprocessing: The files were provided to us in the DICOM format. Upon inspection, we noticed that some files contained artefacts that may mislead the model, impacting its segmentation performance. As a result, we removed frames containing split views and colour Doppler overlays. Then, while running the pipeline, we leveraged SonoNet to retain only the cardiac views. In this step, SonoNet also returned a saliency map of the detected (plane) class which we used to guide image-cropping. We extracted the centroid of this saliency map and cropped the image to a centred region with dimensions 50% of the original width and height. This choice was made to make sure the spine point was visible in the final image. The cropping ensured that the model prioritised the most relevant anatomical features in each view, and also removed irrelevant information like ultrasound controls often present in scans.

ID	CHD
4	TGA with no VSD
5	TGA with VSD
14	TOF with left aortic arch
15	TOF with right aortic arch
41, 42	NC

Table 3.6: Summary of Disease IDs. VSD: Ventricular Septal Defect;

3.5 Metrics

3.5.1 Distance between Spine Points

We did not consider DICE score to be a useful indicator of the performance of the semantic segmentation model on spine point localisation. Instead, we employed distance metrics to gain insights into the spatial accuracy of the predictions in comparison to the

ground truth. Each metric (mean, standard deviation, maximum and minimum) reports the distances in millimetres, between the centroids of the largest predicted spine blob and the corresponding ground truth spine blob across the dataset, accounting for voxel sizes of (1.0mm, 1.0mm) after image processing.

Additionally, we generated confusion matrices to visualise the segmentation performance of the model, categorising image pairs based on specific rules. These rules include:

- True Positives (TPs): These represent the cases where the spine point was correctly predicted for both images. Euclidean distance was computed to measure the distance between the ground truth spine point and the largest predicted spine point.
- True Negatives (TNs): These represent the cases where there was no spine point predicted, and there was no spine point in the ground truth mask. For these pair of images, the distance between the spine points was recorded as 0.
- False Positives (FPs): These represent the cases where the spine point was predicted, but there was no spine point in the ground truth mask. These pairs of images were skipped from the distance calculation.
- False Negatives (FNs): These represent the cases where the spine point was not predicted, but there was a spine point in the ground truth mask. These pairs of images were skipped from the distance calculation.

3.5.2 Accuracy

Accuracy is the most common metric that is a measure of how often the model's predictions align with the actual disease status. It represents the proportion of cases where the model correctly classifies the presence (TP) or absence of the disease (TN) as a fraction of the total number of predictions (TP + TN + FP + FN), given by the formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

3.5.3 Precision

Precision is a measure of the quality of positive cases identified by the model, representing the proportion of true positives (TP) among all instances the model predicted as positive (TP + FP).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3.5.4 Sensitivity/Recall and Specificity

Sensitivity and specificity are two crucial metrics used to evaluate the performance of a disease classification model. They offer complementary insights into the model's ability to correctly identify both healthy and diseased cases. Sensitivity or recall measures the ability of the model to correctly identify individuals diagnosed positive for the disease while specificity measures the ability to correctly identify normal cases.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

3.5.5 F1-Score or DICE

F1-Score, also known as the DICE coefficient, is a metric that balances precision and recall by computing the harmonic mean of the two. It evaluates the overall accuracy of the model by considering both false positives and false negatives. It is particularly useful when there is a class imbalance in the data. The formula is given by:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.5.6 McNemar Test

McNemar Test is a statistical hypothesis test used to determine if there are significant differences in the performance of two classifiers when evaluated on the same data. For binary classification tasks, it specifically checks whether there is a systematic difference in labels predicted by both models.

3.5.7 Mann–Whitney U Test

The Mann–Whitney U Test (or the Wilcoxon Rank Sum Test) is a non-parametric statistical hypothesis test used to compare the distributions of two independent groups. This test is applicable when the dependent variable is continuous and does not follow a normal distribution. It is useful for studying statistical differences in features computed for normal and disease groups, thereby supporting their feasibility for binary classification tasks.

3.6 Technical Details

3.6.1 Pipeline Implementation

We set up a virtual environment to isolate the project dependencies and used Python v3.10 as our coding language. We trained two nnU-Net models on normal 4CH and 3VT datasets respectively. By default, nnUNet trains configurations using a 5-fold cross-validation approach. We utilised fold 0 for our training process. These models were saved as pth files. Several metrics from Scikit-Learn (v1.2) were used to assess the performance of various binary classification models. The final logistic regression models chosen were pickled into binary streams.

To efficiently handle the medical data provided in DICOM, PNG, and NIfTI formats for the various tasks, we employed SimpleITK, an open-source image analysis toolkit. This toolkit provides a comprehensive set of image readers and writers compatible with a wide array of medical image formats.

To build the pipeline, we created an instance of SonoNext and unpickled the binary classifiers. We also instantiated an object of the nnUNetPredictor class for the 4CH and 3VT view with the weights of our saved nnU-Net models. We first used SonoNext to extract cardiac views from the input folder of DICOM files. Each file either represented a single image or a collection of images. When a cardiac view was detected, we used its saliency map to crop the image to a small region of interest. The 14x18 saliency map was resized using the cv2 library to match the image dimensions and guide the cropping process.

For the subsequent tasks of feature extraction and disease diagnosis, we leveraged several popular machine-learning libraries. We visualised the cardiac and vascular angles

using Matplotlib (v3.7) and used scipy.ndimage.label function from SciPy (v1.13) for identifying the largest connected component. Finally, to filter and aggregate the results, we used NumPy (v1.24) for efficient numerical computations.

The modular design of our pipeline, which first instantiates the different models and then separates the tasks of plane detection, feature extraction, and disease diagnosis into distinct functions, ensures scalability and flexibility. The structured approach also allows for easy updates, maintenance, and integration of additional components or functionalities as needed.

3.6.2 Hardware

The nnU-Net models in our study were trained on a 24 GB NVIDIA RTX GPU with a total training time of around 1 hour 30 minutes each with early stopping. For inference within the pipeline, which involves tasks like plane detection and dense segmentation, we used a remotely connected 12 GB NVIDIA TITAN Xp GPU.

3.6.3 Deployment

We packaged our pipeline into a Docker [52] image for Fraiya [53], a start-up company associated with the iFind group. A Docker image is a compact, self-contained, and executable software package that includes all essential libraries and dependent components to run an application within a container. Using Docker isolates our application from the host system, allowing users to deploy the pipeline seamlessly across their local machines, virtual machines, or cloud environments without encountering compatibility issues. This simplifies deployment, enhances portability, and supports scalability, making it ideal for collaborative projects.

Docker images are built from Dockerfiles, which define the steps to create the image. Our Dockerfile includes dependencies specified in the 'requirements.txt' file, the nnU-Net for model execution, and a script to set environment variables and initiate the pipeline. The Docker image is designed to accept runtime arguments, allowing users to specify input data paths, output directories, and the acronym of the CHD for diagnosis ('hlhs', 'tof' or 'tga'). Ultrasound scans can be provided in various formats such as folders containing DICOM files, individual PNG images, or a single MP4 video. This ensures compatibility with diverse imaging modalities and data sources. Note that the presence of split images or Doppler overlays will impact the performance of the pipeline. Therefore, it is preferable that these elements are absent from the input data.

Our pipeline requires GPU acceleration for efficient processing. During development, we utilised Docker Engine (v23.0) with a 24 GB NVIDIA TITAN RTX GPU to build and test our image.

All code developed for this project, including the saved models and implementation of the pipeline, can be found at `https://github.com/ish2002/Masters_HLHS_Unet`
Due to the nature of our dataset, we have made this GitHub repository private.

Chapter 4

Evaluation

In this chapter, we present the results obtained from various experiments performed to form the pipeline, both qualitatively and quantitatively.

To assess the accuracy and reliability of our proposed blob-based spine localisation method, we calculate several distance metrics for the semantic segmentation tasks on the 4-CH and 3VT datasets using nnU-Net. Typically, segmentation maps and their associated Dice scores are used to assess segmentation accuracy. However, our method focuses on identifying key anatomical landmarks through the centroids of the largest connected component rather than full segmentations. Therefore, distance metrics serve as alternative measures to quantify the proximity of our localised points to their true anatomical counterparts.

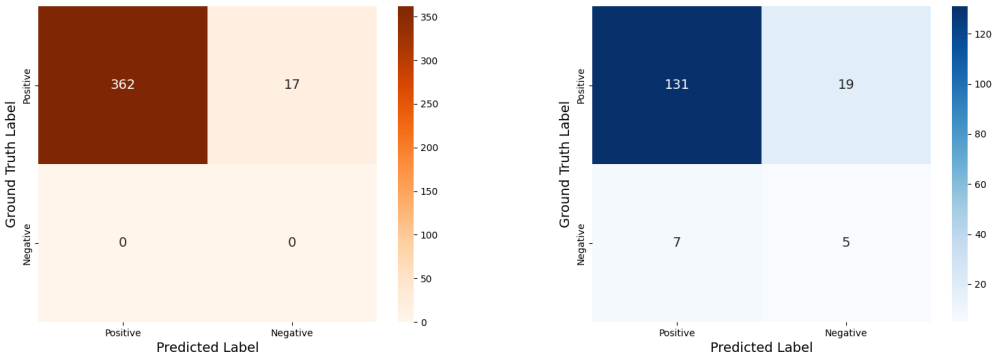
For disease diagnosis, we use patient data with known ground truth disease statuses for HLHS and TGA, comparing the accuracy, sensitivity and specificity of different classifiers. Initially, we evaluate two classifiers for HLHS, a logistic regression model and a pre-trained VGG-16 model. As our results indicate that the logistic regression model outperforms VGG-16, we extend the use of logistic regression to the TGA task. We then compare its performance when trained on different combinations of features. For all classifiers, we use the Mann-Whitney U test to examine correlation between the angles of the two distributions.

Once the models for spine localisation and disease diagnosis are finalised, we proceed to select a plane detection model. We use a small subset of patients with known ground truth plane and disease labels, and run our pipeline to evaluate the consistency of SonoNet and SonoNext. We prioritise consistency over accuracy as it is more suggestive of reliable and predictable performance in identifying the correct cardiac view across varying scan qualities and clinical settings.

Finally, combining all the chosen models, we build the final pipeline as shown in Figure 3.1. To validate our hypothesis whether deep learning based robust image analysis methods improve the early detection rates for CHDs, we demonstrate its performance on the diagnosis of CHDs such as Tetralogy of Fallot and Transposition of the Great Arteries.

4.1 Spine Point Localisation

For the task of spine point localisation, the analysis of confusion matrices (Figure 4.1) revealed that both models achieved high accuracy in spine point detection, 95.5% for 4CH and 83.9% for 3VT. The mean and standard deviation distance metrics (Table 4.1) were lower for the 3VT view, indicating superior localisation performance.



(a) 4-CH (b) 3VT

Figure 4.1: Confusion Matrices illustrating the segmentation accuracy of spine point predictions

Dataset	Mean Dist. ± SD (mm)	Max Dist. (mm)	Min Dist. (mm)
4-CH (362/379)	10.911 ± 12.507	136.618	0.061
3VT (136/162)	4.817 ± 6.252	34.684	0.0

Table 4.1: Summary of Ground Truth and Predicted Spine Point Distance Metrics. Brackets indicate the number of images included in the calculation of the distance metrics.

The high number of true positives for both views (4CH and 3VT) demonstrated the nnU-Net model's ability to correctly predict spine points in most cases, thus confirming our hypothesis of using blobs to achieve spine point localisation. The 3VT view achieved lower mean and standard deviation distances, indicating more accurate and consistent predictions. However, that could have been partly because we had fewer images for this view (162 compared to 379 for 4-CH), which may have reduced variability and enhanced model performance. It also achieved a minimum distance of 0 mm because of our evaluation rule for true negative pairs.

Further, both views had false negatives, as the model sometimes confused the spine with other white specs found on the ultrasound scans, likely due to similar visual features, noise, or resolution and contrast issues. This primarily contributed to the maximum distances reported in the table and can be visualised in the figures below.

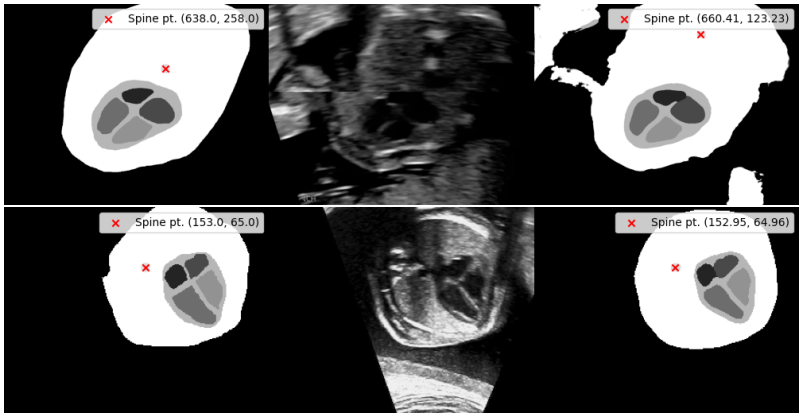


Figure 4.2: Examples of spine localisation for 4-CH dataset. Top: maximum distance (136.618 mm), Bottom: minimum distance (0.061 mm) (ground truth / image / prediction)

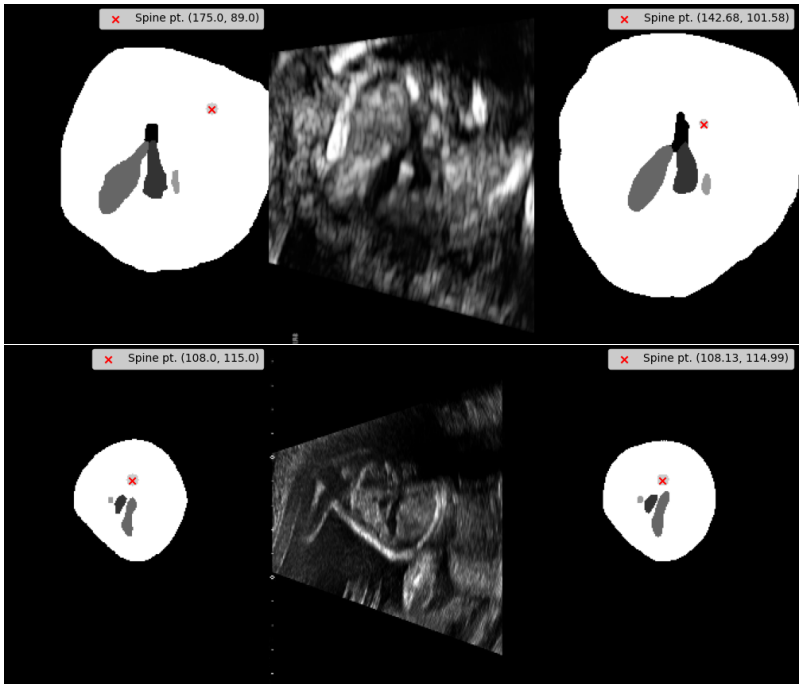


Figure 4.3: Examples of spine localisation for 3VT dataset. Top: maximum distance (34.684 mm), Bottom: minimum (non-zero) distance (0.127 mm) (ground truth / image / prediction)

4.2 Disease Diagnosis

4.2.1 HLHS Classification using 4CH dataset

We drew a box plot to visualise the cardiac angle distribution for NORM and HLHS cases. From Figure 4.4, it is evident that the median cardiac angle for normal hearts (NORM) was lower and the distribution had lesser variability than that for hearts with HLHS (HLH). The Mann-Whitney U test performed at a significance threshold of 0.05, revealed an extremely significant difference between the cardiac angle measurements of normal fetuses and those with HLHS with p-value $<<$ 0.0001 ($= 4.977 \times 10^{-108}$). This result indicates that the cardiac angle is a highly reliable diagnostic indicator for detecting HLHS.

Patient ID	Cardiac Angle Range	Mean ± SD
NORM012 (180/180)	24.7° to 45.6°	32.5 ± 5.3°
NORM019 (123/138)	28.0° to 40.8°	34.5 ± 3.0°
NORM028 (160/160)	17.4° to 26.1°	21.8 ± 2.1°
HLH004 (130/130)	23.8° to 48.3°	39.4 ± 6.2°
HLH031 (12/19)	36.3° to 59.4°	48.3 ± 6.1°
HLH048 (136/136)	3.8° to 162.6°	50.4 ± 13.9°
HLH049 (131/131)	35.4° to 71.4°	50.5 ± 9.3°

Table 4.2: Cardiac Angles Computed per Patient
Images in which the spine was not detected in the corresponding segmentation mask were skipped i.e. not evaluated. Brackets indicate the number of images evaluated per patient.

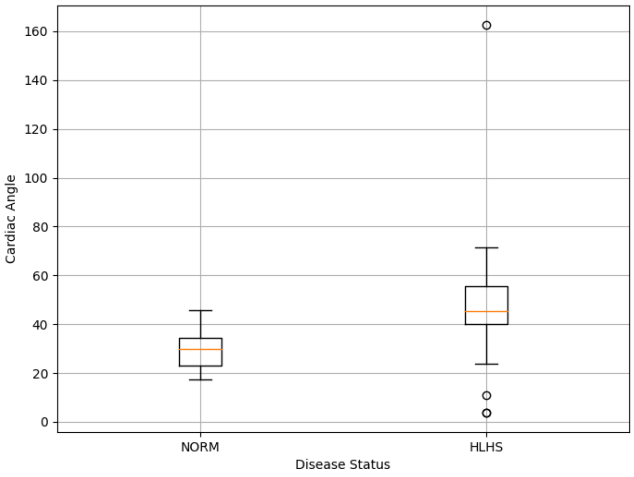


Figure 4.4: Box Plot of Cardiac Angle Distribution

To apply the identified significant difference in cardiac angles for disease diagnosis, we compared two machine-learning models on our dataset for HLHS disease classification: a Logistic Regression model and a VGG-16 model. Both models were passed the same training data, which was shuffled and divided into training and validation sets for the VGG-16 model. It was likely due to the higher number of NORM (318) images than HLHS (280) images in the training set, that both models achieved a higher classification accuracy for patient NORM028. The higher overall accuracy and per-patient accuracies of the logistic regression model suggested that it outperformed the VGG-16 model for the HLHS disease classification task.

Patient/Accuracy	**Logistic Regression Model**	**VGG-16 Model**
Overall (296)	**95.27%**	84.12%
NORM028 (160)	**100%**	85%
HLH048 (136)	**89.71%**	83.09%

Table 4.3: Results for HLHS Classification. Brackets indicate the number of images. The test set includes samples from only two patients: NORM028 and HLH048. Hence, values reported for NORM028 correspond to the Specificity (True Negative rate) of the models, and values reported for HLH048 correspond to the Sensitivity (True Positive rate) of the models.

We used the McNemar test (at significance level of α=0.05) to compare the performance of the logistic regression model and VGG-16 model under the following hypotheses:

- H0: The two classifiers have similar performance.
- H1: There is a difference in performance between the two classifiers.

The test revealed a significant difference in classification outcomes between the two models as the p-value ($p = 8.699 \times 10^{-6}$) was much smaller than α. Thus, we reject the null hypothesis and conclude that there is a significant difference in performance between the two classifiers. This supports the selection of the logistic regression model for HLHS classification in the pipeline.

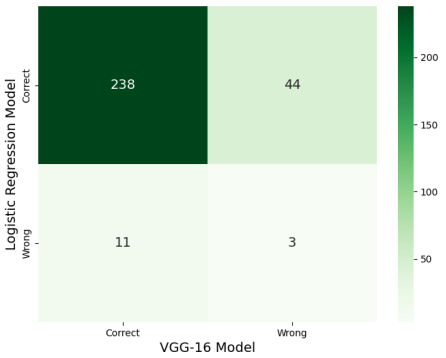


Figure 4.5: Contingency Matrix for McNemar's Test

The high accuracy of 95.27% achieved by the logistic regression model reinforced the finding from the Mann-Whitney U test, indicating a significant decision boundary between HLHS and NORM and further supporting the reliability of using the cardiac angle for HLHS diagnosis. Surprisingly, we noticed that both the lower bound of NORM (around 25°) and the upper bound of HLHS (around 55°) fell within the clinically expected range of 45°±20° that was shown in Figure 2.3. We suppose this could be attributed to differences in the orientation and transformation of the scans in our dataset, which affected the appearance and measurement of the cardiac angles.

Note that the model used for inference was pre-trained on normal cases. Hence, segmentation masks predicted by the model for normal patients appeared satisfactory (Figure A.2) but some anomalies were observed in cases with HLHS (Figure A.3). Despite these anomalies, we proceeded with the computation of cardiac angles between the spino-sternal line and interventricular septum, as these landmarks were still discernible in most cases.

4.2.2 HLHS Classification using RVOT dataset

The box plot of vascular angle distribution for NORM and HLHS cases shown in Figure 4.6 provided interesting insights. There was only a minimal difference of 0.88° between the median vascular angle for normal hearts (21.95°) and for hearts with HLHS (22.83°), indicating similarity in the vascular angles computed for the two groups. The Mann-Whitney U test revealed a p-value of 0.192 at a significance threshold of 0.05, further supporting that there is no statistically significant difference between the vascular angles of normal and HLHS hearts for the RVOT plane.

We can also see that there were quite a few outliers for both NORM and HLHS cases, with the number and range being more for HLHS. In some cases, the vascular angle was also computed incorrectly, as shown in Figure 4.7. These inaccuracies were either due to an inverted or an obtuse vascular angle being measured due to an irregular direction of the aorta and duct skeletons.

To evaluate the performance on the task of HLHS disease classification on our dataset, we trained a logistic regression model using the vascular angles computed for the training set as the input to classify outputs as 1 (HLHS) or 0 (NORM). The model achieved an accuracy of 59.4%, with a precision and recall of 59.4% and 100%, respectively (Table 4.4). However, a specificity of 0% indicated that the model incorrectly classified all negative cases as positive. Due to this misclassification, we did not use this particular model trained on RVOT data for HLHS classification.

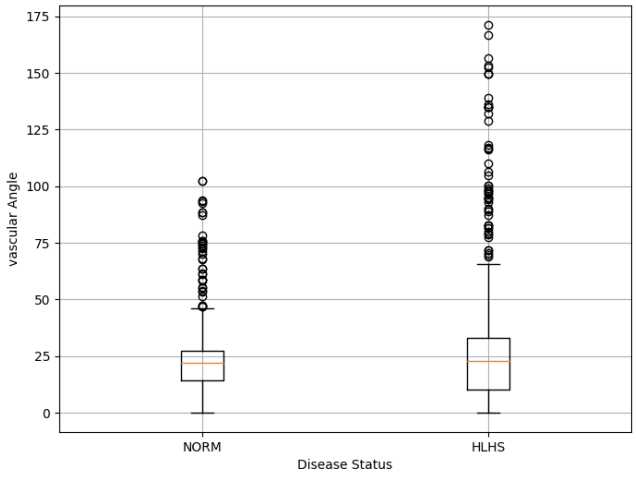


Figure 4.6: Box Plot of Vascular Angle Distribution (RVOT plane).
Median Vascular Angle for NORM cases: 21.95°
Median Vascular Angle for HLHS cases: 22.83°

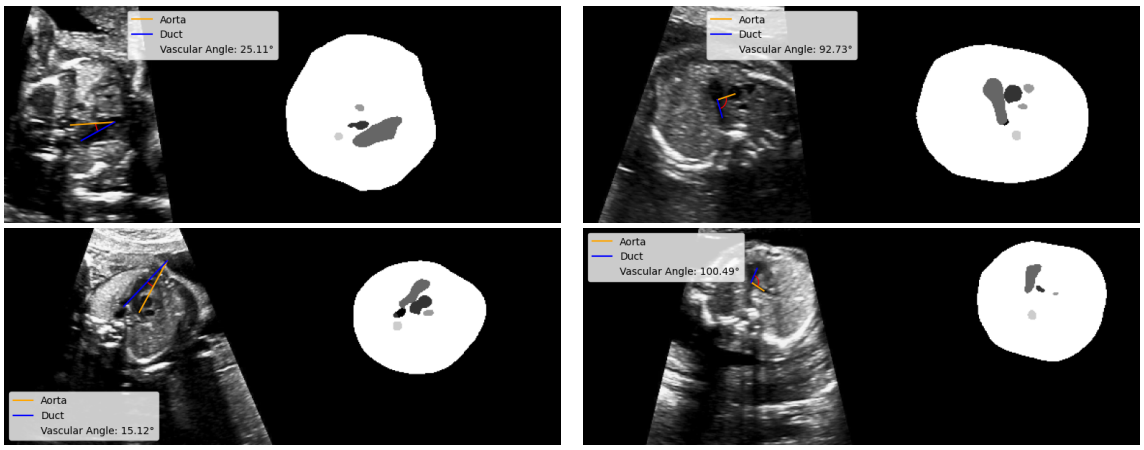


Figure 4.7: Examples of images with incorrect vascular angles. Top: train images. Bottom: test images. Left: images with inverted vascular angles measured in opposite direction. Right: images with obtuse vascular angles due to the orientation of the duct skeleton.

Metric	**Value**
Accuracy	59.4%
Precision	59.4%
Sensitivity	**100%**
Specificity	0%

Table 4.4: Performance of model trained on RVOT data for HLHS Classification.

The low accuracy highlighted the need for robust feature selection and a balanced dataset. Overlapping box plots indicated limited discriminative power of the vascular angle alone, and the slight dataset imbalance, with more HLHS cases, may have biased predictions. Additionally, the RVOT view's limited focus on left heart structures, primarily affected by HLHS, likely contributed to misclassification and low specificity. Thus, underscoring clinical relevance of selecting appropriate imaging modalities for accurate classification.

4.2.3 TGA Classification using 3VT dataset

For TGA classification, we computed the vascular angle using the centroid approach, which incorporates the spine as its vertex. We anticipated that this method would yield more accurate angles and eliminate the issues of inverted and obtuse angles observed in the previous experiment.

From Figure 4.8, we can see that the box plots for NORM and TGA had minimal overlap, suggesting that most cases were distinguishable based on their angles, making this a valuable feature for classification. The Mann-Whitney U test confirmed the significance of our findings, returning a p-value of 2.095×10^{-85}.

The median angle for NORM cases (20.09°) was significantly higher than that for TGA cases (14.43°). Interestingly, the outliers for NORM seemed to fall within the whiskers of TGA, suggesting that some normal cases exhibited characteristics similar to those observed in unhealthy fetuses. Figure 4.9 shows one scenario in which this was observed.

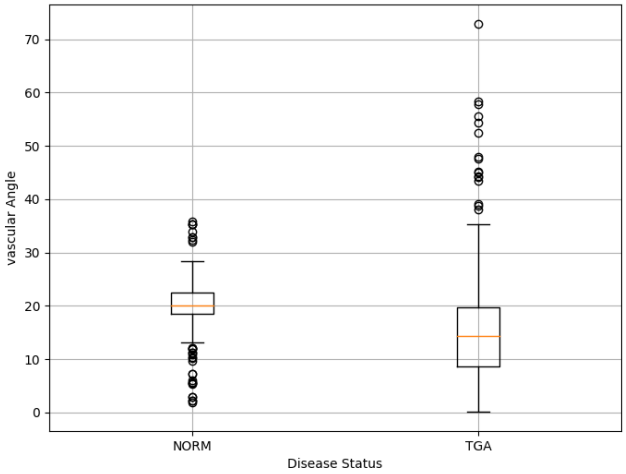


Figure 4.8: Box Plot of Vascular Angle Distribution (3VT plane).

To apply clinical knowledge in feature selection, we introduced the *DA* ratio and distance between centroids in addition to the vascular angle. The findings of a previous study [54] had revealed that a similar metric, the pulmonary artery/aorta ratio, correlates with congenital outflow tract anomalies. To test this, we evaluated the performance of four classifiers for detecting TGA: one using all features (vascular angle, *DA* ratio, and distance between centroids), one using the vascular angle and *DA* ratio, and two others using either the vascular angle or the *DA* ratio alone.

As shown in Table 4.5, the classifier trained on *DA* ratio performed well across all metrics with high results for Accuracy, Precision, and F1-Score. This aligned with the findings of the previous study. But the classifier trained on the vascular angle and *DA* ratio, outperformed this, achieving the highest accuracy, precision, and F1-score. These superior results obtained by using the vascular angle in conjunction with the *DA* ratio highlighted its potential as a novel feature. Therefore, we selected this classifier as the final model for detecting TGA in our pipeline.

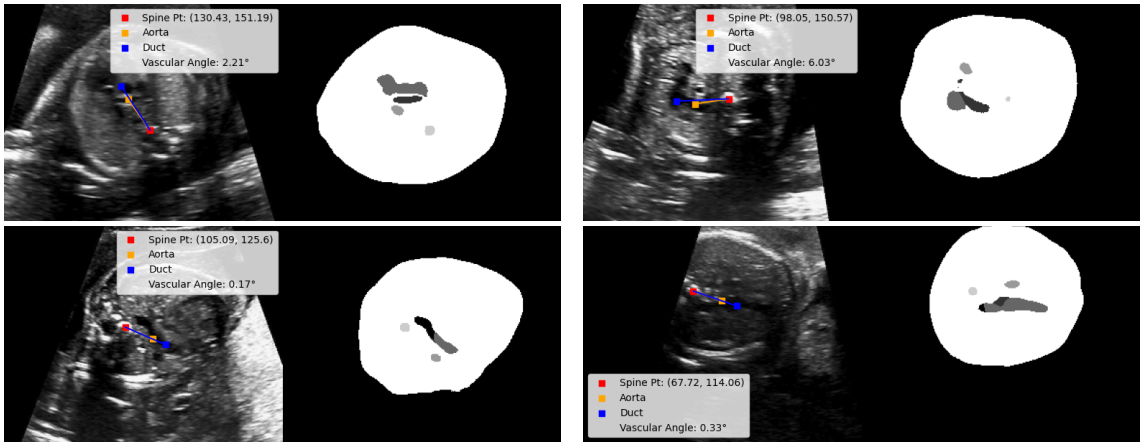


Figure 4.9: Examples of images for which small acute vascular angles were computed. Top: NORM outliers in the box plot. Bottom: TGA cases present within the whiskers. These images illustrate that significant overlap between the aorta and duct masks results in small angles. This occurs in TGA cases due to the aortic arch being positioned beneath the pulmonary trunk or duct.

Metrics/Features	**All Features**	**Angle + Ratio**	**Vascular Angle**	***DA* Ratio**
Accuracy	61.22%	**80.61%**	58.16%	74.49%
Precision	65.71%	**94.12%**	55.40%	92.86%
Sensitivity	46.94%	65.30%	**83.67%**	53.06%
F1-Score	54.76%	**77.10%**	66.67%	67.53%

Table 4.5: Results for TGA Classification.

4.3 Pipeline Performance

To finalise the pipeline, we assessed the performance of both SonoNet and SonoNext at the task of plane detection. This involved running the entire pipeline on a few patients with known ground truth disease statuses to determine the optimal choice. Finally, employing the chosen models, we conducted a comprehensive evaluation of our machine-learning pipeline using specialised patient data sourced from hospitals on a random selection of patients.

For this evaluation, we made a few assumptions based on findings from the previous experiments. For HLHS or TOF classification, we used 4CH and LVOT planes, and for TGA or TOF classification we used 3VT and RVOT planes. This was done due to their morphological similarity as both the LVOT and RVOT views show anatomies which are found in 4CH and 3VT respectively and can thus be used to approximate their features. Further, we noticed that although the boxes representing chosen classifiers did not overlap, there were instances where the whiskers of NORM and HLHS/TGA overlapped, potentially leading to a blurred decision boundary and misclassification. Due to this, we only considered features whose values fell within the interquartile ranges of their distribution. The images were preprocessed to remove split views and Doppler overlays.

4.3.1 Plane Detection: SonoNet vs SonoNext

This experiment specifically aimed at plane detection, hence we have only displayed results for 3 patients in Table 4.6. The dataset is formatted similarly to that used for HLHS

classification, but images from these patients were not used to train or test the classifier, thus constituting a held-out dataset. Detailed results for SonoNet, involving FN/NORM patients as well as all unique planes detected, can be found in Appendix B. For this experiment, we did not crop the images using the saliency map and utilised only 4CH views to compute the cardiac angle, as the ground truth plane was known to be 4CH.

Patient	**Quality**	**SonoNet**			**SonoNext**		
		Planes	**Angle**	**Status**	**Planes**	**Angle**	**Status**
HLH007	Good (141/162)	43 4CH 7 LVOT	40.32°	**True**	75 4CH 2 LVOT	51.46°	**True**
	Medium (79/129)	68 4CH 35 LVOT	36.57°	False	1 4CH 13 LVOT	7.51°	False
	Bad (37/105)	22 4CH 9 LVOT	41.23°	**True**	1 4CH 15 LVOT	37.84°	False
HLH016	Good (0/4)	0 4CH 0 LVOT	N/A	N/A	0 4CH 2 LVOT	N/A	N/A
	Medium (1/78)	0 4CH 0 LVOT	N/A	N/A	0 4CH 1 LVOT	N/A	N/A
	Bad (21/828)	0 4CH 0 LVOT	N/A	N/A	41 4CH 6 LVOT	N/A	N/A
HLH029	Good (1/5)	0 4CH 0 LVOT	N/A	N/A	2 4CH 0 LVOT	21.04°	False
	Medium (3/12)	0 4CH 0 LVOT	N/A	N/A	4 4CH 0 LVOT	36.08°	False
	Bad (0/1)	0 4CH 1 LVOT	N/A	N/A	1 4CH 0 LVOT	N/A	N/A

Table 4.6: Plane Detection Results for SonoNet and SonoNext. 'Quality' denotes image quality. Brackets show the number of images with computed cardiac angles. N/A denotes cases with no detected 4CH views or no computed cardiac angles for identified 4CH views.

The average cardiac angle was computed as the mean of all angles measured for the detected 4CH planes, weighted by confidence scores returned by the model for each plane detection. We used this approach because the confidence score, produced by the softmax layer of the models, represented their certainty that a given frame was indeed a 4CH plane. We assumed that higher confidence scores would likely indicate clearer anatomical features, thus producing more accurate cardiac angle measurements. Therefore, we weighted the angles with their respective confidence scores, to prioritise measurements that the model was more certain about. The formula is as follows:

$$\bar{\theta} = \frac{\sum_{i=1}^{n} w_i \theta_i}{\sum_{i=1}^{n} w_i}$$

where $\bar{\theta}$ is the weighted average of the cardiac angles, n is the number of 4CH planes, θ_i is the cardiac angle for the i-th 4CH plane, and w_i is the confidence score for the i-th 4CH plane.

Logging the unique planes detected highlighted that, despite SonoNext demonstrating lower accuracy in HLHS detection, with 1 true positive and 4 false negatives, it consistently identified 4CH and LVOT planes even in medium and low-quality images. Although

LVOT is not the ground truth plane, it still provides views of the four chambers found in 4CH, which are required to compute the cardiac angle. This suggests that SonoNext is more reliable for the task of plane detection of cardiac views. Thus, we selected SonoNext for plane detection within our pipeline.

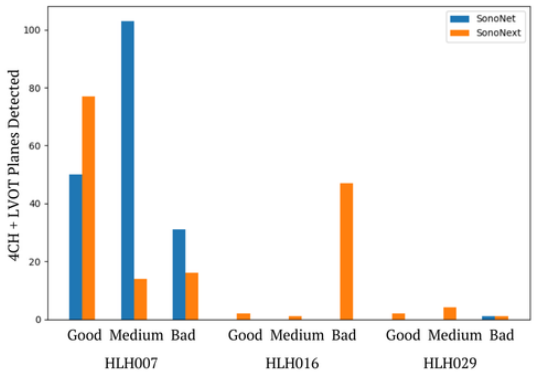


Figure 4.10: Bar chart to showcase consistent 4CH and LVOT plane detections by SonoNext.

4.3.2 TOF Diagnosis

We surprisingly performed a successful diagnosis of TOF using our HLHS classifier in the pipeline, achieving a high accuracy of 90% (Table 4.12). This indicated that an abnormal cardiac angle value could aid in diagnosing both diseases. Table 4.7 presents the results we obtained for both normal fetuses (41 or 42) and those diagnosed positive for Tetralogy of Fallot with either a left (14) or right (15) aortic arch.

Patient ID	Angles within IQR	Avg. Cardiac Angle	TOF Status
349_14	229/252	43.59°	**True**
357_14	418/465	41.02°	**True**
364_14	147/151	51.07°	**True**
372_14	168/213	38.97°	**True**
388_14	441/505	40.15°	**True**
383_14	9/9	62.31°	**True**
399_15	101/118	46.31°	**True**
411_15	225/239	52.59°	**True**
415_15	2/2	52.37°	**True**
413_15	272/285	51.32°	**True**
417_15	848/896	37.65°	False
1099_41	9/10	37.26°	**False**
1110_41	185/210	28.02°	**False**
1374_41	129/130	31.45°	**False**
1454_41	363/391	26.51°	**False**
2333_42	199/200	37.78°	**False**
2378_42	37/38	20.79°	**False**
2011_42	6/6	99.75°	True
2022_42	15/17	32.93°	**False**
2048_42	1249/1322	30.89°	**False**

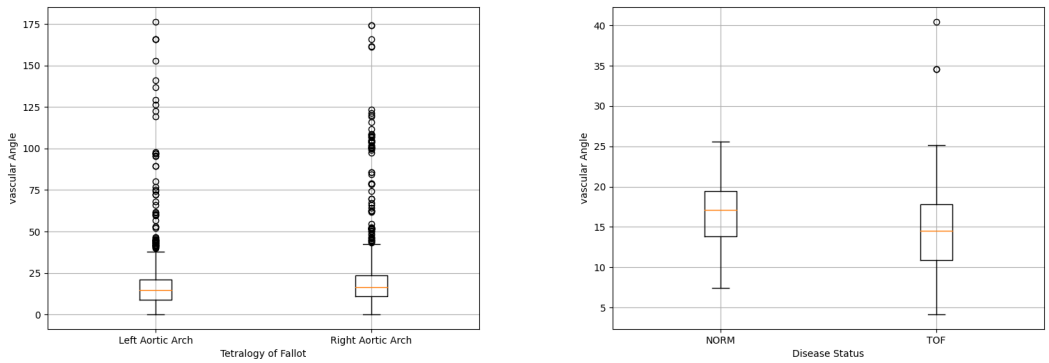
Table 4.7: Results for TOF Classification using HLHS Classifier.

Split	Patient ID	No. of Images
Train	1374_41	317
	1454_41	237
	349_14	104
	372_14	215
	384_14	125
	411_15	38
Test	2378_42	73
	2011_42	27
	415_15	83
	419_15	3

Table 4.8: TOF Classification Dataset
Total TOF images: 568
Total NORM images: 654
Total Train images: 1036 (84.78%)
Total Test images: 186 (15.22%)

Using this same dataset, we attempted to train a TOF classifier on the 3VT and RVOT views. Unlike the HLHS and TGA classifiers, which were trained on images carefully selected by an experienced sonographer, this TOF classifier would rely on planes extracted by SonoNext. Hence, we anticipated that its performance would be influenced by the accuracy of SonoNext's plane detection. To mitigate this potential issue, we extracted vascular angles for all identified planes and retained only those angles that fell within the interquartile range. This approach aimed to protect the classifier from being trained on outliers or incorrectly computed angles, with the hope of enhancing its reliability.

Next, we wanted to study similarities between the two TOF disease groups we had, TOF with left aortic arch and TOF with right aortic arch. Figure 4.11a showed a large overlap of the interquartile ranges of both distributions, thus suggesting similarity in their angle values. The Mann-Whitney test also supported this observation, returning a p-value of 0.294, indicating no significant difference between the angle distributions of the two TOF subgroups. Due to this, we merged both TOF subgroups into a single class labelled TOF (or 1) and passed it to our classifier. The splits for the train and test cases are shown in Table 4.8.



(a) TOF with Left and Right Aortic Arch

(b) NORM and TOF cases

Figure 4.11: Box Plots of Vascular Angle Distribution for two scenarios.

We first trained a logistic regression model. However, due to its limited performance, we explored more sophisticated binary classifiers available in the sklearn library, known for their ability to capture non-linear relationships and intricate patterns in data. Despite these introductions, all four classifiers ultimately exhibited low performance, with the highest accuracy of 57% achieved by the SVM algorithm.

Table 4.9: Classifier Performance

Classifier	**Accuracy (%)**
Logistic Regression	50.54%
Random Forest	54.84%
Support Vector Machine	**57%**
Neural Network	53.23%

This discovery was particularly noteworthy as it contrasted with our previous experiment on HLHS classification using RVOT, where clinical relevance was not a factor. In the case of TOF, a disease involving vessel abnormalities, especially pulmonary artery narrowing, we expected the vascular angle to offer valuable insights. However, our analysis, supported by the box plots of NORM and TOF cases (Figure 4.11b), revealed that the vascular angle possessed limited discriminative power in TOF detection.

4.3.3 TGA Diagnosis

We performed TGA diagnosis with our selected TGA classifier in the pipeline, achieving an accuracy of 60% (Table 4.12). Table 4.10 presents the results we obtained for both normal fetuses (41 or 42) and those diagnosed positive for Transposition of the Great Arteries with either no VSD (4) or having a VSD (5).

Patient ID	**Angles within IQR**	**Avg. Feature**		**TGA Status**
		Angle	**Ratio**	
091_4	213/234	22.05	5.4	**True**
127_4	7/7	48.87	1.99	False
101_4	189/200	12.61	2.44	False
115_4	110/118	18.2	4.09	**True**
123_4	1/1	4.29	1.42	**True**
147_4	2/2	4.4	3.1	**True**
157_5	3/3	35.47	5.27	False
163_5	2/2	19.2	9.95	**True**
173_5	3/4	13.69	4.65	**True**
169_5	2/2	29.06	2.38	False
176_5	3/3	55.1	4.16	False
172_5	13/13	17.43	9.65	**True**
158_5	3/3	28.44	1.06	False
1374_41	536/625	25.25	2.57	**False**
1454_41	449/474	20.18	4.43	True
2333_42	71/74	17.6	1.38	**False**
2378_42	145/1457	11.69	1.26	**False**
2011_42	48/52	22.09	1.977	**False**
2022_42	18/21	18.96	2.38	**False**
2048_42	2162/2491	25.97	6.19	True

Table 4.10: Results for TGA Classification.

4.3.4 Summary Statistics

Tables 4.11 and 4.12 summarise the results obtained by the pipeline from the previous sections for patients with normal hearts, and those diagnosed with TOF, TGA and their subgroups.

Case	n	Pipeline decision correct (n)	%
Normal	16	13	81.25%
TOF	11	10	90.9%
left aortic arch	6	6	**100%**
right aortic arch	5	4	80%
TGA	13	7	53.85%
no VSD	6	4	66.67%
with VSD	7	3	42.86%

Table 4.11: Pipeline decision compared to ground truth for CHDs and their subgroups.

CHD	Accuracy	Precision	Sensitivity	Specificity
TOF	**90%**	**90.91%**	**90.91%**	**88.89%**
TGA	60%	77.78%	53.85%	71.43%
Overall	75%	85%	70.83%	81.25%

Table 4.12: Performance of pipeline on detection of TOF, TGA and overall.

Our pipeline showcased a commendable level of accuracy (90%) and precision (90.91%) for the detection of Tetralogy of Fallot (TOF). Additionally, the sensitivity and specificity metrics, at 90.91% and 88.89% respectively, further affirmed the reliability of our pipeline for TOF diagnosis.

All TOF cases with left aortic arch were classified correctly. Since the classification was performed using a model trained and tested exclusively on HLHS cases, this underscores the versatility and effectiveness of our automated method of spine point localisation and cardiac angle computation as valuable diagnostic indicators for both Tetralogy of Fallot and Hypoplastic Left Heart Syndrome.

On the other hand, the outcomes for Transposition of Great Arteries (TGA) detection revealed a lower level of performance, with an accuracy of 60% and precision of 77.78%. The sensitivity and specificity metrics, at 53.85% and 71.43% respectively, also indicated a comparatively weaker performance in correctly identifying TGA cases. The low sensitivity of 53.85% specifically revealed that for approximately half of the individuals who had the disease, our pipeline produced a negative result, classifying the fetuses as healthy. TGA with VSD proved to be the most challenging subgroup for classification.

The disparity in performance between TOF and TGA detection may have stemmed from various factors. Firstly, the choice of features might have influenced the performance, as some features might have been more indicative of TGA over others. Moreover, the quality of the images utilised for detection could have significantly affected various stages of the pipeline, with clearer and more defined images resulting in more accurate diagnoses.

During the training of our TGA classifier, the 3VT images exhibited a clear demarcation of the constituent anatomies in the majority of cases, which enabled more accurate feature computation. However, during the pipeline evaluation, the data comprised 3VT

and RVOT images detected by SonoNext, which did not consistently depict a distinct delineation. This inconsistency might have impacted the computation of the *DA* ratio and, in some instances, even the vascular angle, which during the aggregation process, could have led to an accumulation of inaccuracies and consequently resulted in incorrect diagnoses.

Furthermore, there were fewer 3VT and RVOT images available for each randomly selected patient in the evaluation dataset, with most having less than 10 images. This scarcity of diagnostic frames may have limited the development of a complete feature representation, leading to variability in diagnostic accuracy.

Overall, while the TOF detection results demonstrate strong performance, the lower accuracy and precision observed in TGA detection highlight potential areas for improvement, such as refining feature selection criteria, enhancing image quality, and optimising plane detection algorithms. Over the next two pages, we show example images from our pipeline evaluation.

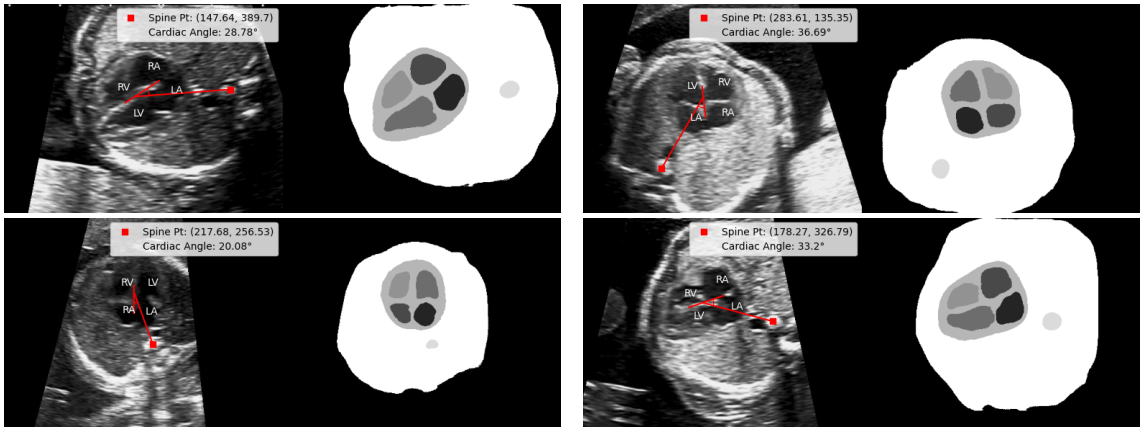


Figure 4.12: Examples of cardiac angles computed for patients with normal fetuses. There is usually a clear delineation of the four chambers with occasional inaccuracies similar to the TOF cases below.

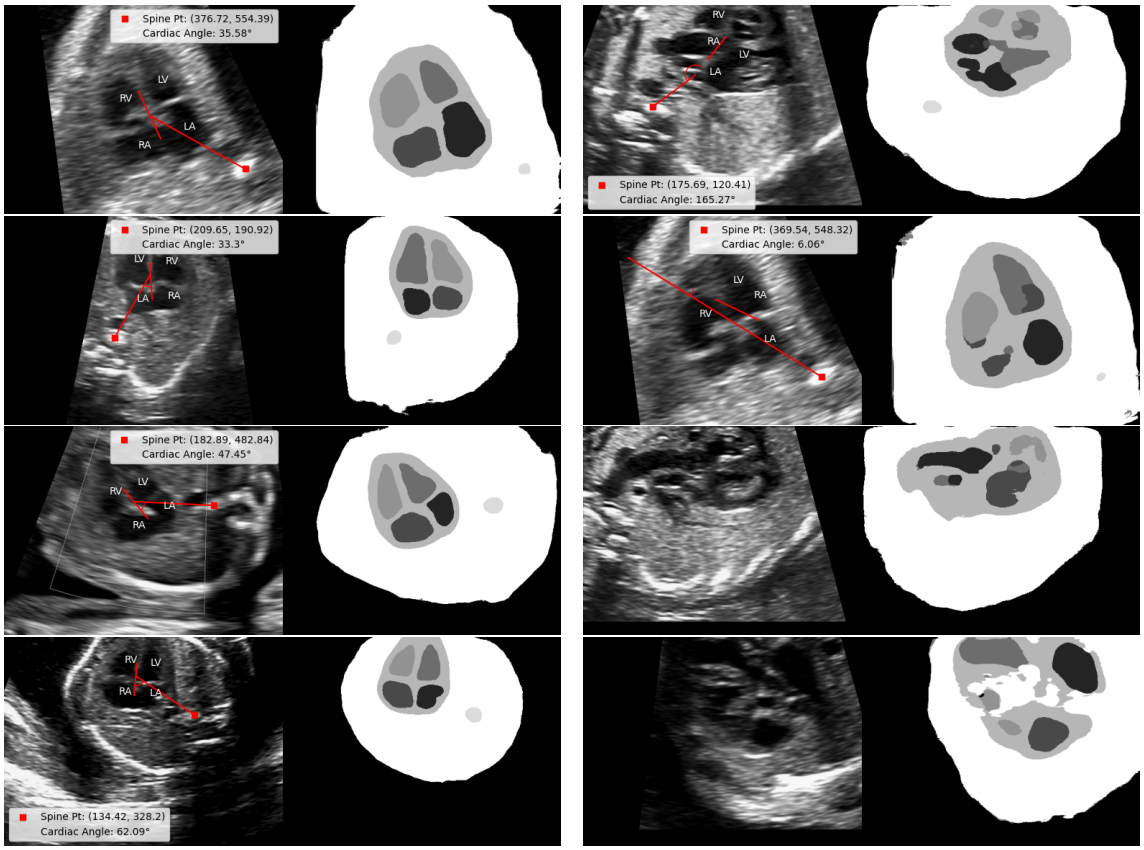


Figure 4.13: Examples of cardiac angles computed by our pipeline for patients with Tetralogy of Fallot. Angles were usually correct (Left) with occasional incorrect or missing angles (Right).

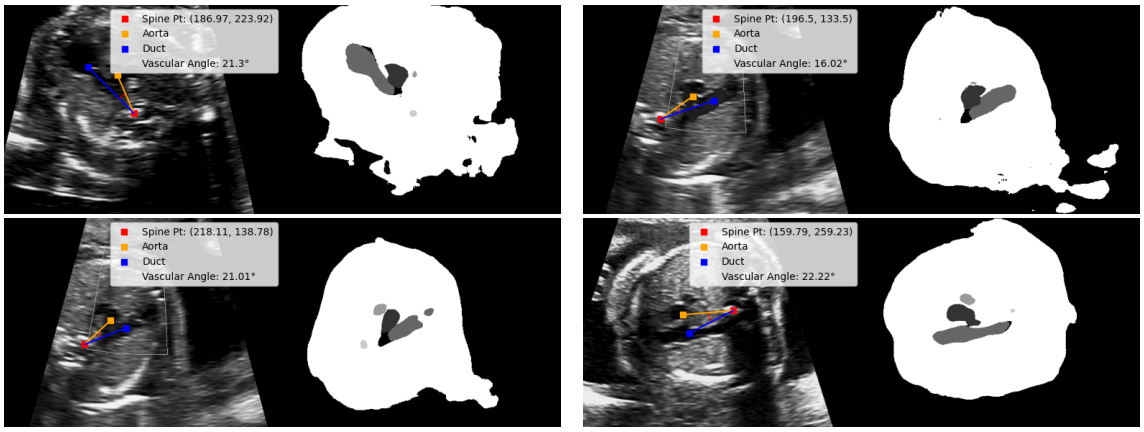


Figure 4.14: Examples of vascular angles computed for patients with normal fetuses. There is a clear delineation of the aorta, duct, and spine in the segmentation masks, similar to the training data, with minimal noise.

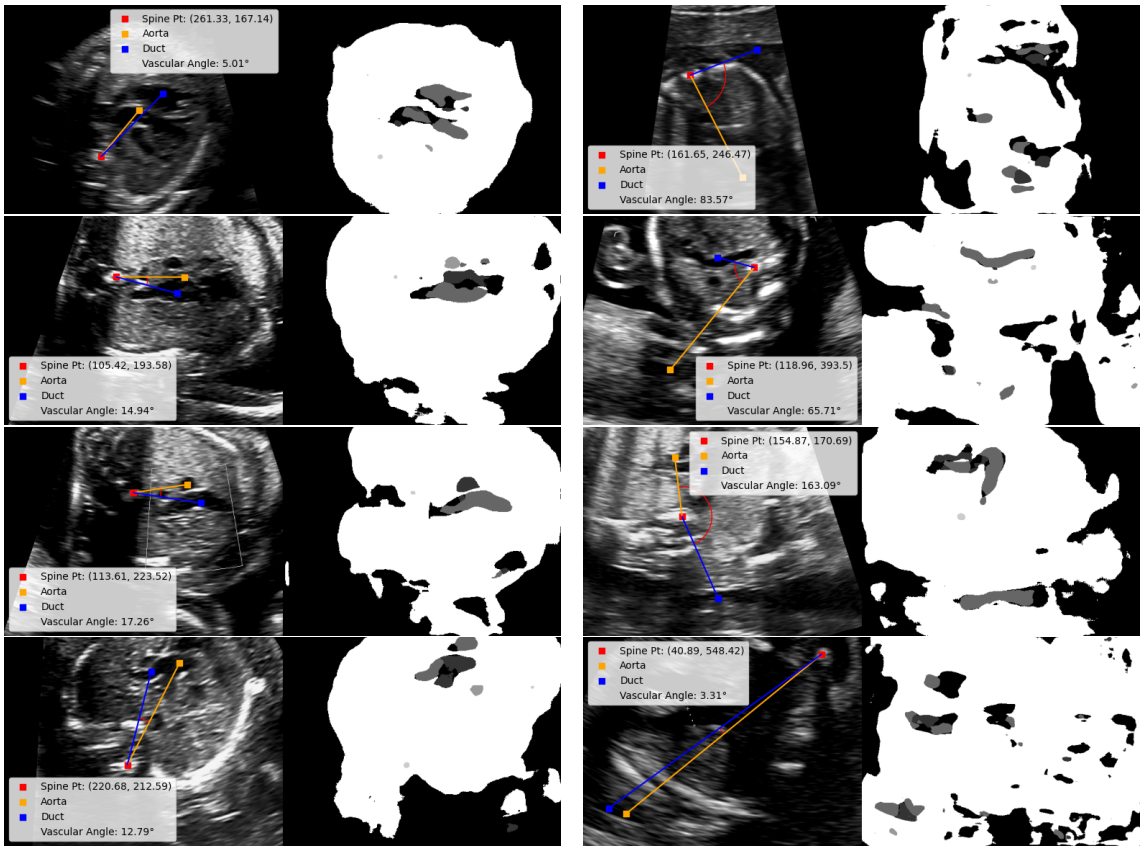


Figure 4.15: Examples of vascular angles computed by our pipeline for patients with Transposition of the Great Arteries. Left: correct angles, Right: incorrect angles. This shows that in some cases, the centroid approach produces decent results despite noisy segmentations.

Chapter 5

Discussion and Conclusion

5.1 Discussion

We have developed and tested an automated deep-learning pipeline for the prenatal detection of congenital heart diseases through detailed experimentation. Our results demonstrate that our pipeline significantly outperforms the current early detection rate for Tetralogy of Fallot (76.6%), achieving an impressive sensitivity of 90.91% and specificity of 88.89%, for subgroups with either a left or right aortic arch. Although the accuracy for Transposition of the Great Arteries was lower at 60% and did not surpass the present rate of diagnosis (84.9%), these findings reveal the clinical potential of deep learning-based automation. They support our hypothesis that robust image analysis methods can streamline the early detection of congenital heart anomalies, leading to improved patient outcomes.

We also hypothesised that converting the spine point into blobs with semantic segmentation could approximate spine point localisation. Our results for TOF classification using our HLHS classifier showed that the cardiac angle, calculated using the predicted spine point, not only classified HLHS with 95.27% accuracy but also identified TOF with a high sensitivity of 90.91% in held-out clinical data. This confirms that our novel approach effectively localises the spine and establishes the automated cardiac axis as a robust multipurpose diagnostic indicator.

We introduced a novel feature, the vascular angle, to examine the spatial alignment of the aorta and duct in 3VT and RVOT scans. This angle, utilised in conjunction with the *DA* ratio, showed promising results for TGA classification. Like the cardiac angle, the vascular angle relied on the spine point, showcasing another application of our proposed spine localisation method. However, we believe that segmentation for 3VT and RVOT planes was sensitive to image quality, and the dataset used for TGA evaluation was limited. This may not have fully reflected the true potential of this angle, and further research will be required to establish its efficacy.

Throughout our research, we made several unique assumptions to enhance the performance of our pipeline by emulating the clinical diagnostic process. For example, similar to how clinicians focus on specific regions of interest when examining ultrasounds, we used a saliency map to crop the images, ensuring that the segmentation model concentrated only on relevant anatomies. This was done to mitigate noise found in real ultrasound data, such as artefacts and varying levels of contrast, which tended to confuse the model in locating the spine. Further, we also restricted the computation of features to only the interquartile range due to observations made from the box plots for classification tasks.

While we are pleased with our achievements within the given timeline, we believe there is scope for further improvement of our study. We could explore methods to assess the quality of anatomical representations in the scans that might enhance the accuracy of feature computations and improve the diagnosis of TGA. We can also try experimenting with different sets of features or employing ensembling techniques to boost the robustness and clinical viability of our pipeline.

5.2 Conclusion

Overall, we have met our objectives by contributing a novel end-to-end pipeline for early diagnosis of fetuses, which has shown notable performance on real clinical data obtained from patients diagnosed with two commonly occurring CHDs, TOF and TGA. We also established a new method of localising the spine in scans, which was used as a landmark to automate the computation of valuable biometric parameters. Additionally, we introduced unique features, based on the clinical definitions of the disease, that have not been previously correlated with these conditions. Furthermore, we recognise opportunities for future research to enhance the accuracy of our pipeline and ultimately lead to more reliable and widespread clinical adoption, improving prenatal care and outcomes for fetuses with congenital heart anomalies.

5.3 Future Work

Automating Pre-processing: While most stages of our pipeline were automated using deep learning techniques to replicate the clinical diagnostic process, real ultrasound data often includes challenges such as split views and Doppler overlays that require manual intervention. Although we initially experimented with basic functions to detect split views, future work could focus on automating this process more robustly.

Identifying Inaccurate Feature Measurements: In our pipeline, we aggregated features by computing their mean and then refined this approach to consider only angles within the interquartile range, which improved our results. However, there remains the challenge of developing mechanisms that can identify and mitigate incorrect measurements across the numerous computed features. We could explore deep learning models for image quality assessment. This will ensure a comprehensive representation of features, unaffected by potential inaccuracies.

Incorporating Unique Identifiable Traits: A robust diagnostic model should effectively accommodate diverse fetal morphologies, which can vary significantly and impact congenital anomaly detection rates. Our datasets were anonymised for ethical reasons, omitting patient-specific details. However, previous studies have demonstrated promising results by correlating features such as maternal health history, gestational age, and demographic factors such as race with congenital anomalies. Integrating these characteristics into our existing solution could potentially enhance its diagnostic accuracy and applicability in clinical settings.

Exploring State-of-the-Art Models and Disease Datasets: Due to the modular structure of our pipeline, further exploration of the most advanced models for each component would have been beneficial, given sufficient time. Additionally, we could have evaluated the pipeline on more nuanced congenital heart diseases with complex features to generate potential valuable insights, but data availability limited our current scope.

Chapter 6

Ethical Issues

The dataset we use in this study, is made up of ultrasound scans of volunteers in their 18-24 week gestation period recorded in a fetal cardiology clinic. Since our research involves processing fetal ultrasound images using deep learning, we acknowledge the ethical concerns related to the use of medical information in a nonclinical setting. We would like to highlight that the data is anonymised providing us with no means of tracing the personal information of the volunteers. Further, we use the same resources with similar supervision as that of the research conducted by Budd et al. [39] which received ethical approval from the NHS R&D and ethical review boards such as the NRES (Reference no. 14/LO/1086) and the LMIAI Centre for Value-Based Healthcare Anonymised Database (Reference no. REC 20/ES/0005). Due to similarities in data, supervision and research objectives, we conclude that the approvals extend to our project. Finally, due to the aforementioned reasons, we are confident that the dataset used and hence our project has no associated ethical considerations or risks.

Bibliography

[1] *Important public health indicators: perinatal and infant mortality.* https://digital.nhs.uk/data-and-information/publications/statistical/ncardrs-congenital-anomaly-statistics-annual-data/ncardrs-congenital-anomaly-statistics-report-2021/perinatal-and-infant-mortality. [Accessed 20th January 2024].

[2] Pan American Health Organization. *Birth defects. The importance of early diagnosis.* https://www.paho.org/en/news/3-3-2023-birth-defects-importance-early-diagnosis. [Accessed 27th December 2023].

[3] Lynn L Simpson. "Screening for congenital heart disease". en. In: *Obstet. Gynecol. Clin. North Am.* 31.1 (Mar. 2004), pp. 51–59.

[4] tiny tickers. *CHD Statistics and Research.* https://www.tinytickers.org/media-centre/chd-statistics-and-research/. [Accessed 25th May 2024].

[5] Mengfang Li et al. "Medical image analysis using deep learning algorithms". In: *Frontiers in Public Health* 11 (2023). ISSN: 2296-2565. DOI: 10.3389/fpubh.2023.1273253. URL: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1273253.

[6] Zhe Guo et al. "Deep Learning-Based Image Segmentation on Multimodal Medical Imaging". In: *IEEE Transactions on Radiation and Plasma Medical Sciences* 3.2 (2019), pp. 162–169. DOI: 10.1109/TRPMS.2018.2890359.

[7] Florina Nela Oşvar et al. "Cardiac axis evaluation as a screening method for detecting cardiac abnormalities in the first trimester of pregnancy". en. In: *Rom. J. Morphol. Embryol.* 61.1 (2020), pp. 137–142.

[8] Youn-Joon Jung, Bo-Ra Lee, and Gwang Jun Kim. "Efficacy of fetal cardiac axis evaluation in the first trimester as a screening tool for congenital heart defect or aneuploidy". en. In: *Obstet. Gynecol. Sci.* 63.3 (May 2020), pp. 278–285.

[9] Florina Nela Oşvar et al. "Cardiac axis evaluation as a screening method for detecting cardiac abnormalities in the first trimester of pregnancy". en. In: *Rom. J. Morphol. Embryol.* 61.1 (2020), pp. 137–142.

[10] C. Athalye et al. "Deep-learning model for prenatal congenital heart disease screening generalizes to community setting and outperforms clinical detection". In: *Ultrasound in Obstetrics & Gynecology* 63.1 (2024), pp. 44–52. DOI: https://doi.org/10.1002/uog.27503. eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.27503. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.27503.

[11] iFind Research Group. https://www.ifindproject.com/about-us/. [Accessed 27th December 2023].

[12] *Medical imaging: what you need to know.* https://www.gov.uk/government/publications/medical-imaging-what-you-need-to-know/medical-imaging-what-you-need-to-know--2. [Accessed 20th January 2024].

[13] Public Health England. *NHS Fetal Anomaly Screening Programme (FASP): programme overview*. https://www.gov.uk/guidance/fetal-anomaly-screening-programme-overview. [Accessed 20th January 2024].

[14] NHS. *Congenital heart disease*. https://www.nhs.uk/conditions/congenital-heart-disease/. [Accessed 27th December 2023].

[15] CINCINNATI CHILDREN'S HOSPITAL MEDICAL CENTER. *Hypoplastic Left Heart Syndrome (HLHS)*. https://www.cincinnatichildrens.org/health/h/hlhs. [Accessed 27th December 2023].

[16] CINCINNATI CHILDREN'S HOSPITAL MEDICAL CENTER. *Transposition of the Great Arteries*. https://www.cincinnatichildrens.org/health/t/transposition. [Accessed 27th December 2023].

[17] Nicholas Aldridge et al. "Detection rates of a national fetal anomaly screening programme: A national cohort study". en. In: *BJOG* 130.1 (Jan. 2023), pp. 51–58.

[18] NICOR. *National Congenital Heart Disease Audit (NCHDA)*. https://www.nicor.org.uk/interactive-reports/national-congenital-heart-disease-audit-nchda. [Accessed 12th June 2023].

[19] British Heart Foundation. *Understanding your child's heart - Tetralogy of Fallot*. https://www.bhf.org.uk/informationsupport/publications/children-and-young-people/understanding-your-childs-heart---tetralogy-of-fallot. [Accessed 25th May 2024].

[20] STANFORD MEDICINE CHILDREN'S HEALTH. *Fetal Circulation*. https://www.stanfordchildrens.org/en/topic/default?id=fetal-circulation-90-P01790. [Accessed 27th December 2023].

[21] Helena Gardiner and Rabih Chaoui. "The fetal three-vessel and tracheal view revisited". In: *Seminars in Fetal and Neonatal Medicine* 18.5 (2013). Perinatal Cardiology, pp. 261–268. ISSN: 1744-165X. DOI: https://doi.org/10.1016/j.siny.2013.01.007. URL: https://www.sciencedirect.com/science/article/pii/S1744165X13000085.

[22] JS Carvalho et al. "ISUOG Practice Guidelines (updated): sonographic screening examinationof the fetal heart". In: *Ultrasound in Obstetrics & Gynecology* 41.3 (2013), pp. 348–359. DOI: https://doi.org/10.1002/uog.12403. eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.12403. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.12403.

[23] Paul D. Russ and Julia A. Drose. *Cardiac Malposition*. https://radiologykey.com/cardiac-malposition/. [Accessed 27th December 2023].

[24] J. S. Carvalho et al. "ISUOG Practice Guidelines (updated): fetal cardiac screening". In: *Ultrasound in Obstetrics & Gynecology* 61.6 (2023), pp. 788–803. DOI: https://doi.org/10.1002/uog.26224. eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.26224. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.26224.

[25] *The Essential Guide to Neural Network Architectures*. https://www.v7labs.com/blog/neural-network-architectures-guide. [Accessed 20th January 2024].

[26] *Convolutional Neural Network Coupled with a Transfer-Learning Approach for Time-Series Flood Predictions - Scientific Figure on ResearchGate*. https://www.researchgate.net/figure/Structures-of-artificial-neural-network-ANN-model-that-show-a-data-flows-in-the-ANN_fig2_338190342. [Accessed 20th January 2024].

[27] Titus J. Brinker et al. "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task". In: *European Journal of Cancer* 113 (2019), pp. 47–54. ISSN: 0959-8049. DOI: https://doi.org/10.1016/j.ejca.2019.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0959804919302217.

[28] Monika Grewal et al. "RADNET: Radiologist Level Accuracy using Deep Learning for HEMORRHAGE detection in CT Scans". In: *CoRR* abs/1710.04934 (2017). arXiv: 1710.04934. URL: http://arxiv.org/abs/1710.04934.

[29] Sai Balaji. "Binary Image classifier CNN using TensorFlow". In: *Techiepedia* (2020).

[30] *Convolutional Neural Networks cheatsheet*. https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks. [Accessed 21st January 2024].

[31] Mahmoud Khodier, Sabah Ahmed, and Mohammed Sayed. "Complex Pattern Jacquard Fabrics Defect Detection Using Convolutional Neural Networks and Multispectral Imaging". In: *IEEE Access* 10 (Jan. 2022), pp. 1–1. DOI: 10.1109/ACCESS.2022.3144843.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[33] Fabian Isensee et al. *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation*. 2018. arXiv: 1809.10486 [cs.CV].

[34] Alexander Weichert et al. "Semi-automatic measurement of fetal cardiac axis in fetuses with congenital heart disease (CHD) with fetal intelligent navigation echocardiography (FINE)". en. In: *J. Clin. Med.* 12.19 (Oct. 2023).

[35] Y. Zhao et al. "Fetal cardiac axis in tetralogy of Fallot: associations with prenatal findings, genetic anomalies and postnatal outcome". In: *Ultrasound in Obstetrics & Gynecology* 50.1 (2017), pp. 58–62. DOI: https://doi.org/10.1002/uog.15998. eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.15998. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.15998.

[36] Aline Wolter et al. "Fetal Cardiac Axis in Fetuses with Conotruncal Anomalies." In: *Ultraschall in der Medizin* 38 (1980), pp. 198–205.

[37] Prabu Pachiyannan et al. "A Cardiac Deep Learning Model (CDLM) to Predict and Identify the Risk Factor of Congenital Heart Disease". In: *Diagnostics* 13 (July 2023). DOI: 10.3390/diagnostics13132195.

[38] Christian F. Baumgartner et al. *SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound*. 2017. arXiv: 1612.05601 [cs.CV].

[39] Samuel Budd et al. *Detecting Hypo-plastic Left Heart Syndrome in Fetal Ultrasound via Disease-specific Atlas Maps*. 2021. arXiv: 2107.02643 [eess.IV].

[40] Lukasz Jakubowski. *In Search of the Heart's Map: Deep Learning for Anatomical Structure Contouring in Fetal Cardiac Analysis*. 2023.

[41] Denny Loevlie. *Logistic Regression with PyTorch*. https://towardsdatascience.com/logistic-regression-with-pytorch-3c8bbea594be. [Accessed 25th May 2024].

[42] opencv. https://opencv.org/. [Accessed 20th January 2024].

[43] T. V. Vigneswaran et al. "Assessment of cardiac angle in fetuses with congenital heart disease at risk of 22q11.2 deletion". In: *Ultrasound in Obstetrics & Gynecology* 46.6 (2015), pp. 695–699. DOI: https://doi.org/10.1002/uog.14832. eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.14832. URL: https://obgyn.onlinelibrary.wiley.com/doi/abs/10.1002/uog.14832.

[44] D Carrasco and L Guedes-Martins. "Cardiac axis in early gestation and congenital heart disease". en. In: *Curr. Cardiol. Rev.* 20.1 (Jan. 2024).

[45] Alisa Arunamata et al. "Abstract 13945: Evaluation of a Deep Neural Network for Detection of D-Transposition of the Great Arteries on Fetal Echocardiograms". In: *Circulation* 148.Suppl_1 (2023), A13945–A13945. DOI: 10.1161/circ.148.suppl_1.13945. eprint: https://www.ahajournals.org/doi/pdf/10.1161/circ.148.suppl_1.13945. URL: https://www.ahajournals.org/doi/abs/10.1161/circ.148.suppl_1.13945.

[46] Rima Arnaout et al. *Deep-learning models improve on community-level diagnosis for common congenital heart disease lesions*. 2018. arXiv: 1809.06993 [cs.CV].

[47] Thomas G Day et al. "Artificial Intelligence to Assist in the Screening Fetal Anomaly Ultrasound Scan (PROMETHEUS): A Randomised Controlled Trial". In: *medRxiv* (2024). DOI: 10.1101/2024.05.23.24307329. eprint: https://www.medrxiv.org/content/early/2024/05/25/2024.05.23.24307329.full.pdf. URL: https://www.medrxiv.org/content/early/2024/05/25/2024.05.23.24307329.

[48] rdroste. *SonoNet$_{PY}$Torch*. https://github.com/rdroste/SonoNet_PyTorch. 2020.

[49] Fabian Isensee. *nnUNet*. https://github.com/MIC-DKFZ/nnUNet/. 2024.

[50] Jia Deng et al. *ImageNet: A large-scale hierarchical image database*. 2009. DOI: 10.1109/CVPR.2009.5206848.

[51] LabelBox. https://labelbox.com/. [Accessed 20th January 2024].

[52] Docker Inc. *Docker overview*. https://docs.docker.com/guides/docker-overview/. [Accessed 16th June 2023].

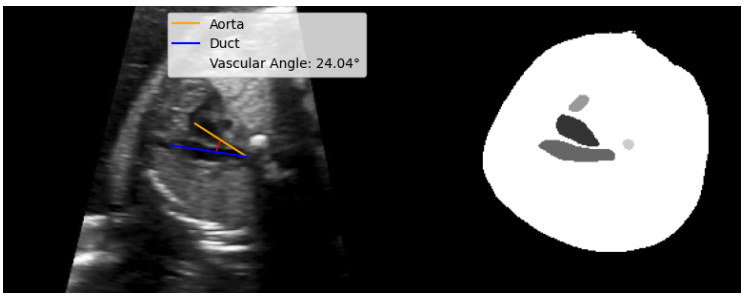
[53] Fraiya. *Fraiya Ultrasound*. https://fraiya.com/. [Accessed 16th June 2023].

[54] S F Wong et al. "Pulmonary artery/aorta ratio in simple screening for fetal outflow tract abnormalities during the second trimester". en. In: *Ultrasound Obstet. Gynecol.* 30.3 (Sept. 2007), pp. 275–280.

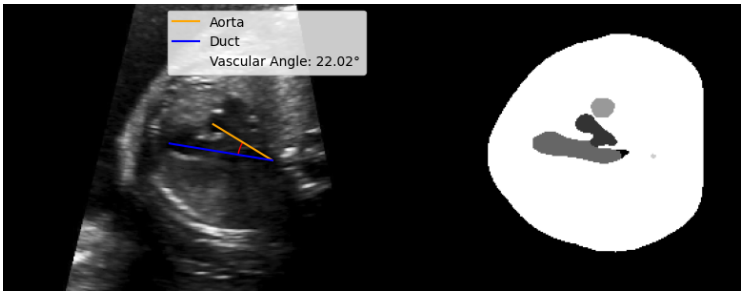
Appendix A

Example Images Illustrating Model Performance

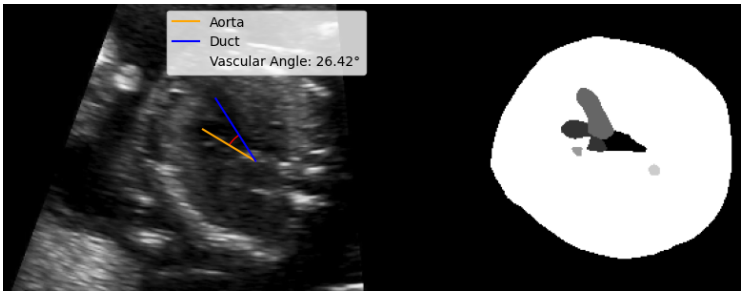
A.1 Vascular Angle for Different Image Qualities



(a) Good



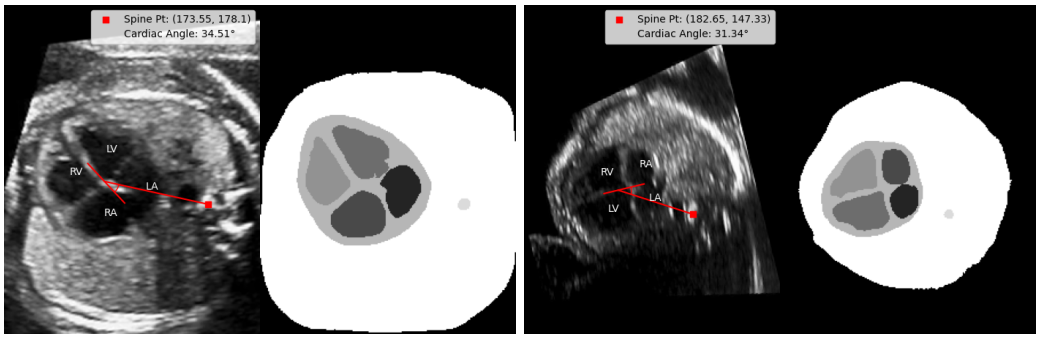
(b) Medium



(c) Bad

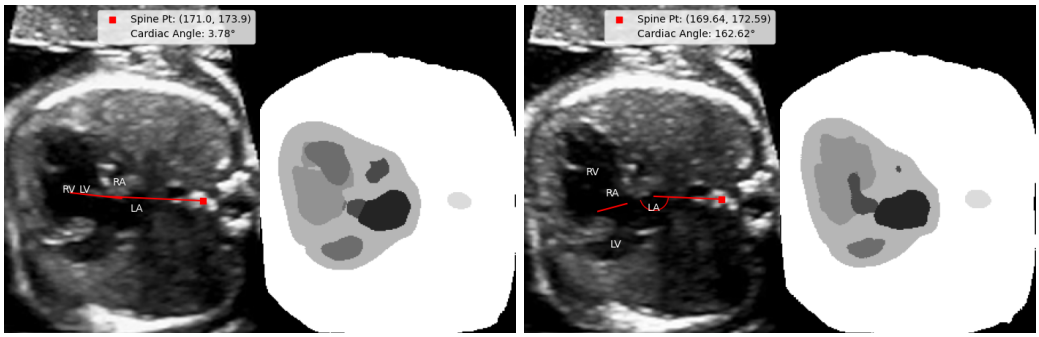
Figure A.1: Figure shows examples of randomly selected good, medium and bad quality images obtained from patient FN008. Vascular angles computed in the three images are similar regardless of the ultrasound scan quality. This consistency supports our decision to include medium and lower-quality images in the dataset of normal cases.

A.2 Cardiac Angles for Normal and HLHS cases



(a) (b)

Figure A.2: Examples of segmentations for NORM cases with clear delineation of anatomical landmarks.



(a) (b)

Figure A.3: Examples of segmentations for HLHS cases showing irregular boundaries and incomplete delineation of cardiac anatomy. These cardiac angles are observed as outliers in the box plot.

A.3 Vascular Angles using Spine Centroid

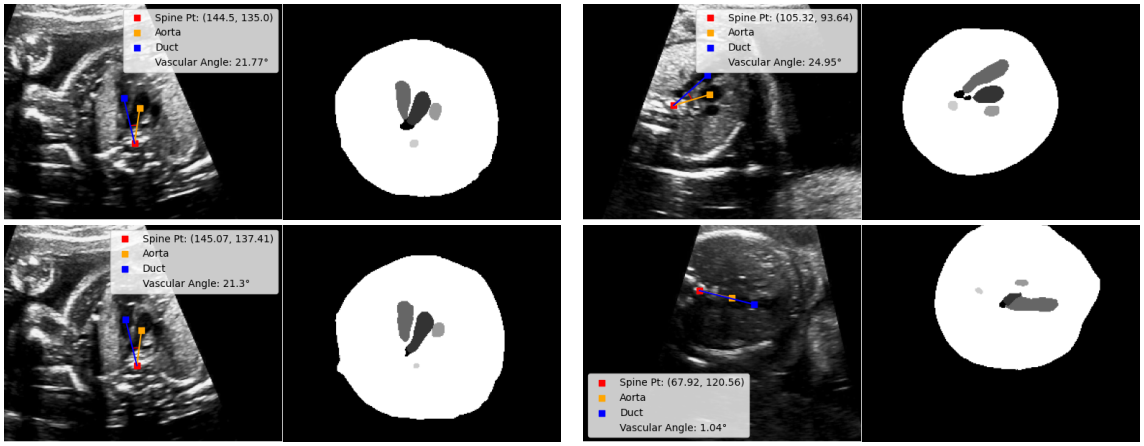


Figure A.4: Examples of 3VT images for which vascular angles was computed as intended using the centroid approach.

Appendix B

Plane Detection Results on 4CH View for HLHS & NORM Patients

B.1 SonoNet

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (141/162)	43 4CH 99 Abdominal 12 Background 1 Kidneys 7 LVOT	40.32°	True
4CH-Medium (79/129)	68 4CH 20 Abdominal 6 Background 35 LVOT	36.57°	False
4CH-Bad (37/105)	22 4CH 14 Abdominal 53 Background 9 LVOT 1 Lips 5 RVOT 1 Spine (sag.)	41.23°	True

Table B.1: Results for Patient HLH007

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (1/5)	4 Abdominal 1 RVOT	N/A	N/A
4CH-Medium (3/12)	4 Background 8 RVOT	N/A	N/A
4CH-Bad (0/1)	1 LVOT	N/A	N/A

Table B.3: Results for Patient HLH029

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (0/4)	3 Abdominal 1 Lips	N/A	N/A
4CH-Medium (1/78)	9 Abdominal 69 Background	N/A	N/A
4CH-Bad (21/828)	85 Abdominal 715 Background 1 Brain (Cb.) 7 Femur 13 Lips 5 Profile 2 Spine (sag.)	N/A	N/A

Table B.2: Results for Patient HLH016

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Medium (0/14)	4 4CH 5 Background 3 LVOT 2 RVOT	N/A	N/A
4CH-Bad (0/7)	7 Background	N/A	N/A

Table B.4: Results for Patient HLH033

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (254/469)	1 4CH 386 Abdominal 78 Background 4 Lips	N/A	N/A
4CH-Medium (385/1114)	452 Abdominal 593 Background 69 Lips	N/A	N/A
4CH-Bad (302/1457)	286 Abdominal 984 Background 11 Brain (Cb.) 1 Femur 4 LVOT 171 Lips	N/A	N/A

Table B.5: Results for Patient HLH037

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (2/2)	2 LVOT	N/A	N/A
4CH-Medium (7/7)	2 3VV 4 4CH 1 RVOT	47.46°	True
4CH-Bad (2/2)	2 4CH	44.38°	True

Table B.6: Results for Patient HLH039

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Medium (4/15)	2 4CH 1 Abdominal 12 Background	45.27°	True
4CH-Bad (4/110)	4 Abdominal 105 Background 1 RVOT	N/A	N/A

Table B.7: Results for Patient HLH046

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (0/5)	3 Abdominal 2 Background	N/A	N/A
4CH-Medium (0/16)	16 Abdominal	N/A	N/A
4CH-Bad (5/106)	10 Abdominal 93 Background 3 Brain (Tv.)	N/A	N/A

Table B.8: Results for Patient HLH060

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (2/435)	151 4CH 55 Abdominal 217 Background 2 Brain (Tv.) 1 LVOT 9 Lips	N/A	N/A
4CH-Medium (3/204)	115 4CH 4 Abdominal 33 Background 2 Femur 1 LVOT 49 Lips	33.13°	False
4CH-Bad (78/502)	162 4CH 91 Abdominal 239 Background 3 Femur 7 Lips	46.7°	True

Table B.9: Results for Patient FN006

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (289/944)	1 3VV 483 4CH 202 Abdominal	37.24°	False
4CH-Medium (152/359)	22 4CH 35 Abdominal 133 Background 2 Femur 167 Lips	38.52°	True
4CH-Bad (150/397)	90 4CH 70 Abdominal 137 Background 2 Brain (Tv.) 5 LVOT 92 Lips 1 Profile	41.92°	True

Table B.10: Results for Patient FN008

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (23/86)	19 Abdominal 65 Background 2 Profile	N/A	N/A
4CH-Medium (0/29)	6 Abdominal 23 Background	N/A	N/A
4CH-Bad (20/39)	13 Abdominal 26 Background	N/A	N/A

Table B.11: Results for Patient NORM007

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (39/170)	6 4CH 102 Abdominal 7 Background 55 Lips	38.41°	True
4CH-Medium (29/224)	72 4CH 58 Abdominal 1 Background 93 Lips	40.47°	True
4CH-Bad (12/31)	14 Abdominal 1 Background 16 Lips	N/A	N/A

Table B.12: Results for Patient NORM014

B.2 SonoNext

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (141/162)	75 4CH 56 Background 2 LVOT 29 SPINE-SAGITTAL	51.46°	True
4CH-Medium (79/129)	2 3VT 1 4CH 103 Background 13 LVOT 8 RVOT 2 SPINE-SAGITTAL	7.51°	False
4CH-Bad (37/105)	1 4CH 83 Background 15 LVOT 3 RVOT 3 Spine (sag.)	37.84°	False

Table B.13: Results for Patient HLH007

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (0/4)	2 Background 2 LVOT	N/A	N/A
4CH-Medium (1/78)	77 Background 1 LVOT	N/A	N/A
4CH-Bad (21/828)	1 3VT 41 4CH 3 Abdominal 713 Background 2 Lips 6 LVOT 1 RVOT 61 Spine (sag.)	N/A	N/A

Table B.14: Results for Patient HLH016

Plane-Quality	Planes Detected	Avg. Cardiac Angle	HLHS Status
4CH-Good (1/5)	2 4CH 3 Background	21.04°	False
4CH-Medium (3/12)	4 4CH 4 Background 4 RVOT	36.08°	False
4CH-Bad (0/1)	1 4CH	N/A	N/A

Table B.15: Results for Patient HLH029