Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

Supervised Fetal Cardiac Anomaly Detection: Handling Label Noise and Temporal Dynamics with Vision Foundation Models

Author:
Abhivir Singh

Supervisor:
Professor Bernhard Kainz

Second Marker:
Professor Yingzhen Li

January 22, 2026

Contents

Chapter 1

Introduction

Congenital heart disease (CHD) is the leading cause of infant mortality, affecting roughly 8 to 10 in 1000 live births worldwide, with early prenatal detection critical for improving outcomes through timely intervention and specialised care planning [1]. The Fetal Anomaly Screening Programme (FASP) in the UK requires routine ultrasound screening at 18-21 weeks of gestation, where sonographers examine the fetal heart to identify structural abnormalities [2].

Studies report that detection rates typically range from 40-80% in the UK [3] and are as low as 28% in some places [4]. Variability in detection rates arises primarily because sonography is inherently challenging: fetal cardiac structures are small, dynamic, and often obscured by maternal tissue, requiring highly skilled sonographers with extensive training. Unlike CT or MRI scans, where image acquisition is standardised, ultrasound requires the operator to simultaneously acquire the image, optimise the physics (gain, depth, focus), and interpret the anatomy in real-time. In addition, ultrasound screening is not a static scan, but a video sequence. For many conditions, an accurate diagnosis requires analysing temporal changes, and in some cases, the key sign of pathology may appear only fleetingly across a few frames. Cognitive factors such as fatigue and inattentional blindness play a significant role. Furthermore, the rarity of anomalies contributes to the "prevalence effect", where low target prevalence leads to higher miss-rates [5].

This variability, combined with the increasing demand for screening services and the limited availability of expert fetal cardiologists, creates a pressing need for automated decision support systems that can assist sonographers in identifying potential abnormalities.

Current approaches to automated fetal cardiac analysis face technical limitations. Traditional deep learning methods typically require training end-to-end models from scratch on domain-specific medical imaging datasets, which demands substantial computational resources and large volumes of labelled data—both scarce in medical imaging contexts [6]. Moreover, fetal ultrasound presents unique challenges: videos contain multiple frames per subject, labels are provided at the subject level rather than the frame level, and the dataset exhibits severe class imbalance with healthy cases vastly

outnumbering disease cases. Existing domain-specific models, while effective, lack the generalisability and transfer learning capabilities that have revolutionised natural image analysis. The domain gap between natural images (on which most foundation models are trained) and medical ultrasound images further complicates direct application of state-of-the-art vision models.

Recent advances in self-supervised learning (SSL) have produced foundation models such as DINOv2, which generate high-quality visual features from large-scale natural image datasets without human supervision [7]. These models demonstrate remarkable transfer learning capabilities, achieving competitive performance across diverse computer vision tasks when used as frozen feature extractors, without requiring task-specific fine-tuning. In medical imaging, preliminary evidence suggests that foundation model embeddings can effectively capture semantic information even when applied to domains distinct from their training distribution [8]. This leads to the motivation for exploring whether generalist foundation model embeddings can support supervised classification tasks in fetal ultrasound imaging, potentially offering a more efficient and scalable alternative to training domain-specific models from scratch.

Chapter 2

Background

2.1 Congenital Heart Disease

2.1.1 Prenatal Screening & Standard Views

The NHS FASP defines a standardised mid-trimester cardiac protocol comprising four key anatomical views [9]:

Four-Chamber View (4CV / 4CH): It's 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. In normal fetuses, atria and ventricles are equal in size, with an intact septum and a normal atrioventricular (AV) valve offset. [10]

Three Vessels and Trachea View (3VT): This view shows the spatial arrangement of major vessels: Pulmonary Trunk (anterior/left), Aorta (central) and Superior Vena Cava (posterior/right). In a normal scan, the three vessels are arranged in a straight line in decreasing order of their diameters from left to right. This view is used to assess vessel alignment, arch crossover, and great vessel relationships. It detects anomalies such as coarctation, right aortic arch, or transposition.

Left Ventricular Outflow Tract (LVOT) View : The LVOT is a muscular channel in the left ventricle that transports blood outward and into the aorta. This view is acquired by angling the ultrasound transducer anteriorly towards the fetal right shoulder from the 4CH plane. This view confirms the continuity between the left ventricle and the ascending aorta, with the aortic valve opening freely and the aorta arising from the left ventricle.

Right Ventricular Outflow Tract (RVOT) View: The RVOT is the pathway through which blood flows out of the right ventricle and into the pulmonary trunk. This view is used to assess the relationship between the aorta (posterior) and the pulmonary trunk (anterior and to the left).

In Figure 2.1 below, all four views are shown for reference along with a data quality score. The score is based on landmark clarity, artifact absence, and adequate visu-

alisation (with a score of 1 indicating poor quality and 5 indicating high quality). [11].

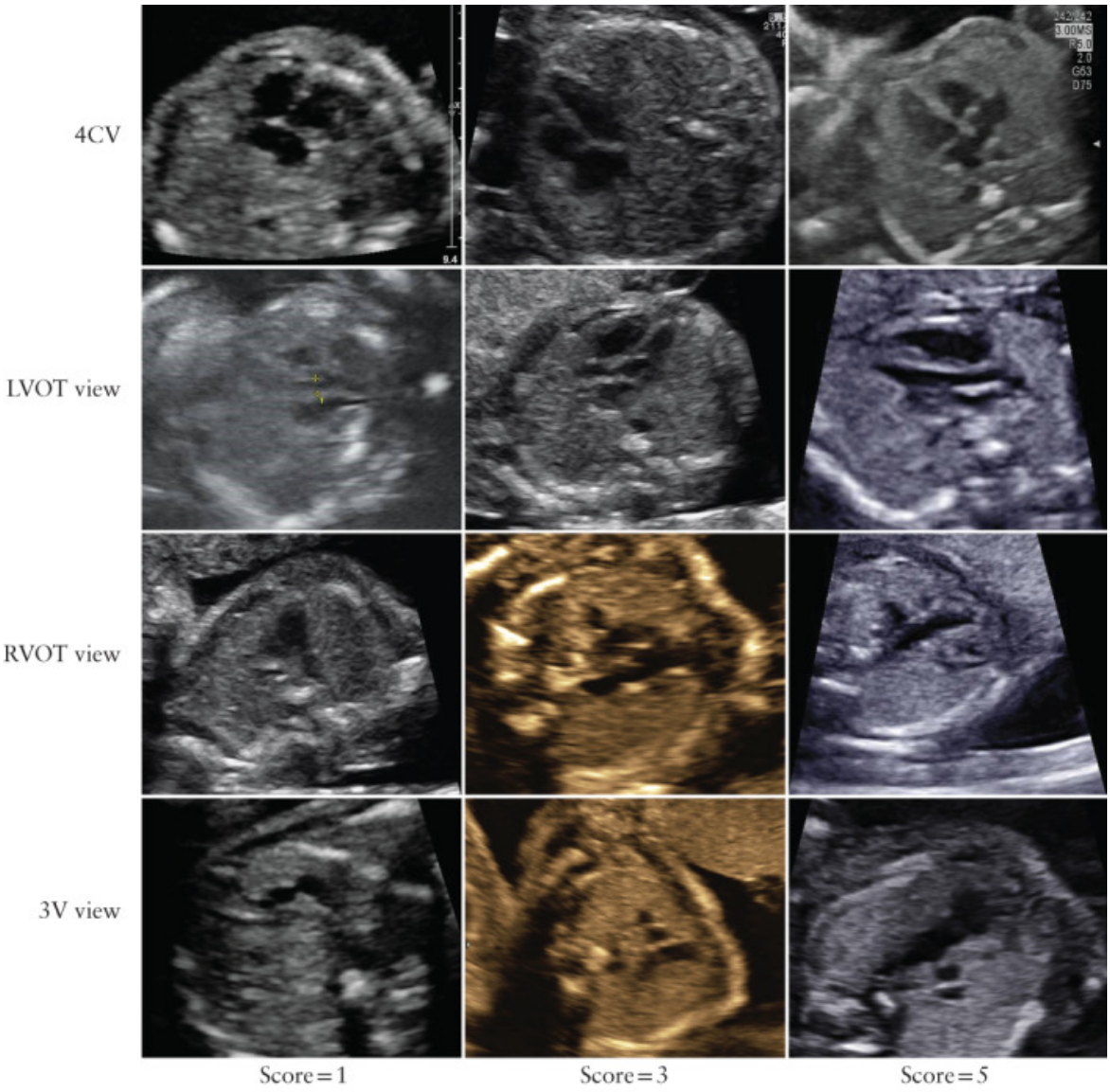

Figure 2.1: Examples of ultrasound images of fetal heart in four-chamber (4CV), left ventricular outflow tract (LVOT), right ventricular outflow tract (RVOT) and three-vessel (3V) views, that obtained quality score of 1, 3 or 5, in cases with severe congenital heart disease at birth.

Summarised in Table 2.1 below.

2.1.2 Technical Challenges in Computational Fetal Ultrasound

Fetal ultrasound video represents one of the most challenging data modalities from a computer vision perspective:

- **Speckle Noise**: Images are formed by the constructive and destructive inter-

Table 2.1: Standard fetal echocardiography views and clinical assessment.

VIEW	ANATOMICAL ASSESSMENT	DETECTABLE CONDITIONS
Four-Chamber View (4CV)	Ventricular symmetry, AV valve offset, crux of heart	Hypoplastic Left Heart Syndrome (HLHS), Atrioventricular septal defect (AVSD), Single Ventricle
Left Ventricular Outflow Tract (LVOT)	Septal-aortic continuity	Transposition of the Great Arteries (TGA), VSD
Right Ventricular Outflow Tract (RVOT)	Pulmonary artery crossing aorta	Pulmonary Stenosis, TGA
Three-Vessel and Trachea View (3VT)	Vessel alignment, aortic arch	Coarctation, Right Aortic Arch

ference patterns produced by echoes scattering back from tissues, resulting in granular texture that confuse edge detection algorithms and standard convolutional neural networks (CNNs).

- **Acoustic Shadowing**: High-density structures like the fetal spine or ribs reflect the majority of sound waves, creating signal dropout behind them. In cardiac screening, a shadow cast by a rib can obscure the interventricular septum, mimicking a ventricular septal defect (false positive) or hiding a real defect (false negative).
- **View Dependency**: Unlike 3D volumetric modalities, 2D ultrasound provides planar cross-sections. The appearance of the heart is highly dependent on the angle of sonar sensor (a slight tilt of the probe can make a normal heart appear abnormal or vice versa).

These challenges motivate the use of robust feature representations that can generalise across the noise and variability inherent in ultrasound imaging.

2.2 SSL and Vision Foundation Models

2.2.1 SSL as Feature Learning for Supervised Tasks

The field of computer vision has undergone a transformation analogous to the "BERT moment" in natural language processing, driven by foundation models. Foundation models are large-scale models pretrained on massive datasets using self-supervised objectives that learn generalisable feature representations without explicit human annotation [12].

The key insight is that self-supervised learning can learn rich visual features from unlabelled data, which can then be used for supervised learning tasks with limited labelled examples. This hybrid approach is particularly valuable in medical imaging, where expert annotation is expensive and time-consuming, but large volumes of unlabelled imaging data is used to train general purpose large scale vision models. Rather than replacing supervised learning, self-supervised pretraining provides a powerful feature extraction stage that enables effective supervised classification with much less labelled examples than would be required to train from scratch.

Self-supervised learning creates pretext tasks that generate supervision signals from the data itself. The two main paradigms are:

- **Contrastive Learning** (e.g., SimCLR, MoCo): Learns representations by pulling together different augmented views of the same image while pushing apart views from different images. While effective, these methods require large batch sizes and careful negative sampling [13].
- **Self-Distillation** (e.g., DINO, iBOT): Uses a student-teacher framework where both networks share the same architecture but different parameters. The student learns to match the teacher's outputs across different augmented views, with the teacher updated as an exponential moving average of the student. [14]

2.2.2 DINO: Self-Distillation with No Labels

DINO (self-*DI*stillation with *NO* labels) introduced a seminal approach to self-supervised learning for Vision Transformers [15]. The key innovation lies in its student-teacher framework applied to image patches.

Given an input image, multiple augmented views are generated: global crops (large field of view) and local crops (small field of view). The student network processes all views, while the teacher sees only global views. The training objective minimises the cross-entropy between teacher and student output distributions:

$$L_{DINO} = -\sum_{x \in \{global\}} \sum_{x' \in V, x' \neq x} P_t(x) \log P_s(x') \tag{2.1}$$

The teacher parameters are not updated via gradient descent but as an exponential moving average of the student:

$$\theta_t \leftarrow \lambda \theta_t + (1 - \lambda)\theta_s \tag{2.2}$$

Emergent Properties: A remarkable finding of DINO is that the self-attention maps of the trained model correspond to semantic regions of the image, effectively learning object segmentation without any segmentation supervision. The features cluster by semantic category in embedding space, suggesting that self-supervised objectives can learn semantically meaningful representations.

2.2.3 DINOv2: Scaling Self-Supervised Learning

DINOv2 represents a significant advancement over its predecessor, trained on a curated dataset of 142 million images (LVD-142M) and producing features that work as they are (frozen) across different tasks and image distributions without task-specific fine-tuning [7].

Key Technical Contributions:

- **Data Curation Pipeline**: Automatic curation from 1.2 billion web images using self-supervised retrieval to find images similar to high-quality seed sets, followed by deduplication and diversity balancing.
- **Combined Training Objectives**:
  1. *Image-level (DINO)*: Cross-entropy loss between student and teacher on global views
  2. *Patch-level (iBOT)*: Masked patch prediction where student predicts teacher's features for masked patches
  3. *KoLeo Regulariser*: Encourages uniform distribution of features within a batch, improving retrieval performance
- **Efficiency Optimisations**: FlashAttention, sequence packing, improved stochastic depth, and knowledge distillation from larger to smaller models.

Performance Characteristics:

- Achieves 86.5% top-1 accuracy on ImageNet-1k with a linear probe (+4.2% over previous SSL methods)
- Drastically improved robustness on out-of-distribution benchmarks (+29.6% on ImageNet-A over iBOT)
- State-of-the-art frozen features for dense tasks (depth estimation, semantic segmentation)

Table 2.2: ViT model variants with parameter counts and embedding dimensions.

MODEL	PARAMETERS	EMBEDDING DIMENSION
ViT-S/14	˜22M	384
ViT-B/14	˜86M	768
ViT-L/14	˜300M	1024
ViT-g/14	˜1.1B	1536

The frozen feature paradigm is particularly attractive for medical imaging applications where labelled data is scarce but high-quality feature representations are critical.

2.2.4 Recent Developments: DINOv3

DINOv3, released in 2025, extends the framework with improved video understanding capabilities [16]. Of particular relevance to this project is the approach to video classification: rather than extracting a single per-frame embedding, DINOv3 extracts patch-level representations from each frame, augments them with spatial and temporal positional embeddings, and feeds the resulting spatiotemporal sequence through a lightweight transformer for classification.

This approach preserves fine-grained spatial information across time enabling the model to identify motion patterns and temporal relationships that frame-level averaging would obscure, which is a crucial capability for cardiac ultrasound where the beating heart's motion patterns carry diagnostic information.

2.3 Video Classification Methods

2.3.1 The Evolution of Video Understanding

Video classification has evolved through several architectural paradigms, representing different trade-offs between computational cost and temporal modelling capability:

Frame-Level Classification with Aggregation: The simplest approach applies an image classifier to each frame independently and aggregates predictions (e.g., mean, max, voting). This strategy is computationally efficient and allows direct application of pretrained image models, but it fundamentally ignores temporal relationships: the output is invariant to frame order [17].

CNN-RNN Hybrids: Combines convolutional neural networks (for spatial feature extraction) with recurrent neural networks such as LSTMs or GRUs (for temporal modelling). The CNN processes each frame to produce a feature vector and the sequence of features is fed to the RNN. While theoretically good, these models suffer from vanishing gradients in long sequences and high computational costs [18].

3D Convolutions: Architectures such as C3D [19] and I3D [20] extend 2D convolutions to the temporal dimension, processing spatiotemporal volumes directly. These models can capture local motion patterns (e.g., the motion between adjacent frames) and have achieved strong results on action recognition benchmarks. However, they require substantially more parameters and computation than 2D alternatives and are limited in their ability to capture long-range temporal dependencies.

Video Transformers: The self-attention mechanism naturally generalises to sequences of arbitrary length. Video transformers treat the input as a sequence of patch tokens drawn from multiple frames. Positional encodings capture both spatial (within-frame) and temporal (across-frame) position. DINOv3 [16] is state-of-the-art (SOTA) in spatial feature extraction from videos and InternVideo2 is SOTA on video benchmarks.

Applies self-attention across spatiotemporal patches, treating video as a sequence of patch tokens with positional encodings for both space and time. Captures long-range

dependencies but has quadratic complexity in sequence length.

2.3.2 Aggregation Methods for Frame Embeddings

When using pretrained image encoders for video classification (as this project proposes with DINOv2), the choice of aggregation method significantly impacts both performance and the types of patterns the model can recognise:

Mean Pooling: Element wise averaging of frame embeddings. This approach is simple, parameter-free, and stable, but it loses information about variation across frames. Consider the beating heart: embeddings of systolic frames may differ systematically from those of diastolic frames, but mean pooling collapses this temporal variation into a single "average" representation. Diagnostic information encoded in the dynamics of the cardiac cycle would be lost.

Max Pooling: Takes element-wise maximum across frames. In principle, this captures the most "salient" features from each embedding dimension. However, the semantics of maximum in a learned latent space are unclear. Unlike class logits, where maximum has a natural interpretation, the maximum of arbitrary embedding dimensions may not correspond to anything meaningful. Max pooling is also highly sensitive to outliers: a single noisy frame can dominate the aggregated representation.

Attention Pooling: A learned weighted combination of frame embeddings, where the weights are computed by a small network (typically a linear layer followed by softmax). This allows the model to learn which frames are most informative, down-weighting uninformative or noisy frames. Attention pooling has shown strong results in multiple-instance learning problems, which share structural similarity with video classification: both involve aggregating variable-length sets of instances (frames or patches) to a single prediction.

Temporal Transformers: Treats frame embeddings as tokens in a sequence and applies self-attention to learn complex temporal relationships. A temporal transformer can attend to distant frames, model the dependencies between cardiac phases, and potentially learn to recognise motion patterns. This is the most expressive option but computationally expensive and needs more training data to learn the additional parameters.

2.3.3 Temporal Considerations for Cardiac Imaging

Fetal cardiac ultrasound presents specific temporal requirements. With fetal heart rates of 120-160 BPM, a complete cardiac cycle spans approximately 375-500 milliseconds. At 30 frames per second, this corresponds to 11-15 frames.

Implications:

- Sampling fewer frames risks missing important phases of the cardiac cycle
- Some conditions may only be visible during specific phases (e.g., valve motion abnormalities)

- Motion patterns themselves carry diagnostic information

The STUD framework [21] addressed this by using 64-frame clips (approximately 2 seconds at standard rates), ensuring capture of multiple complete cardiac cycles. For efficiency, they sampled every 3rd frame, trading temporal resolution for computational tractability.

An aggregation method like mean pooling over a cardiac cycle may converge to a central average representation, losing the motion component that distinguishes healthy from pathological hearts. This is the reason for using attention-based or transformer-based aggregation that can learn to focus on diagnostically relevant temporal patterns.

2.4 Handling Data Challenges

2.4.1 Learning from Noisy Labels

A distinctive challenge in this project is that labels are available only at the subject level, not at the frame or video level. When these labels are propagated to individual frames for training, significant label noise is introduced where many frames in an "unhealthy" subject may show perfectly normal anatomy, while the pathology may only be visible in specific frames or views.

The Noisy vs Hard Sample Problem: A key challenge is distinguishing between:

- *Noisy samples*: Incorrectly labelled (should be down-weighted or excluded). If given high weight, it will teach the model incorrect associations and degrade performance. Example: a frame that happens not to show the heart.
- *Hard samples*: Difficult but correctly labelled (may benefit from additional attention). If down-weighted, may cause the model to fail on the very cases it most needs to learn. Example: a frame showing a subtle defect.

In both cases, the model struggles to learn, but optimal handling differs.

Techniques for Noisy Label Learning:

1. **Sample Reweighting**: Assign lower weights to samples likely to be mislabelled, often identified through high loss values or inconsistent predictions across training epochs.
2. **Curriculum Learning**: Train on "clean" (high-confidence) samples first, gradually introducing noisier samples as the model develops robust feature representations.
3. **Co-Teaching**: Train two networks simultaneously, each selecting low-loss samples for the other to learn from, exploiting the observation that different networks memorise different noisy samples.
4. **Neighbourhood Consistency**: Use embedding similarity to identify likely mislabelled samples, where if a sample's nearest neighbours have inconsistent labels, the sample may be noisy.

Project Context: With subject-level labels, we expect substantial label noise at the frame level. The hierarchical evaluation strategy (frame → clip → subject) naturally addresses this: while frame-level performance may be limited by noise, aggregation to clip and subject levels should improve results as noise averages out.

2.4.2 Class Imbalance Techniques

The datasets for CHD generally exhibit severe class imbalance, with unhealthy subjects and classes being very low in number compared to healthy ones. Standard training approaches on imbalanced data tend to be biased toward majority classes, as correctly classifying the abundant healthy cases contributes more to the overall loss than correctly classifying the rare disease cases.

Loss Function Modifications:

1. **Weighted Cross-Entropy**: Assigns higher loss weight to minority classes, typically proportional to inverse class frequency:

$$\mathcal{L}_{\text{WCE}} = -\sum_{c} w_c \cdot y_c \log(\hat{y}_c) \tag{2.3}$$

   where w_c is inversely proportional to class frequency.

2. **Focal Loss** [22]: Down-weights the contribution of easy examples to focus training on hard cases:

$$\mathcal{L}_{\text{FL}}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \tag{2.4}$$

   where p_t is the model's estimated probability for the correct class, $\alpha_t \in [0, 1]$ is a balancing factor (often used to give more weight to the positive class), and $\gamma \geq 0$ is the focusing parameter.

   When a sample is correctly classified ($p_t \approx 1$), the modulating factor $(1 - p_t)^{\gamma}$ approaches zero. So the easy examples are down-weighted and training gets focused on hard samples that were misclassified.

3. **Class-Balanced Loss**: Weights samples by effective number of samples per class, accounting for diminishing marginal benefit of additional samples.

Sampling Strategies:

1. **Oversampling**: Sample minority classes more frequently during training, either through simple duplication or synthetic generation (SMOTE).

2. **Undersampling**: Reduce majority class samples to balance the dataset, trading off data utilisation for balance.

3. **Hard Example Mining**: Prioritise samples that the model currently classifies incorrectly or with low confidence.

Evaluation Considerations: Standard accuracy metrics are misleading under class imbalance (as 80% accuracy could be achieved by predicting "healthy for all samples). Appropriate metrics include:

1. **Macro-averaged AUROC**: Averages per-class AUROC, treating all classes equally
2. **Precision and Recall**: Particularly important for minority (disease) classes
3. **Matthews Correlation Coefficient (MCC)**: Balanced measure that accounts for all four confusion matrix quadrants

2.4.3 Multi-Label Classification

The problem is fundamentally multi-label rather than multi-class: subjects may present with multiple concurrent conditions. This affects both architecture and loss function choices:

Architecture: Use sigmoid activations rather than softmax, producing independent probabilities for each class.

Loss Function: Binary cross-entropy applied independently to each class. The binary cross-entropy loss is defined as:

$$\mathcal{L}_{\text{BCE}} = -\sum_{c} \left[y_c \log(\hat{y}_c) + (1 - y_c) \log(1 - \hat{y}_c) \right], \tag{2.5}$$

where $y_c \in \{0, 1\}$ is the ground-truth label for class c, and $\hat{y}_c \in [0, 1]$ is the predicted probability for class c.

Threshold Selection: Each class may require a different decision threshold, optimised based on the desired precision-recall trade-off for that condition.

2.5 Foundation Models in Medical Imaging

2.5.1 The Domain Gap Challenge

A fundamental question for applying vision foundation models to medical imaging is whether representations learned from natural images transfer effectively to the medical domain. Natural images and medical images differ substantially in their visual characteristics, semantic content, and relevant features for downstream tasks. The visual characteristics of natural photographs and medical images differ substantially: medical images often have lower contrast, different texture statistics, and semantically relevant features at scales or locations that differ from those in photographs. The objects of interest in medical imaging (tumours, anatomical structures, subtle abnormalities) have little visual resemblance to the objects that dominate natural image datasets (faces, cars, animals).

Despite these concerns, accumulating evidence suggests that foundation model features provide surprisingly strong performance across medical imaging modalities. Studies have demonstrated that DINOv2 features achieve competitive results on classification tasks across MRI, CT, X-ray, and dermoscopic images when used with simple linear

classifiers [8]. The robustness of these features to the change in distribution [23] is particularly valuable given the well documented variability in medical imaging equipment, protocols, and patient populations across institutions.

One explanation for this transferability is that foundation models learn representations capturing fundamental visual primitives (edges, textures, shapes, spatial relationships) that are relevant across domains. The object-centric nature of DINOv2 representations, where features naturally separate foreground structures from background, aligns well with medical imaging tasks that require isolating anatomical regions from surrounding tissue. Whether this transfers specifically to ultrasound, with its characteristic speckle texture and view-dependent appearance for CHD, is an empirical question that this project aims to address.

2.5.2 Unsupervised Anomaly Detection with Foundation Model Embeddings

While this project focuses on supervised classification, it is useful to examine how foundation model embeddings have been applied to unsupervised tasks in medical imaging, as this demonstrates the general utility of these features. Schulthess and Konukoglu [24] explored whether DINOv2 embeddings could support unsupervised anomaly detection in medical imaging, proposing a method that combines frozen DINOv2 features with a Dirichlet Process Mixture Model (DPMM). The approach:

1. Extracts DINOv2 embeddings from normative (healthy) images
2. Fits a DPMM to model the distribution of normal embeddings
3. At inference, computes similarity between test embeddings and DPMM cluster centres
4. Lower similarity indicates higher anomaly score

Key Finding: Normalised DINOv2 embeddings align with anatomical structures even in the presence of anomalies, suggesting that the features capture semantically meaningful information despite the domain gap from natural images.

The method achieved competitive AUROC while halving inference time compared to memory-bank approaches like PatchCore (39ms vs 422ms per image). However, reviewers noted a significant gap in AUPR performance, suggesting that while the approach identifies many anomalies, it may struggle with precise localisation.

2.5.3 L-FUSION: Uncertainty-Aware Segmentation

L-FUSION (Laplacian Fetal US Segmentation with Integrated FoundatiON models) represents the state-of-the-art in fetal ultrasound analysis [25]. The framework addresses the need for uncertainty quantification in clinical AI systems:

Architecture: Uses a frozen Universal Ultrasound Foundation Model (USFM) as encoder with two parallel decoder heads:

- Laplacian Head: Models segmentation logits as Gaussian distribution, predicting mean and covariance
- Dropout Head: Uses Monte Carlo Dropout to sample multiple plausible segmentation hypotheses

Uncertainty Types:

- Aleatoric Uncertainty: Inherent data noise (e.g., acoustic shadows)
- Epistemic Uncertainty: Model knowledge gaps, critical for out-of-distribution detection

Performance: Achieved 95.5% Dice on fetal head segmentation (HC18 dataset) while running at 40-45 FPS, and boosted out-of-distribution detection by up to 20% through uncertainty calibration using inter-path variance (comparing disagreement between the two heads).

Relevance: The "frozen encoder + trainable heads" paradigm aligns directly with the approach proposed in this project. The authors mathematically prove (via Sufficieny Theorem 2.1) [25] that if a foundation model (like USFM) preserves all segmentation-relevant information in its embeddings, those embeddings are a sufficient statistic for both the segmentation mask and its uncertainty. This supports the use of frozen features for downstream classification.

2.5.4 Zero-Shot CHD Detection

The STUD (Sparse Tube Ultrasound Distillation) framework [21] represents a recent breakthrough in Congenital Heart Disease detection.

Approach: Formulates CHD detection as normality modelling, where training is only done on healthy fetal cardiac ultrasound videos and detecting anomalies as deviations from the learned normal distribution.

Technical Innovation: Uses sparse 3D space-time tubes for efficient video tokenisation, reducing tokens to 9.39% of dense sampling while preserving spatiotemporal detail. Training follows a DINO-style self-distillation objective adapted for video.

Privacy-Preserving Model Merging: Proposes DiVMerge (Divergence Vector-guided Model Merging) to combine models trained at different hospital sites without data sharing:

1. Compute geometric median of local models as robust centre
2. Weight each model's contribution based on divergence from median
3. Retain site-specific parameters only if they confidently deviate

Results: On external test sites with unseen anomalies (RAA, LSVC, VSD, CM), the merged model achieved 81.0% accuracy and 78.1% F1-score, outperforming centralised training by 20% accuracy and 54.6% F1-score.

Critical Analysis: While zero-shot detection elegantly sidesteps the need for labelled disease data, it has limitations:

- Requires large volumes of healthy data for training the normality model
- May flag normal anatomical variants as abnormal
- The "any clip abnormal" video-level aggregation rule may not be optimal
- Performance varies substantially across sites (89.7% vs 72.2% accuracy)

A supervised approach trained on labelled disease examples, as proposed in this project, may achieve higher precision by learning disease-specific features rather than relying solely on deviation from normality.

2.6 Summary

2.6.1 State of the Art

The current state of the art in fetal cardiac ultrasound analysis is characterised by several parallel research directions:

- **Zero-Shot Anomaly Detection**: The STUD framework [21] achieves 78.1% F1-score on CHD detection using only healthy training data, demonstrating that normality modelling can detect pathology without labelled disease examples.
- **Uncertainty-Aware Analysis**: L-FUSION [25] provides state-of-the-art segmentation with dual uncertainty estimation, enabling out-of-distribution detection and clinical feedback.
- **Foundation Model with Domain Adaptation**: MM-DINOv2 [26] demonstrates effective domain adaptation of vision foundation models for medical imaging, with semi-supervised extensions that leverage unlabelled data.
- **Efficient Video Processing**: Sparse tube sampling [21] and efficient architectures [27] enable processing of ultrasound video at clinically acceptable frame rates.

2.6.2 Identified Gaps

1. **Supervised Classification with Foundation Models**: While zero-shot anomaly detection avoids the need for labelled disease data [21], it may achieve lower precision than supervised approaches trained on disease examples. No work has systematically evaluated supervised classification using foundation model embeddings on fetal cardiac ultrasound.
2. **Optimal Feature Aggregation**: The choice between frame-wise processing with post-hoc aggregation versus proper temporal modelling of video clips remains unexplored for this domain. How much does incorporating temporal information improve over simple mean pooling?

3. **Handling Subject-Level Labels**: The challenge of training with noisy labels propagated from subject-level annotations has not been specifically addressed in the fetal ultrasound literature.
4. **Practical Deployment Considerations**: Most work focuses on achieving high accuracy metrics; fewer studies address the precision-recall trade-offs relevant for clinical screening workflows.

2.6.3 Project Aim

This project addresses these gaps through a step by step investigation of supervised classification using DINOv2 embeddings for fetal cardiac ultrasound analysis. The key contributions will include:

1. **Comprehensive Baseline Evaluation:** Establishing the performance of frozen DINOv2 embeddings with various classifiers (logistic regression, MLP) and aggregation methods (mean, max, attention) as a foundation for further improvements.
2. **Temporal Modelling Investigation**: Comparing frame-wise classification with post-hoc aggregation against proper video classification approaches that incorporate temporal information through attention mechanisms or transformers.
3. **Noisy Label Handling**: Evaluating techniques for learning with subject-level labels that introduce frame-level noise, including sample reweighting and curriculum learning.
4. **Class Imbalance Strategies**: Systematically evaluating loss weighting, focal loss, and other techniques for handling the severe class imbalance in disease detection.
5. **Multi-Level Evaluation**: Reporting results at frame, clip, and subject levels to provide a complete picture of model performance and the benefits of aggregation.

The project bridges the gap between zero-shot anomaly detection (which requires no disease labels but may lack precision) and end-to-end supervised training (which requires extensive labelled data and computational resources). By leveraging foundation model embeddings with custom classifiers, it aims to achieve high classification performance while remaining computationally efficient and robust to the data challenges inherent in clinical ultrasound screening.

Chapter 3

Project Plan

3.0.1 Completed Work

Data Exploration & Baseline Results: Conducted an exploratory data analysis. Used Pandas library to analyse CSVs and custom scripts to analyse dataset videos' metadata and content.

Found and reported a) the number of subjects (healthy, unhealthy & of specific disease type), b) statistics on individual disease types, c) video dataset characteristics (average video length, videos per subject, frames per second, resolution, Doppler overlay percentage).

Established baseline results using Frozen DINOv2 and Logistic Regression classifier. Configuration used for baseline: 10 frames per video, maximum 20 videos per subject, mean pooling aggregation across frames and videos, balanced class weighting.

3.0.2 Future work (in order of implementation)

Classifier Architecture Experiments: Replace logistic regression with a multi-layer perceptron to learn non-linear decision boundaries. Experiment with fine-tuning the last layers of DINOv2 to check whether domain adaptation bridges the gap between natural images and ultrasound.

Held-Out Disease Evaluation: Train the model with one disease class entirely held out, then evaluate whether the model can detect this unseen anomaly as "unhealthy" at test time. This tests generalisation to novel conditions

Aggregation Method Comparison: Mean pooling provides a simple baseline but may lose important information. Alternative aggregation strategies of Attention pooling and Transformer aggregation will be evaluated.

Temporal Sampling Experiments: The current approach samples only 10 frames (0.33 seconds at 30 FPS), which may not capture a complete cardiac cycle. Given fetal heart rates of 120-160 BPM, a full cycle requires approximately 0.4-0.5 seconds.

Experiments with Increased frame count, Temporal stride, and Full video sampling will be evaluated.

Patch-Level Temporal Modelling: Rather than extracting a single embedding per frame, this approach preserves spatial structure: 1) Extract patch-level representations from DINOv2 (not the final pooled embedding), 2) Add learnable positional embeddings encoding both spatial (x, y) and temporal (frame number) positions. 3) Feed the sequence of spatiotemporal patches through a lightweight transformer and 4) The transformer learns to aggregate information across both space and time. This allows the model to a) Track how specific image regions change over time, b) Identify motion patterns that may indicate cardiac abnormalities, c) Learn which patches are informative for classification.

Attention-Based Interpretability: The transformer's attention weights provide a natural mechanism for model interpretability: 1) Analyse which frame patches receive high attention for different predictions, 2) Visualise attention maps overlaid on original video frames, 3) Compare attention patterns between correctly and incorrectly classified samples

Dealing with Data Imbalance and Noisy Labels: 1) Evaluate and build pipeline for multi label classification and experiment with the different techniques used to handle class imbalance, mentioned in Section 2.4.2. 2) Conduct View-Specific Analysis as different ultrasound views are optimised for detecting different conditions. 3) Implement techniques to handle noisy labels and to make the distinction between hard and noisy labels.

Test-Time Augmentation: Apply sliding window sampling across videos at test time and ensemble predictions. This may improve robustness without requiring model changes.

Conduct Interpretability Analysis: Detailed analysis of attention patterns and false positive/negative cases: a) Extract attention maps for misclassified samples. b) Analyse whether attention focuses on clinically relevant structures. c) Investigate specific failure modes (e.g., shadow artifacts causing false positives for certain conditions)

Uncertainty Estimation: (potential extension): Incorporate uncertainty quantification to identify cases where the model is uncertain: epistemic uncertainty and aleatoric uncertainty. This could flag cases for human review, and be important for clinical deployment.

Chapter 4

Evaluation Plan

4.1 Key Questions

The evaluation is designed to answer the following research questions:

1. **Do frozen DINOv2 embeddings provide discriminative features for fetal cardiac anomaly detection?**
   - *Measured by:* AUROC significantly above random baseline.
2. **Does aggregating predictions improve performance?**
   - *Measured by:* Subject-level AUROC > Clip-level > Frame-level.
3. **Does incorporating temporal information help?**
   - *Measured by:* Video-based methods (e.g., Transformer aggregation) outperform frame-wise methods with simple aggregation.
4. **Can the model handle the severe class imbalance?**
   - *Measured by:* Acceptable per-class AUROC and recall for minority diseases.
5. **How does the approach compare to existing methods?**
   - *Measured by:* Comparison with zero-shot methods such as STUD [21] and other prior literature.
6. **Is the model learning clinically meaningful features?**
   - *Measured by:* Qualitative attention analysis and view-specific performance patterns.

4.2 Success Criteria

The project success is defined by the following primary and secondary metrics:

Criterion	Metric / Target	Rationale
Discriminative Power	Macro-AUROC $\geq$ 0.85	Validates frozen foundation model efficacy.
Minority Detection	Recall (Unhealthy) $\geq$ 0.80	Minimizes clinically critical false negatives.
Temporal Value	$> 5\%$ AUROC improvement	Validates video vs. frame-wise modeling.
Multi-label Extension	9 disease classes	Assesses specificity beyond binary detection.
Interpretability	Qualitative attention	Ensures focus on clinical cardiac structures.

4.3 Evaluation Metrics

Given the severe class imbalance (84.5% healthy), standard accuracy is discarded in favor of:

- **AUROC (Macro-averaged):** Primary metric for discrimination across thresholds, treating minority disease classes equally.
- **AUPR:** Highly informative for imbalanced data; the random baseline is equal to class prevalence (≈ 0.155).
- **Recall (Sensitivity):** prioritized to minimize missed diagnoses of ductal-dependent lesions.
- **Multi-label Metrics:** Per-class AUROC, Hamming Loss, and Exact Match Ratio for the 9-disease classification task.

Detailed analysis will include confusion matrices and violin plots of prediction scores.

4.4 Hierarchical Evaluation Structure

Evaluation is conducted at three levels to address the granularity mismatch between labels (subject-level) and predictions (frame/clip-level):

1. **Frame-Level:** Establishes a baseline for comparison with prior literature. Labels are propagated from the subject level, introducing inherent noise.
2. **Clip-Level:** Assesses methodological contributions in temporal modeling. We compare mean, max, attention, and Transformer-based pooling.

3. **Subject-Level (Clinical Target):** Final diagnostic output. Clip-level predictions are aggregated to provide a per-patient diagnosis, averaging out uninformative video segments.

Chapter 5

Declarations

5.1 Ethical Considerations

This project uses anonymised fetal cardiac ultrasound video data. All data used in this project are existing datasets with research ethics committee approvals in place, including approvals from the London - West London & GTAC Research Ethics Committee (12/LO/1247, 14/LO/1805, 14/LO/1806, 19/LO/1957, 22/LO/0163) and other relevant committees (IRAS 164026, 257568, 292223; REC 20/ES/0005). The development of the proposed methods has no ethical implications. The use of anonymised patient image data and public data has been classified as ethically uncritical by the NHS National Institute for Health Research (NIHR) and will follow ICO standards. All validation experiments are non-invasive and safe, involving retrospective analysis of anonymised images and quantitative evaluation of numerical data.

Bibliography

[1] Hoffman JI, Kaplan S. The incidence of congenital heart disease. Journal of the American College of Cardiology. 2002;39(12):1890-900. pages 1

[2] Public Health England. Fetal Anomaly Screening Programme: Handbook; 2021. Accessed: 2024-01-01. https://www.gov.uk/government/publications/fetal-anomaly-screening-programme-handbook. pages 1

[3] Tegnander E, Eik-Nes S. Prenatal detection of heart defects in a non-selected population of 30 149 fetuses—detection rates and outcome. Ultrasound in Obstetrics & Gynecology. 2006;27(3):252-65. pages 1

[4] Ciulpan A, Lacatușu A, Pop LL, Paul C, Lungeanu D, Iacob D, et al. Incidence and Antenatal Detection of Congenital Heart Malformations—Data from a Tertiary Obstetric Romanian Center. Diagnostics. 2024;14(15):1659. pages 1

[5] Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. Journal of experimental psychology: General. 2007;136(4):623. pages 1

[6] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical image analysis. 2017;42:60-88. pages 1

[7] Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:230407193. 2023. pages 2, 7

[8] Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, et al. Big self-supervised models are strong semi-supervised learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 14878-88. pages 2, 13

[9] Public Health England. Screening for fetal anomalies: programme overview. London, UK; 2021. Accessed: 2024-05-20. Available from: https://www.gov.uk/guidance/fetal-anomaly-screening-programme-overview. pages 3

[10] Mahatme I. Deep Learning Pipeline for Automated Diagnosis of Fetal Congenital Cardiac Anomalies [Distinguished MEng Individual Project].

Imperial College London; 2024. Department of Computing. Available from: https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/distinguished-projects/2324-ug-projects/Mahatme,-Ishita-(im620).pdf. pages 3

[11] Carvalho JS, Axt-Fliedner R, Chaoui R, Copel JA, Cuneo BF, Goff D, et al. ISUOG Practice Guidelines (updated): fetal cardiac screening. Ultrasound in Obstetrics & Gynecology. 2023;61(6):788-803. First published: 02 June 2023. Open Access. Available from: https://doi.org/10.1002/uog.26224. pages 4

[12] Bommasani R. On the opportunities and risks of foundation models. arXiv preprint arXiv:210807258. 2021. pages 5

[13] Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. In: International Conference on Machine Learning. PMLR; 2020. p. 1597-607. pages 6

[14] Zhou J, Wei C, Wang H, Shen W, Xie C, Yuille A, et al. iBOT: Image BERT Pre-Training with Online Tokenizer. In: International Conference on Learning Representations; 2022. . pages 6

[15] Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 9650-60. pages 6

[16] Meta AI. DINOv3: Advancing self-supervised learning for vision. arXiv preprint. 2025;arXiv:2508.10104. Available from: https://arxiv.org/abs/2508.10104. pages 8

[17] Ng JYH, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G. Beyond Short Snippets: Deep Networks for Video Classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. . pages 8

[18] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, et al. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 2625-34. pages 8

[19] Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning Spatiotemporal Features with 3D Convolutional Networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015. . pages 8

[20] Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. . pages 8

[21] Saha P, Mishra D, Hernandez-Cruz N, Patey O, Papageorghiou AT, Asano YM, et al. Self-supervised normality learning and divergence vector-guided model merging for zero-shot congenital heart disease detection in fetal ultrasound videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2025. p. 560-71. pages 10, 14, 15, 19

[22] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2980-8. pages 11

[23] Lu MY, Chen B, Williamson DF, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. Nature medicine. 2024;30(3):863-74. pages 13

[24] Schulthess N, Konukoglu E. Anomaly Detection by Clustering DINO Embeddings Using a Dirichlet Process Mixture. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2025. p. 46-56. pages 13

[25] Müller JP, Wright R, Day TG, Venturini L, Budd SF, Reynaud H, et al. L-FUSION: Laplacian fetal ultrasound segmentation and uncertainty estimation. In: International Workshop on Advances in Simplifying Medical Ultrasound. Springer; 2025. p. 164-73. pages 13, 14, 15

[26] Scholz D, Erdur AC, Ehm V, Meyer-Baese A, Peeken JC, Rueckert D, et al. MM-DINOv2: Adapting Foundation Models for Multi-modal Medical Image Analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2025. p. 320-30. pages 15

[27] Pu B, Li K, Cheng N, Li S, Huang N, Liao G, et al. Semantic segmentation of the four-chamber view of fetal echocardiography based on MobileUNet-FPN. IEEE Journal of Biomedical and Health Informatics. 2022;26(9):4437-47. pages 15