International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Gynosight AI: A Trustworthy Explainable Ultrasound Assistant for Obstetrics and Gynaecology

Aditya Prakash, Anand S, Kiran U, Muhammed Swalih N, Asst. Professor Mrs. L. Nithya, Professor Dr. S. Pathur Nisha, Professor Dr. A. B. Arockia Christopher
Computer Science and Engineering
Nehru Institute of Technology, Coimbatore

Abstract- Ultrasound is the main imaging test in obstetrics and gynaecology, used to monitor fetal growth and to examine the ovaries and uterus. Existing artificial intelligence tools each handle only one condition, give an answer even when the image is blurred or unfamiliar, and show heatmaps that are judged only by eye. This project proposes GynoSight AI, a trustworthy explainable ultrasound assistant with four modules: fetal growth (head circumference measurement, with an amniotic fluid check from the amniotic fluid index entered by the doctor), ovarian mass classification, polycystic ovary syndrome (PCOS) detection and uterine fibroid detection. The modules use DenseNet121, EfficientNet-B0 and U-Net with transfer learning, and Grad-CAM++ highlights the region behind each finding. A trust gate combines image quality, Monte Carlo Dropout uncertainty, an energy-based familiarity score and a measurable heatmap plausibility score. Conformal risk control sets the acceptance threshold so that the error rate among accepted cases stays below a chosen limit, and every rejected case is referred to a gynaecologist with a plain-language reason. Phase–I covers the literature survey, gap analysis, requirements, system design and evaluation plan; implementation and testing on public datasets are planned for Phase–II.

Keywords- Obstetric Ultrasound, Explainable AI, Grad-CAM++, Uncertainty Estimation, Conformal Risk Control.

I. INTRODUCTION

1. Background and Context

Obstetrics and gynaecology (OB/GYN) is one of the busiest departments in any hospital. It cares for women throughout pregnancy and childbirth (obstetrics) and treats disorders of the female reproductive organs such as the uterus and ovaries (gynaecology). Ultrasound is its main imaging test because it is safe, inexpensive, portable and does not use ionizing radiation. This section explains the role of ultrasound in the department, the growth of artificial intelligence (AI) in medical imaging, and why trust is the central issue for AI support in this field.

Ultrasound in Obstetrics and Gynaecology

Ultrasound imaging works by sending high-frequency sound waves into the body through a probe (transducer) and recording the echoes that return from tissue boundaries. A computer converts the time and strength of the echoes into a grey-scale image in real time. Fluid appears black because it reflects little sound, soft tissue appears in shades of grey, and bone and calcification appear bright white.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Because no radiation is involved, ultrasound can be repeated as often as needed during pregnancy, which makes it the standard imaging test for the fetus.

In obstetrics, ultrasound is used to confirm and date a pregnancy, to check the number of fetuses, to screen for structural abnormalities, to check the position of the placenta and to monitor fetal growth and the amount of amniotic fluid. Growth is assessed by recognizing standard views (planes) of the fetus and measuring structures such as the head circumference, the abdominal circumference and the femur length. These measurements are compared with reference charts for the gestational age.

In gynaecology, ultrasound is performed through the abdomen (trans-abdominal) or through the vagina (trans-vaginal) to examine the uterus and ovaries. It is the first test for pelvic pain, abnormal bleeding, infertility and suspected masses. Common findings include ovarian cysts and tumours, the polycystic ovarian pattern seen in polycystic ovary syndrome (PCOS), and uterine fibroids. Because the same machine and largely the same staff handle both obstetric and gynaecological scans, a single assistant that supports both types of scan fits the way the department actually works.

Ultrasound also has well-known weaknesses. The image depends on how the probe is held, the machine settings (gain, depth and focus), the patient's body habitus and the skill of the operator. Images often contain speckle noise, acoustic shadows behind bone or gas, and on-screen text and measurement markers. Two operators can obtain quite different images of the same patient. In many Indian hospitals a small number of trained sonologists handle a large number of scans, and less experienced staff often perform the first examination. These factors make ultrasound one of the hardest imaging types for both humans and computers to interpret consistently.

Artificial Intelligence in Medical Imaging

Deep learning, and in particular the convolutional neural network (CNN), has transformed automatic image analysis over the last decade. A CNN learns to recognize patterns directly from labelled images instead of relying on hand-crafted rules. Surveys of deep learning in medical image analysis show applications in almost every imaging type and organ, and landmark studies have shown expert-level performance, for example in classifying skin cancer from photographs. Reviews describe this as the convergence of human and artificial intelligence in medicine, while also warning about bias, poor generalization and lack of transparency.

In OB/GYN ultrasound, CNNs have been used to detect standard fetal planes in real time [2], to recognize common maternal–fetal planes with accuracy close to experts [4], to measure fetal head circumference [22], to detect fibroids [23] and PCOS [7], and to separate benign from malignant ovarian tumours at a level comparable to expert examiners in a large multicentre study [6]. Most of these models are trained by transfer learning: a network pre-trained on millions of natural photographs is fine-tuned on a smaller set of medical images.

The Need for Trustworthy AI

High accuracy on a test set is not the same as safety in the clinic. A blurred, shadowed or wrongly angled ultrasound image can lead a model to a confident but wrong answer, and a model trained at one hospital may behave differently on another hospital's machines. Explainable AI methods such as Grad-CAM and Grad-CAM++ [5] can show which region influenced a prediction, but a benchmark on chest X-rays found that such heatmaps often disagree with the regions drawn by experts [19]. Modern networks are also known to be over-confident, so their probability outputs cannot be taken at face value.

For these reasons, an AI assistant for the OB/GYN department must do more than predict. It must check that the image is usable, recognize when an image is unlike its training data, verify that its explanation

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

points to the right structure, and know when to stop and ask a gynaecologist. These four abilities together are what this report calls trustworthiness, and they form the core of the proposed system.

Clinical Conditions Addressed

GynoSight AI focuses on the common conditions for which public ultrasound data exists. Figure 1.1 shows how they are grouped under obstetrics and gynaecology.

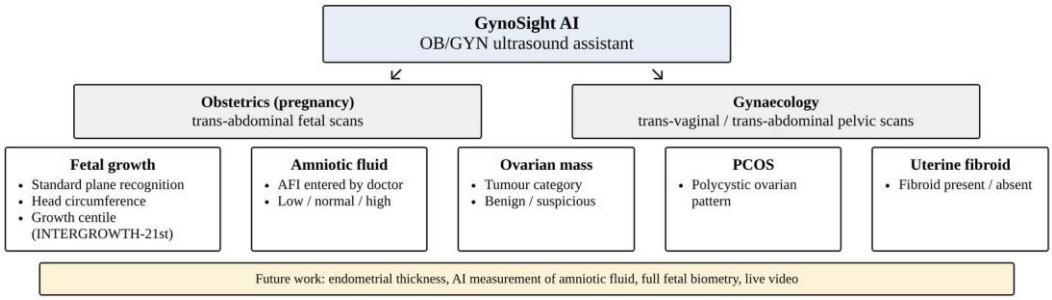


Figure 1 Clinical conditions addressed by GynoSight AI

Fetal Growth

Poor fetal growth (fetal growth restriction) and excessive growth are both linked to complications at birth. Growth is judged by measuring the fetus and comparing the value with international standards such as INTERGROWTH-21st, which convert a measurement into a centile for the gestational age [15]. Values below the 3rd or above the 97th centile are commonly flagged for closer review. The head circumference is one of the core measurements and is measured on the trans-thalamic plane of the fetal head.

Amniotic Fluid

Amniotic fluid surrounds and protects the fetus. Too little fluid (oligohydramnios) or too much fluid (polyhydramnios) requires closer monitoring. The amniotic fluid index (AFI) is measured by dividing the uterus into four quadrants and adding the deepest vertical pocket of fluid in each [16]. In this project the AFI is measured by the sonographer and entered by the user, and the system classifies the level as low, normal or high.

Ovarian Masses

Ovarian cysts and tumours range from harmless functional cysts to cancer. Expert examiners use structured rules, such as the International Ovarian Tumor Analysis (IOTA) simple rules, which look at features like solid parts, papillary structures, irregular walls, shadowing and blood flow [21]. Correctly separating benign from suspicious masses decides whether a woman needs specialist surgery.

Polycystic ovary syndrome (PCOS). PCOS is a common hormonal disorder in women of reproductive age. Under the Rotterdam criteria, it is diagnosed when two of three features are present: irregular or absent ovulation, signs of excess androgen, and polycystic ovarian morphology on ultrasound [18]. The ultrasound part, many small follicles or an enlarged ovarian volume, is what the PCOS module aims to recognize.

Uterine Fibroids

Fibroids (leiomyomas) are non-cancerous growths of the uterine muscle and are very common. They can cause heavy bleeding, pain and fertility problems. The FIGO classification describes fibroids by their position, from submucosal (inside the cavity) to subserosal (on the outer surface) [14]. Ultrasound is the first test used to detect them.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

2. Problem Statement

Practical Problems with Existing AI Tools

Existing AI tools for OB/GYN ultrasound have the following practical problems:

- Each tool handles only one task, such as fetal planes, ovarian tumours or fibroids, so a department would need many separate tools with different interfaces.
- A prediction is always returned, even for blurred, shadowed or wrongly taken images, although ultrasound quality depends strongly on the operator.
- Heatmaps are missing or judged only by eye, so there is no check that the model is looking at the fetal head, the tumour or the uterus.
- Referral to an expert is at best simulated or suggested, with no guaranteed error limit and no reason given for each referral.
- Public datasets are small, often from one hospital, and in some cases full of duplicate images, which inflates the reported accuracy.

Clinical Consequences

These problems matter in practice. A junior sonographer may not realize that an image is of poor quality and may receive a confident but wrong result. A heatmap that highlights the bladder instead of an ovarian mass could mislead a doctor who trusts it. An assistant that shows every answer with the same confidence gives the gynaecologist no way of knowing which cases need careful review, so it saves little time. In the worst case, over-trust in an unreliable tool could delay the diagnosis of a growth problem or a suspicious mass.

Problem Definition

The problem addressed in this project is to design a single ultrasound assistant for the OB/GYN department that covers its common checks, rejects unusable images, shows where each finding is, verifies its own heatmaps, and refers uncertain cases to a gynaecologist with a clear reason while keeping the error among accepted cases below a chosen limit.

3. Motivation and Need for the Project

Evidence from Recent Studies

The Nature Medicine study by Christiansen et al. showed that AI can match or exceed expert examiners in classifying ovarian tumours on ultrasound and, in a simulated triage, can reduce referrals to experts by 63 percent [6]. This shows the value of AI support in gynaecology, but the dataset is not public and the study does not describe image-quality checks or heatmap verification. Smaller studies on public data report accuracy of 90 to 99 percent [7], [23], yet a 2026 study found that only 225 of 4,668 images in a popular PCOS dataset were unique [3], which means that such numbers can be misleading. Recent ovarian studies suggest sending uncertain cases to experts but do not test it [10].

Workload and Skill Gap

Ultrasound demand in the OB/GYN department is high because every pregnancy needs several scans and gynaecological complaints are common. Senior sonologists are few, especially outside large cities, so a large share of first examinations is performed by less experienced staff. An assistant that gives immediate feedback on image quality, highlights findings and points out which cases need a senior opinion can support these staff and help the senior doctor focus on the difficult cases.

Research Opportunity

Methods for uncertainty estimation [9], out-of-distribution detection [13], selective prediction and conformal risk control [1] are mature in machine learning research, and deferral to clinicians has been shown to help in medical screening [8]. However, the literature review in Chapter 2 found no OB/GYN ultrasound system that combines these methods with heatmap verification across several conditions. There is therefore a clear need, and a publishable research opportunity, for an assistant that brings the

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

common OB/GYN ultrasound checks together, evaluates them honestly with duplicate-free, patient-level splits, and adds a principled safety layer that knows when to stop and ask a gynaecologist. Figure 1.2 contrasts this approach with a conventional AI tool.

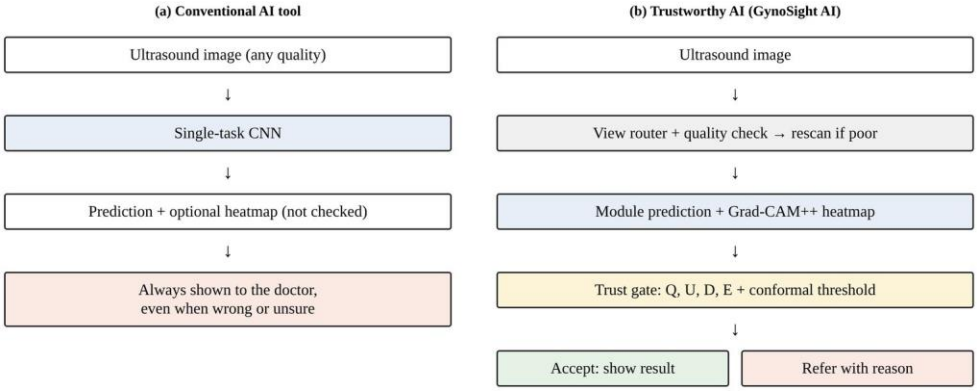


Figure 2 Conventional AI tool compared with the proposed trustworthy approach

4. Aim and Objectives

Aim

To design a trustworthy explainable ultrasound assistant for the obstetrics and gynaecology department that analyses fetal growth, ovarian masses, PCOS and uterine fibroids, highlights each finding with a verified heatmap, and safely refers unreliable cases to a gynaecologist with a reason.

Specific Objectives

- To analyze existing deep learning methods for OB/GYN ultrasound and for explanation and safe referral, and identify their limitations.
- To design a view router that identifies fetal, ovarian and uterine views and sends each image to the correct module.
- To design four analysis modules: fetal growth (plane recognition, head circumference and amniotic fluid level), ovarian mass classification, PCOS detection and fibroid detection.
- To design a heatmap module using Grad-CAM++ and a heatmap plausibility score that measures how much of the heatmap lies on the target structure.
- To design a four-signal trust gate (quality, uncertainty, familiarity and heatmap plausibility) whose threshold is set by conformal risk control.
- To plan an honest evaluation with duplicate removal, patient-level splits and testing on images from an unseen hospital, to be carried out in Phase–II.

Overview of the Proposed Solution

Figure 1.3 gives a simple overview of what GynoSight AI does with one ultrasound image, from upload to the final report. Step 7, the trust check, is the main new contribution of this project.

The image is first cleaned and its view is identified. Images of an unsupported type or of poor quality are rejected immediately with a message. Accepted images are analyzed by the matching module, which produces a result and a heatmap. The trust check then combines four signals; trusted results are shown with the heatmap, and untrusted ones are referred to the gynaecologist with a reason. Every case can be saved as a PDF report.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

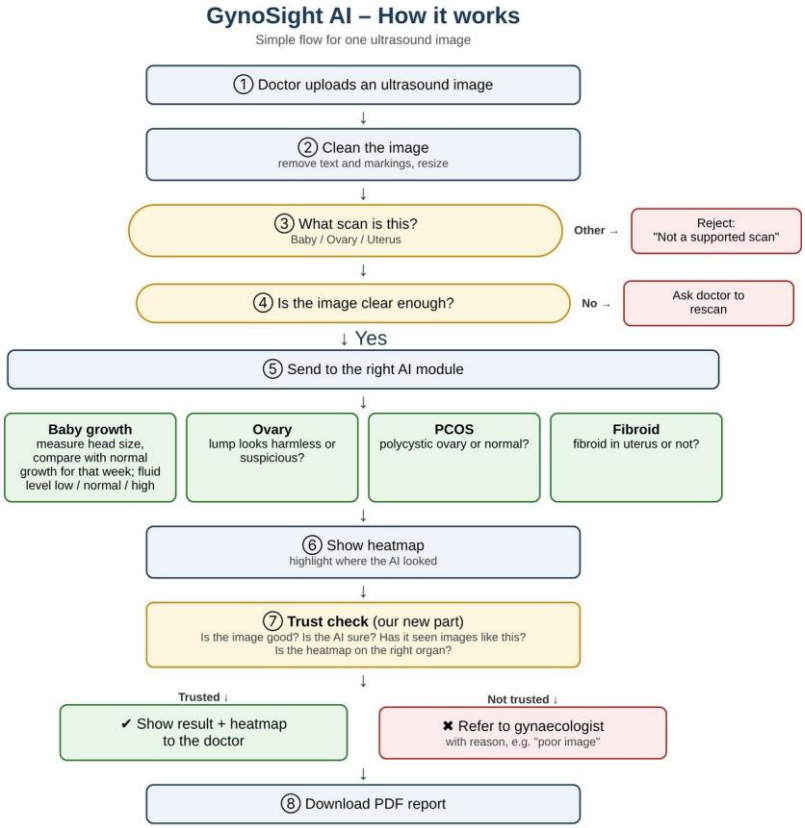


Figure 3 Simple overview of how GynoSight AI works

5. Scope and Limitations

Scope of Phase–I

- Literature survey of AI for OB/GYN ultrasound and of explainability, uncertainty and selective prediction methods.
- Identification of research gaps and definition of the problem, requirements and constraints.
- Selection of public datasets for the four modules and design of the data pipeline.
- Design of the system architecture, modules, algorithms, user interface and evaluation plan.
- Detailed plan, timeline and technology stack for Phase–II.

Scope of Phase–II

- Implementation of the data pipeline, view router, quality check, four modules, heatmaps and trust gate.
- Development of the Streamlit dashboard and PDF report.
- Evaluation on public datasets, including duplicate-free splits and an unseen-hospital test.
- Preparation of the final report and an IEEE conference paper.

Limitations

- Implementation, training, testing and results are reserved for Phase–II. No performance results are reported in this document.
- The system is a decision-support prototype for research and education, not a certified medical device, and does not replace a gynaecologist.
- Only public, de-identified still images are used; live video scanning, endometrial thickness measurement and automatic amniotic fluid measurement from images are kept as future work

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

because suitable public data was not found. In Phase–II, the amniotic fluid level is classified from the amniotic fluid index (AFI) entered by the doctor, not measured by AI.

- Fetal growth assessment uses the head circumference only; full biometry (abdominal circumference, femur length, estimated fetal weight) is future work.
- The conformal guarantee holds only when new images resemble the calibration images; performance under strong distribution shift will be measured but cannot be guaranteed.

6. Organization of the Report

Chapter 1 introduces the domain, the clinical conditions, the problem, motivation, aim, objectives and scope. Chapter 2 gives background on deep learning and trustworthy AI, reviews seven key studies, compares them and identifies the research gap. Chapter 3 analyses the problem, compares the existing and proposed systems, and lists stakeholders, use cases, requirements, constraints, risks and feasibility. Chapter 4 presents the proposed methodology, architecture, data flow, workflow, modules, model architectures, training strategy, algorithms, interface design, tools and evaluation plan. Chapter 5 gives the work breakdown, schedule, deliverables, expected outcomes and the detailed Phase–II plan with its timeline, technology stack, testing and deployment strategy. Chapter 6 concludes the report and outlines future work, followed by the references and appendices.

Figure 1.4 shows how the chapters build on one another. The first three chapters establish the problem and the requirements, Chapter 4 answers them with a design, Chapter 5 plans how the design will be built and tested, and Chapter 6 closes the report.

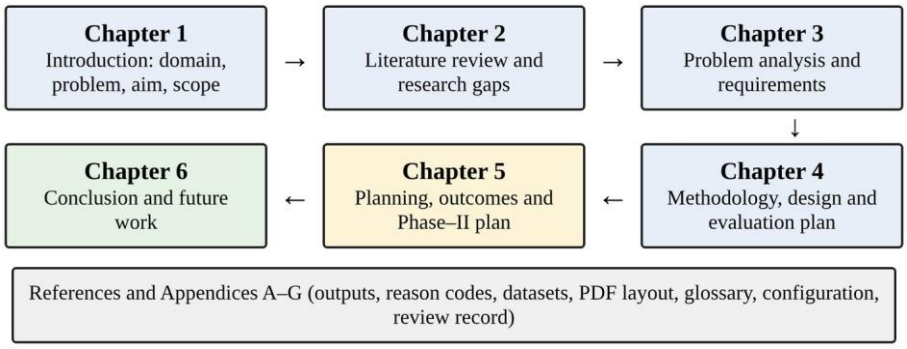


Figure 4 Organization of the report

II. LITERATURE REVIEW

1. Review of Existing Approaches

This section first gives the technical background needed to understand the reviewed work. It then reviews seven key studies covering the four planned modules (fetal growth, ovarian masses, PCOS and fibroids) and recent work on uncertainty and referral in gynaecological ultrasound. Finally, it reviews the supporting techniques for explanation, uncertainty, unfamiliar-input detection, safe referral and image quality that are used in the design.

Deep Learning Background for Medical Image Analysis

A convolutional neural network processes an image through a sequence of layers. Convolution layers slide small learnable filters over the image to detect local patterns such as edges and textures; activation functions such as ReLU add non-linearity; pooling layers reduce the spatial size; and deeper layers combine simple patterns into complex shapes. A final fully connected layer with a softmax function converts the learned features into class probabilities. Figure 2.1 illustrates this pipeline and the transfer learning strategy used throughout the reviewed work.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

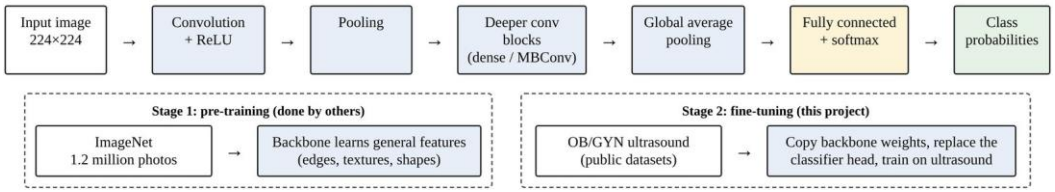


Figure 5: Convolutional neural network pipeline and transfer learning

Important CNN families include VGG, which showed the benefit of depth using small 3×3 filters; ResNet, which introduced residual (shortcut) connections that allow very deep networks to be trained; DenseNet, in which every layer receives the feature maps of all earlier layers [12]; and EfficientNet, which scales depth, width and input resolution together for a better accuracy–efficiency trade-off [20]. For segmentation, that is, labelling every pixel, U-Net uses an encoder–decoder shape with skip connections and works well with limited medical data [17].

Medical datasets are small compared with natural image datasets. Transfer learning addresses this: a network is first trained on ImageNet, a large database of over a million labelled photographs, and its weights are then fine-tuned on the medical images. The early layers, which detect general features, are reused, and only the later layers need to adapt to ultrasound. Almost all of the studies reviewed below use this approach.

This project reviews seven key studies in detail. For each study the objective, data, method, findings and limitations are summarized, followed by its relevance to GynoSight AI.

Automated Measurement of Fetal Head Circumference Using 2D Ultrasound Images
Author(s): T. L. A. van den Heuvel, D. de Bruijn, C. L. de Korte and B. van Ginneken [22]
Year: 2018
Objective. To measure the fetal head circumference automatically from two-dimensional ultrasound images, a measurement that is normally made by hand by a sonographer placing an ellipse on the skull.
Data. The study introduced the HC18 dataset of 1,334 ultrasound images of the fetal head from 551 pregnant women, acquired across all trimesters, each with a manually drawn head outline and the pixel size needed to convert pixels to millimetres.
Method. The authors proposed an automatic method that classifies pixels belonging to the fetal skull and fits an ellipse to them, from which the head circumference is computed. The dataset was released as a public challenge so that other methods could be compared on the same images.

Findings and limitations. The measured head circumference was close to that of experienced sonographers, and the dataset has since become a widely used benchmark. However, the method only measures; it does not check whether the image is of acceptable quality or whether the measurement should be trusted, and it does not relate the measurement to the gestational age.

Relevance. GynoSight AI uses HC18 to train its head outline model and adds a quality check, a reliability check and a comparison with international growth standards [15].

Evaluation of Deep Convolutional Neural Networks for Automatic Classification of Common Maternal Fetal Ultrasound Planes
Author(s): X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz et al. [4]
Year: 2020
Objective. To evaluate whether deep CNNs can recognize the standard maternal–fetal ultrasound planes used in routine pregnancy scans, a first step towards automatic biometry.

Data. The work released FETAL_PLANES_DB, a dataset of about 12,400 images from 1,792 patients collected at two hospitals by several operators on different machines. Images are labelled into the fetal abdomen, brain, femur and thorax, the maternal cervix and other views, with the brain further divided into three standard views.
Method. Many CNN architectures were compared under the same training protocol, and their performance was compared with that of human experts.
Findings and limitations. The best networks reached performance close to that of human experts. The study is important because the data comes from two hospitals and is publicly available. Its limitation is that it only recognizes the plane; it does not measure growth, check heatmaps or refer uncertain images. Relevance. GynoSight AI uses FETAL_PLANES_DB for its fetal plane classifier and for an unseen-hospital test, which most OB/GYN studies cannot perform.

Deep Learning Based Uterine Fibroid Detection in Ultrasound Images
Author(s): H. Xi and W. Wang [23]
Year: 2024
Objective. To detect uterine fibroids automatically in ultrasound images and so support early diagnosis.
Data. The study used a public dataset of 1,990 uterine ultrasound images labelled as fibroid or non-fibroid [24], which was augmented to about 10,000 images for training.
Method. An EfficientNet-B0 network with an added attention mechanism was fine-tuned on the augmented data.
Findings and limitations. An accuracy of 0.99 was reported. However, the training curves suggested overfitting, which the authors themselves noted, and the work includes no heatmap, no uncertainty estimate and no external test. Because augmented copies of the same image can end up in both training and test sets, very high accuracy on such data should be treated with caution.
Relevance. GynoSight AI uses the same public dataset for its fibroid module, but with duplicate checking, splitting before augmentation, Grad-CAM++ heatmaps and the trust gate.

Leveraging AI for Automatic Classification of PCOS Using Ultrasound Imaging
Author(s): A. Divekar and A. Sonawane [7]
Year: 2024
Objective. To classify ovarian ultrasound images as healthy or showing PCOS as part of the Auto-PCOS classification challenge.
Data. The work used the PCOSGen dataset [11], which contains ultrasound frames labelled as healthy or unhealthy and collected from online sources with the help of a gynaecologist.
Method. InceptionV3 with transfer learning was compared with ResNet101, EfficientNet-B7 and a RadImageNet DenseNet, and LIME was used to explain a few predictions.
Findings and limitations. InceptionV3 achieved 90.5 percent accuracy and placed third in the challenge. The limitations include data drawn from web videos, a class imbalance, explanations shown for only a few examples, and no uncertainty estimate.
Relevance. GynoSight AI uses PCOSGen for its PCOS module but removes duplicate frames first, following the finding of Bhuiyan et al. [3].

International Multicenter Validation of AI-Driven Ultrasound Detection of Ovarian Cancer
Author(s): F. Christiansen, E. Konuk, A. R. Ganeshan, R. Welch et al. [6]
Year: 2025
Objective. To test whether AI models can distinguish benign from malignant ovarian lesions on ultrasound across many hospitals and countries, and whether they could reduce the number of referrals to expert examiners.
Data. More than 17,000 ultrasound images from 3,652 patients collected at 20 hospitals in eight countries.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Method. Neural network models were trained and validated across centres and compared with expert and non-expert examiners. A simulated triage tested how many cases could be handled without an expert.
Findings and limitations. The models reached an accuracy of 86.3 percent compared with 82.6 percent for experts and 77.7 percent for non-experts, and the simulated triage reduced referrals to experts by 63 percent. This is strong evidence that AI can support gynaecological ultrasound. However, the data is not public, and the published summary does not describe image-quality checks, heatmap verification or a guaranteed error limit for the triage.
Relevance. GynoSight AI builds on the triage idea but makes the referral rule explicit, multi-signal and error-controlled, and explains each referral.

Early-Fusion Hybrid CNN-Transformer Models for Multiclass Ovarian Tumor Ultrasound Classification
Author(s): I. Garcia-Atutxa, J. Martínez-Más, A. Bueno-Crespo and F. Villanueva-Flores [10]
Year: 2025
Objective. To classify ovarian tumours into histopathological categories from ultrasound and to study how prediction uncertainty relates to errors.
Data. 1,469 ovarian tumour ultrasound images from the public MMOTU dataset [25], with eight tumour categories, using patient-level cross-validation.
Method. EfficientNet-B7 and a Swin Transformer were combined by early fusion. Prediction entropy was analyzed as an uncertainty measure, and calibration was applied.
Findings and limitations. Errors were rare among low-entropy cases and much more common among the most uncertain cases, and the authors proposed automating about 90 percent of cases and sending the most uncertain 10 percent to experts. However, this triage was not tested, Grad-CAM was judged only visually, no out-of-distribution or quality check was used, and all data came from one centre.
Relevance. GynoSight AI turns this suggestion into a tested design: a trust gate with four signals and a conformal threshold, with heatmaps checked against the MMOTU tumour outlines.

PelviNeXt: A Modality-Agnostic Hybrid Network for Pelvic Imaging in Women's Health
Author(s): S. T. Bhuiyan, R. Rahman, S. Wasi et al. [3]
Year: 2026
Objective. To build a single hybrid network for women's pelvic imaging that works across imaging types.
Data. PCOS ultrasound images (PCOSGen) and pelvic X-ray images. Before training, the authors audited PCOSGen for exact and near-duplicate images.
Method. A hybrid convolution and attention network was trained on the de-duplicated data.
Findings and limitations. Only 225 of the 4,668 PCOS images were found to be unique. On the de-duplicated data the model reached 92 percent accuracy. The authors note the very small dataset, the absence of cross-site evaluation, qualitative-only Grad-CAM and no uncertainty, abstention or out-of-distribution handling.
Relevance. This finding shows why GynoSight AI includes duplicate removal in its data pipeline and why the trust gate is needed for small, noisy OB/GYN datasets.

The remaining sub-sections review the supporting techniques used in the design. Figure 2.2 groups them into five families.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

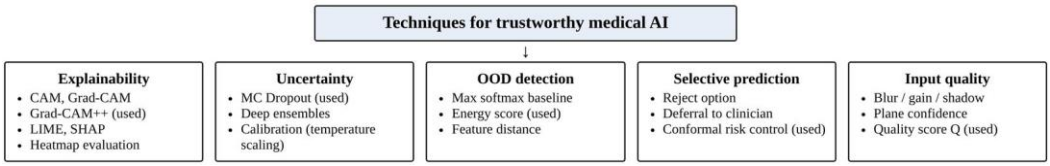


Figure 6: Families of techniques for trustworthy medical AI

Explainable AI and Heatmap Methods

Class activation mapping (CAM) showed that a CNN with global average pooling can highlight the image regions responsible for a class. Grad-CAM generalized this to almost any CNN by using the gradients flowing into the last convolutional layer to weight its feature maps. Grad-CAM++ uses higher-order gradient terms and localizes small or multiple regions better [5], which suits ultrasound, where findings such as follicles or small fibroids can be small or multiple.

Model-agnostic methods explain predictions without looking inside the network. LIME fits a simple local model around one prediction by perturbing the input, and SHAP assigns each input feature a contribution based on game theory. These methods are slower and, for images, depend on how the image is divided into regions.

A key concern is that heatmaps can look convincing while being wrong. Saporta et al. benchmarked several saliency methods on chest X-rays against expert-drawn regions and found that they often fell short of human localization [19]. This motivates the heatmap plausibility score in GynoSight AI, which measures how much of the heatmap falls on the expected structure instead of trusting it by eye.

Uncertainty Estimation and Calibration

Uncertainty can come from noise in the data (aleatoric uncertainty) or from the model's lack of knowledge (epistemic uncertainty). Monte Carlo Dropout keeps dropout active at test time and averages several forward passes; the spread of the predictions approximates Bayesian model uncertainty [9]. Deep ensembles train several networks independently and combine them, which often gives better uncertainty but multiplies training cost.

Calibration asks whether a predicted probability of 0.9 is correct about 90 percent of the time. Guo et al. showed that modern deep networks are often over-confident and that a simple post-processing step, temperature scaling, improves calibration. GynoSight AI uses Monte Carlo Dropout because it adds no training cost, and will report the expected calibration error of each module.

Out-of-Distribution Detection

An out-of-distribution (OOD) input is one that differs from the training data, such as an image from a different organ, a different machine or a non-ultrasound image. The maximum softmax probability is a simple baseline for detecting such inputs, but networks can give high softmax scores to unfamiliar images. The energy score, computed from the network's raw output values (logits), separates in-distribution and OOD inputs better than softmax confidence and needs no retraining [13]. GynoSight AI uses the energy score as its familiarity signal D.

Selective Prediction, Deferral and Conformal Risk Control

Selective classification allows a model to abstain on difficult inputs; the trade-off between the fraction of cases accepted (coverage) and the error on those cases (risk) is shown by a risk–coverage curve. In medicine, CoDoC learned when an AI system should defer to clinicians and showed that combining AI and clinician decisions reduced false positives in screening while keeping accuracy [8].

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Conformal prediction gives statistical guarantees that hold for any model, provided that the calibration and test data are exchangeable, that is, drawn from the same distribution. Conformal risk control extends this idea to control the expected value of any bounded loss, such as the error rate among accepted cases [1]. This is the basis of the trust threshold in GynoSight AI: the threshold is chosen on a calibration set so that the expected error among accepted cases stays below a chosen level α.

Ultrasound Image Quality Assessment

Ultrasound quality depends on the operator. Simple image measures can detect common problems: blur reduces the variance of the Laplacian (a measure of edge sharpness), wrong gain produces large areas that are too dark or saturated, and acoustic shadowing produces dark bands inside the scan sector. For fetal scans, the confidence of a standard-plane classifier also indicates whether a correct plane was captured, which is how real-time plane detection systems such as SonoNet guide the operator [2]. GynoSight AI combines these measures into a quality score Q.

Clinical Reference Standards

The INTERGROWTH-21st project published international standards for fetal growth based on serial ultrasound measurements of healthy pregnancies across several countries [15], which allow a measured head circumference to be converted into a centile for the gestational age. The four-quadrant amniotic fluid index provides a simple, widely used measure of amniotic fluid [16]. The IOTA simple rules describe ultrasound features of benign and malignant ovarian masses [21], the Rotterdam criteria define polycystic ovarian morphology as part of PCOS diagnosis [18], and the FIGO system classifies fibroids by location [14]. These standards guide the outputs and messages that GynoSight AI shows to the gynaecologist.

2. Comparative Analysis of Related Work

Comparison of Key Studies

Table 2.1 compares the reviewed works in terms of the problem addressed, method, key finding, limitation and relevance to this project.

Table 1: Comparative summary of literature

Author / Year	Problem addressed	Method / technology	Key finding	Limitation / gap	Relevance to this project
van den Heuvel et al., 2018 [22]	Fetal head circumference	Pixel classifier + ellipse fitting; HC18 dataset	HC close to sonographers	Measurement only; no quality or trust check	Data and baseline for head outline
Burgos-Artizzu et al., 2020 [4]	Fetal plane recognition	Many CNNs on 12,400 images, two hospitals	Close to human experts	Planes only; no heatmap check or referral	Plane classifier and unseen-hospital test
Xi and Wang, 2024 [23]	Fibroid detection	Attention EfficientNet-B0	99% accuracy reported	Possible overfitting; no heatmap or uncertainty	Dataset for fibroid module
Divekar and Sonawane, 2024 [7]	PCOS detection	InceptionV3, LIME	90.5% accuracy	Web-video frames; few explanations; no uncertainty	Dataset for PCOS module

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Christiansen et al., 2025 [6]	Benign vs malignant ovarian lesions	Neural networks, 20 centres, 8 countries	86.3% vs 82.6% experts; 63% fewer referrals	Private data; no quality or heatmap check described	Supports AI triage; we add error control
Garcia-Atutxa et al., 2025 [10]	Ovarian tumour type	EfficientNet-B7 + Swin; entropy	Uncertain cases hold most errors	Triage untested; visual Grad-CAM; one centre	Basis for tested trust gate
Bhuiyan et al., 2026 [3]	PCOS and pelvic X-ray	Hybrid network; de-duplication	Only 225 of 4,668 PCOS images unique	Tiny data; no uncertainty or abstention	Motivates duplicate removal

The comparison shows that existing work either targets a single OB/GYN task, reports accuracy without checking image quality or heatmaps, or proposes expert referral without testing it or bounding its error. None combines the department's common checks with heatmap verification and a guaranteed, explained referral rule.

Comparison of Supporting Techniques

Table 2.2 compares the candidate techniques for each trust signal and shows which one was selected and why.

Table 2 Comparison of candidate techniques for the trust signals

Signal	Candidate techniques	Strengths and weaknesses	Selected for GynoSight AI
Explanation	CAM, Grad-CAM, Grad-CAM++, LIME, SHAP	CAM needs a special architecture; LIME and SHAP are slow; Grad-CAM++ is fast and handles small or multiple regions	Grad-CAM++ [5]
Uncertainty	Softmax entropy, MC Dropout, deep ensembles	Entropy is over-confident; ensembles multiply training cost; MC Dropout adds no training cost	MC Dropout [9]
Familiarity (OOD)	Max softmax, energy score, feature distance	Max softmax is weak; feature distance needs class statistics; energy is simple and stronger	Energy score [13]
Image quality	Expert scoring, learned quality network, simple measures	Expert scores are not available for public data; simple measures are explainable	Sharpness, gain, shadow, coverage, plane confidence

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Referral threshold	Fixed confidence cut-off, selective classification, conformal risk control	Fixed cut-offs give no guarantee; conformal risk control bounds the expected error	Conformal risk control [1]

Feature-wise Comparison with the Proposed System

Table 2.3 compares the reviewed studies with the proposed system feature by feature. A tick (✓) means the feature is present, a dash (–) means it is absent and "P" means it is partial or only suggested.

Table 3: Feature-wise comparison of existing studies and the proposed system

Feature	[22]	[4]	[23]	[7]	[6]	[10]	Ours
More than one OB/GYN task	–	–	–	–	–	–	✓
Public data	✓	✓	✓	✓	–	✓	✓
Heatmap shown	–	–	–	P	–	P	✓
Heatmap checked	–	–	–	–	–	–	✓
Uncertainty	–	–	–	–	–	✓	✓
Image-quality check	–	–	–	–	–	–	✓
OOD check	–	–	–	–	–	–	✓
Referral with error bound	–	–	–	–	P	P	✓
Reason for referral	–	–	–	–	–	–	✓
Multi-site test	–	✓	–	–	✓	–	✓

3. Research / Design Gap Identified

Gaps in Existing Work

- Single task: each tool covers only one OB/GYN check [4], [22], [23],
- [7].
- No input checking: no image-quality or unfamiliar-scan check, although ultrasound quality depends on the operator [7], [23].
- Unverified heatmaps: heatmaps are missing or judged by eye [3], [10], [23].
- Untested or unbounded referral: expert referral is simulated or suggested, without an error guarantee or reason [6], [10].
- Optimistic evaluation: small, single-source and duplicate-heavy data inflate accuracy [3], [23].

How GynoSight AI Addresses the Gaps

GynoSight AI addresses these gaps with one assistant covering four modules behind a view router, quality and familiarity checks, a heatmap plausibility score, a four-signal trust gate with conformal risk control and reason codes, and an evaluation plan with duplicate removal, patient-level splits and an unseen-hospital test. Table 2.4 maps each gap to the design element that addresses it.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 4: Mapping of research gaps to proposed solutions

Research gap	Proposed solution	Where described
Single task	View router and four modules in one assistant	Sections 4.2 and 4.4
No input checking	Quality score Q and energy-based familiarity score D	Sections 4.4.3 and 4.5
Unverified heatmaps	Heatmap plausibility score E against target masks	Sections 4.4.8 and 4.5
Untested or unbounded referral	Trust gate with conformal risk control and reason codes R1–R5	Section 4.4.8, Appendix 2
Optimistic evaluation	De-duplication, patient-level splits, unseen-hospital test	Sections 4.4.1 and 4.7

4. Summary

The literature shows that deep learning can match experts on individual OB/GYN ultrasound tasks and that AI triage can reduce the load on experts. It also shows that heatmaps are unreliable, that network confidence is often miscalibrated, that uncertainty-based referral has been proposed but not tested in this area, and that public datasets need careful cleaning. Mature techniques exist for each missing piece, namely Grad-CAM++ for explanation, Monte Carlo Dropout for uncertainty, the energy score for familiarity and conformal risk control for a guaranteed referral threshold, but no OB/GYN system combines them. These insights lead directly to the design of GynoSight AI.

Figure 2.3 places the reviewed work on a timeline. It shows that the building blocks needed for a trustworthy assistant, namely segmentation and classification networks, uncertainty estimation, heatmap explanations, out-of-distribution scores and conformal guarantees, have appeared over the last decade, while OB/GYN ultrasound datasets became public mainly after 2018. GynoSight AI is positioned at the point where these two lines of work can be joined.

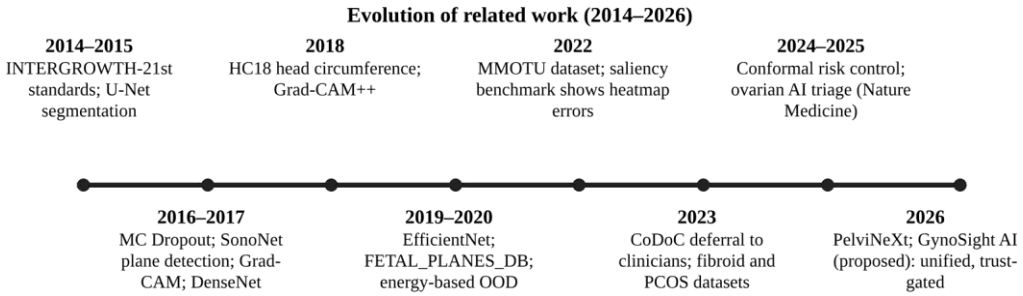


Figure 7 Evolution of the related work from 2014 to 2026

III. PROBLEM ANALYSIS AND REQUIREMENTS

1. Detailed Problem Definition

Real-World Scenario

Consider a district hospital where one senior gynaecologist supervises antenatal scans and gynaecological ultrasound performed by junior doctors and sonographers. On a busy day the department may perform dozens of scans: routine growth scans for pregnant women, follow-up scans for women with suspected growth problems, and pelvic scans for pain, bleeding or infertility. An assistant that pre-reads the images could flag a small fetal head for the gestational age, a suspicious ovarian mass or a fibroid, but only if the gynaecologist can trust it. If it gives a confident answer on a shadowed image, or highlights the bladder as evidence of a tumour, it could mislead rather than help.

The gynaecologist therefore needs three things from such an assistant: a clear result with a visual explanation when the assistant is reliable, an immediate request to rescan when the image is not usable,

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

and an honest statement of "please review this yourself, and here is why" when the assistant is not sure. The design problem is to provide these three outcomes in a principled and measurable way.

Formal Problem Formulation

The problem is defined as follows. Given an ultrasound image x from the OB/GYN department, the system must (i) identify the view (fetal, ovarian or uterine), (ii) check image quality, (iii) run the appropriate module to classify or measure, (iv) produce a heatmap or outline of the finding, (v) compute a trust score from several independent signals, and (vi) either accept the result or refer the case to a gynaecologist with a reason, so that the error among accepted cases does not exceed a chosen limit α on data similar to the calibration data.

Formally, let f be the module that maps an image to a prediction $\hat{y} = f(x)$, and let T(x) be the trust score. For a threshold λ, the system accepts the case if $T(x) \geq \lambda$ and refers it otherwise. The coverage is the fraction of cases accepted, and the risk is the error rate among accepted cases. The goal is to choose λ so that the expected risk is at most α while keeping the coverage as high as possible. This is a selective prediction problem with a statistical guarantee.

System Inputs and Outputs

Table 3.1 lists the inputs that the system receives and the outputs that it produces for each case.

Table 5: System inputs and outputs

Type	Item	Description
Input	Ultrasound image	PNG, JPEG or DICOM still image from a fetal, ovarian or uterine scan
Input (optional)	Gestational age	Weeks of pregnancy, entered for fetal scans to compute the growth centile
Input (optional)	Amniotic fluid index	AFI in centimetres, measured by the sonographer, for the fluid level check
Output	View and quality	Detected view and quality status, or a rejection message
Output	Finding	Class and probability, or head circumference and centile; fluid level
Output	Explanation	Grad-CAM++ heatmap or head outline overlaid on the image
Output	Trust decision	Accepted or referred, with the four signal values and a reason code
Output	Report and log	Downloadable PDF report and an entry in the audit log

2. Stakeholders / Application Context

Existing System and Its Drawbacks

At present, ultrasound images in most OB/GYN departments are read manually. The sonographer acquires the images and makes measurements, and the gynaecologist reviews them and writes the report. Where AI tools are used, they are usually separate products or research prototypes, each for one task. The main drawbacks of this situation are:

- Reading depends heavily on the skill and experience of the operator, and experienced staff are few.
- Each AI tool covers only one condition, and none checks image quality before predicting.
- AI tools give an answer for every image, including blurred or unusual ones, with no indication of when they may be wrong.
- Heatmaps, where present, are not verified, and no reason is given when a case needs an expert.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Reported accuracies are often based on small, single-hospital or duplicate-heavy datasets.

Proposed System and Its Advantages

The proposed system, GynoSight AI, is a single assistant that pre-reads OB/GYN ultrasound images and supports the gynaecologist, who always makes the final decision. Figure 3.1 compares the existing and proposed systems.

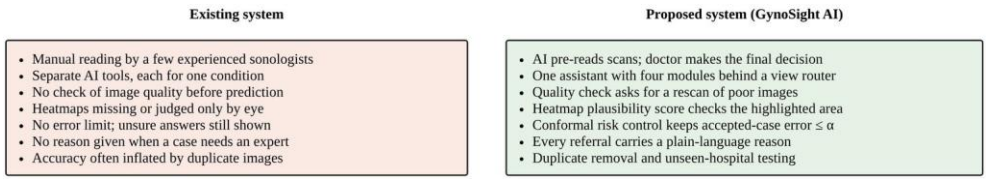


Figure 8 Existing system compared with the proposed system

The main advantages of the proposed system are that it covers four common checks in one interface, asks for a rescan of poor images, checks whether its heatmaps are on the right structure, keeps the error among accepted results below a chosen limit, gives a reason for every referral and is evaluated honestly on cleaned data and an unseen hospital. Table 3.2 summarizes the comparison.

Table 6: Comparison of the existing and proposed systems

Aspect	Existing system	Proposed system
Coverage	One tool per condition	Four modules behind one view router
Image quality	Not checked	Quality score Q; rescan request
Unfamiliar images	Not detected	Energy-based familiarity score D
Explanation	Missing or unverified heatmap	Grad-CAM++ heatmap with plausibility score E
When unsure	Answer shown anyway	Referred to the gynaecologist with a reason
Error control	None	Conformal risk control, error ≤ α among accepted cases
Evaluation	Often duplicate-heavy, single site	De-duplicated, patient-level, unseen-hospital test

Stakeholders

Table 3.3 lists the stakeholders of the system, their role and what they expect from it.

Table 7 Stakeholders and their expectations

Stakeholder	Role	Expectations
Gynaecologists and obstetricians	Primary users and final decision makers	Reliable results, clear heatmaps, honest referrals with reasons
Sonographers and junior doctors	Acquire and upload images	Immediate feedback on image quality and view
Pregnant women and patients	Indirect beneficiaries	Faster, safer and more consistent reporting
Hospital administrators	Decide on adoption	Reduced backlog, audit trail, low cost
Medical students and trainees	Learners	Visual examples of standard planes and findings
Project team and guide	Developers and supervisors	A working, well-evaluated prototype and a publication

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Application Environment

The intended environment is an OB/GYN department or teaching laboratory, with the system accessed through a web browser on a local network and running on a GPU workstation or, more slowly, on a CPU. Images are exported from the ultrasound machine as still images or DICOM files and uploaded through the dashboard. Because no patient data leaves the local network, the system can be used for teaching and research without sending images to an external cloud service.

Use Case Analysis

Figure 3.2 shows the use case diagram. The sonographer or junior doctor uploads images and enters details; the gynaecologist reviews results, referrals and reports; and the developer or administrator trains and calibrates the models and sets the allowed error level α.

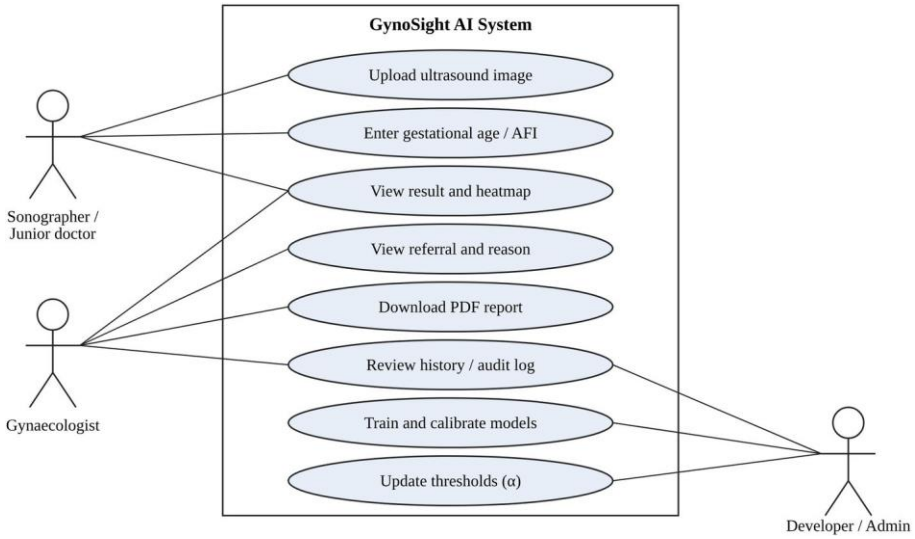


Figure 9 Use case diagram of GynoSight AI

Tables 3.4, 3.5 and 3.6 describe the three main use cases in detail.

Table 8 Use case UC1 – Analyze an ultrasound image

Field	Description
Actors	Sonographer or junior doctor (primary), gynaecologist
Pre-condition	The models and the trust threshold λ∗ are loaded
Main flow	1. User uploads an image. 2. User enters gestational age and AFI if it is a fetal scan. 3. System preprocesses the image, detects the view and checks quality. 4. System runs the module, computes the heatmap and trust signals. 5. System shows the accepted result with heatmap.
Alternative flow	3a. Unsupported view: system shows reason R5. 3b. Poor quality: system asks for a rescan (R1). 4a. Trust score below λ∗: system shows a referral with the reason (R2–R4).
Post-condition	The decision is stored in the audit log

Table 9 Use case UC2 – Review a referral and download the report

Field	Description
Actors	Gynaecologist

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Pre-condition	UC1 has produced a result or a referral
Main flow	1. Gynaecologist opens the case. 2. System shows the image, heatmap, result, signal values and reason. 3. Gynaecologist reviews the image and adds remarks. 4. System generates the PDF report for download.
Alternative flow	2a. For a referred case the system shows which signal failed and suggests manual review.
Post-condition	A PDF report is available and the review is logged

Table 10 Use case UC3 – Train and calibrate the models

Field	Description
Actors	Developer or administrator
Pre-condition	Datasets are downloaded and de-duplicated
Main flow	1. Split data by patient. 2. Train the router and modules. 3. Fit the trust-score weights on the validation set. 4. Choose $\lambda*$ on the calibration set for the chosen α. 5. Evaluate on the test set and save the models and configuration.
Alternative flow	5a. If the error on the test set exceeds α, investigate distribution shift and recalibrate.
Post-condition	New model checkpoints and threshold are available to the dashboard

3. Functional and Non-Functional Requirements

Functional Requirements

The functional requirements describe what the system must do. They are listed in Table 3.7.

Table 11 Functional requirements

ID	Requirement
FR1	The system shall accept ultrasound images in PNG, JPEG and DICOM formats through a web interface.
FR2	The system shall remove on-screen text and calipers, crop the ultrasound sector and resize images.
FR3	The system shall identify the view (fetal, ovarian, uterine) and route the image to the correct module.
FR4	The system shall reject images of unacceptable quality and request a rescan.
FR5	The fetal growth module shall recognize the fetal plane and, for head views, outline the head and compute the head circumference.
FR6	The system shall compare the head circumference with growth standards when the gestational age is entered, and classify the amniotic fluid level as low, normal or high when the AFI is entered.
FR7	The system shall classify ovarian masses, detect PCOS and detect fibroids with a probability for each class.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

FR8	The system shall generate a Grad-CAM++ heatmap or outline overlay for each finding.
FR9	The system shall compute the four trust signals and accept or refer each case with a reason code.
FR10	The system shall generate a PDF report and log every decision with a timestamp.
FR11	The system shall allow the administrator to set the allowed error level α and reload the calibrated threshold.
FR12	The system shall display the trust signal values so that the gynaecologist can see why a case was accepted or referred.
FR13	The system shall allow previous cases in the session to be reviewed from the audit log.

Non-Functional Requirements

The non-functional requirements describe how well the system must perform. They are listed in Table 3.8.

Table 12 Non-functional requirements

ID	Category	Requirement
NFR1	Performance	Analysis of one image should complete within about 5 seconds on a GPU workstation (target, to be measured in Phase–II).
NFR2	Safety	Error among accepted cases should remain at or below the chosen limit α on in-distribution test data.
NFR3	Explainability	Every accepted result must have a heatmap or outline, and every referral a reason.
NFR4	Usability	The interface should be usable by a gynaecologist with no AI background after a short demonstration.
NFR5	Privacy	Only de-identified public data will be used; uploaded images are not stored beyond the session unless saved.
NFR6	Maintainability	Each module is independent so that new tasks (e.g. endometrial thickness) can be added later.
NFR7	Portability	The system should run on Windows and Linux with Python 3.10 or later.
NFR8	Reliability	The application should handle invalid files and missing inputs with clear messages instead of crashing.
NFR9	Reproducibility	Random seeds, data splits, model versions and configuration are recorded so that results can be repeated.

4. Constraints, Assumptions and Risks

Constraints

- Time: implementation and testing must be completed within Semester VIII (about five months).
- Compute: training will use a single consumer GPU or free cloud GPU sessions, which limit session length.
- Data: public OB/GYN datasets are small; some have no patient identifiers or region outlines.
- Regulatory: the system cannot be used for real clinical decisions without regulatory approval.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Team: four members share the work alongside other coursework.

Assumptions

- Dataset labels are sufficiently accurate for training and evaluation.
- Calibration and test data come from the same distribution, as the conformal guarantee requires; the unseen-hospital test checks what happens when this fails.
- Gestational age is entered by the user for the growth comparison, since it cannot be read reliably from a single image.
- The amniotic fluid index (AFI) is measured by the sonographer using the standard four-quadrant method and entered by the user, since no public dataset with fluid annotations was found.
- Public dataset licences permit academic research use, as stated by their publishers.

Risks and Mitigation

Table 3.10 lists the main project risks, their impact and the planned mitigation.

Table 13 Risks and mitigation plan

Risk	Impact	Mitigation
Small PCOS dataset after de-duplication	High	Strong augmentation, transfer learning, cross-validation and honest reporting of confidence intervals.
Insufficient GPU time	Medium	Transfer learning, 224×224 inputs, mixed precision and free cloud GPUs.
No region outlines for PCOS and fibroid images	Medium	Use annotation files where available; otherwise check the heatmap lies inside the ultrasound sector and uterus/ovary region.
Conformal guarantee fails under distribution shift	Medium	Test on the second hospital in FETAL_PLANES_DB and report the gap; recalibrate per site.
Data leakage between train and test	High	Duplicate removal and patient-level splits wherever patient IDs exist.
Integration delays	Low	Independent modules with fixed interfaces, tested early.
Team member unavailable	Low	Shared code repository, documentation and pair ownership of each module.

Ethical and Legal Considerations

The project uses only public, de-identified datasets released for research, and each dataset is cited as required by its licence. No new patient data is collected in Phase–II. If a local hospital later provides images for feedback, ethics committee approval and patient consent will be obtained first. The system will display a clear notice that it is a decision-support research prototype and not a diagnostic device, and that every result must be reviewed by a qualified gynaecologist. Referral reasons are written in plain language so that the doctor is never asked to trust an unexplained decision.

5. Feasibility Analysis

Technical Feasibility

All components are well established and available as open-source libraries. Pretrained DenseNet121 and EfficientNet-B0 weights are available in PyTorch, U-Net and Grad-CAM++ implementations exist, and Monte Carlo Dropout, energy scores and conformal risk control can be implemented in a

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

few hundred lines of Python. The datasets are publicly downloadable. The team has studied Python and machine learning in earlier semesters.

Operational Feasibility

The web interface needs only a browser, and the outputs (finding, heatmap, measurement, trust decision and reason) are designed for gynaecologists without AI training. The workflow matches the existing one: the sonographer uploads exported images and the gynaecologist reviews results, so no change in clinical routine is needed for a trial.

Economic Feasibility

All datasets and software are free, and training can use free cloud GPUs or a college laboratory GPU. Table 3.11 gives an approximate cost estimate.

Table 14 Approximate cost estimate

Item	Source	Estimated cost
Datasets (FETAL_PLANES_DB, HC18, MMOTU, PCOSGen, fibroid)	Public repositories	Free
Software (Python, PyTorch, Streamlit, OpenCV)	Open source	Free
GPU computing	Google Colab / Kaggle free tier or college lab	Free
Development computer	Existing student laptops / lab PCs	No new purchase
Optional cloud GPU upgrade for faster training	Paid tier, if needed	Low (optional)
Printing and binding of reports	Local	Nominal

Schedule Feasibility

The work is split into modules allocated to four members (Chapter 5), and the modules can be trained in parallel within the five-month Phase–II. The schedule includes buffer time in the final month for fixes, documentation and the research paper.

Legal and Ethical Feasibility

Because only public research datasets are used and the system is positioned as a research and teaching prototype, there are no legal barriers to Phase–II. Any future clinical use would require regulatory approval and prospective validation, which are outside the scope of this project.

IV. PROPOSED METHODOLOGY AND DESIGN

1. Overall Proposed Approach

This chapter describes the planned design of GynoSight AI; it is not an implementation report. Values such as the number of training epochs, the trust-score weights and the threshold will be finalized in Phase–II.

Design Principles

- Doctor in control: the system supports and never replaces the gynaecologist; every output is a suggestion with an explanation.
- Fail safe: when any check fails, the system refers the case instead of guessing.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Explain everything: every accepted result has a heatmap or outline, and every referral has a plain-language reason.
- Measurable safety: the error among accepted cases is controlled by a statistical method, not a hand-picked cut-off.
- Modularity: each clinical task is an independent module so that tasks can be added or improved separately.
- Honest evaluation: data is de-duplicated and split by patient, and generalization is tested on an unseen hospital.

Step-by-Step Approach

- Data collection: obtain the datasets in Table 4.2, remove duplicates and split at patient level into training, validation, calibration and test sets.
- Preprocessing: remove on-screen text and calipers, crop the ultrasound sector, resize to 224×224 and reduce speckle noise.
- View routing: train a small classifier to identify fetal, ovarian and uterine views and reject unsupported images.
- Modules: train the fetal plane classifier and head outline model, the ovarian mass classifier, the PCOS classifier and the fibroid classifier.
- Heatmaps: apply Grad-CAM++ to each classifier and use the head outline as the explanation for measurements.
- Trust signals: compute quality (Q), uncertainty (U), unfamiliarity (D) and heatmap plausibility (E).
- Trust gate: combine the signals with logistic regression and set the threshold with conformal risk control.
- User interface: display results in a Streamlit dashboard and generate a PDF report.
- Evaluation: measure classification, measurement, localization and trust performance, including an unseen-hospital test.

Development Methodology

The project follows an iterative, incremental methodology. Each module is built and tested in a short cycle: prepare data, train a baseline, evaluate, and improve. Modules are integrated as soon as they reach a working baseline, so that integration problems are found early. Code is kept in a shared Git repository with one branch per module, and every experiment records its configuration, data split and random seed so that results can be repeated. Progress is reviewed with the guide at each milestone listed in Section 5.5.

2. System Architecture / Conceptual Framework

Layered Architecture

Figure 4.1 shows the proposed architecture. An input ultrasound image is preprocessed and passed to the view router. The quality check rejects unusable images with a rescan request. Accepted images go to one of four modules. Each module produces a result and a Grad-CAM++ heatmap or measurement overlay, which also feeds the heatmap plausibility signal (E) to the trust gate. The trust gate either accepts the result or refers the case to a gynaecologist with a reason, and both paths end at the dashboard and PDF report.

- Input and preprocessing layer: standardizes images from different machines and removes on-screen markings that a model could learn as shortcuts.
- Routing and quality layer: a lightweight CNN decides which module should handle the image, and the quality score Q decides whether the image is usable.
- Analysis layer: four independent models so that each can be trained and improved separately.
- Explanation layer: Grad-CAM++ heatmaps and head outlines show the doctor where the finding is and provide the E signal.
- Trust layer: the safety gate that decides whether a result can be shown or must be referred.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Presentation layer: the dashboard and report present the result and keep a record.

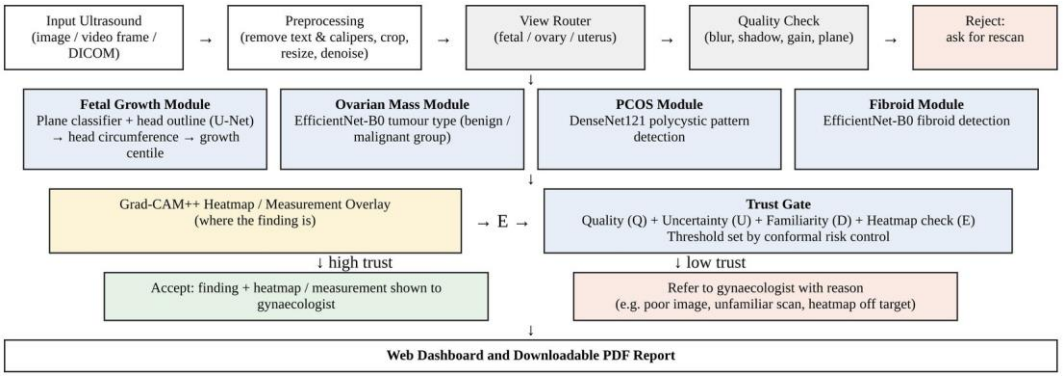


Figure 10 Proposed system architecture of GynoSight AI

Data Flow Diagrams

A data flow diagram (DFD) shows how data moves through the system. Figure 4.2 is the context-level (level 0) DFD, which treats the whole system as one process exchanging data with the user, the datasets, the audit log and the model store.

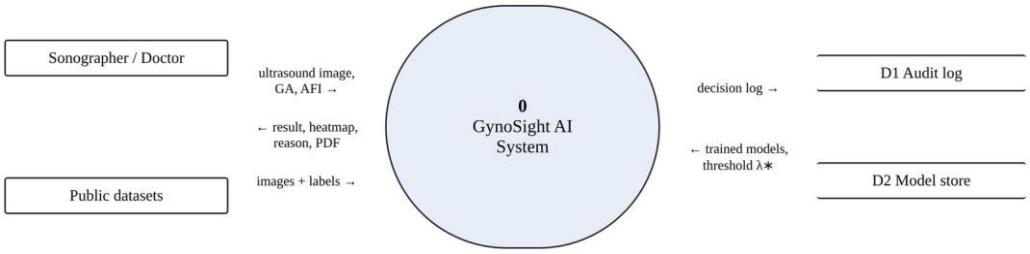


Figure 11 Level 0 (context) data flow diagram

Figure 4.3 is the level 1 DFD, which breaks the system into six processes: preprocessing, view routing and quality checking, module analysis, explanation, the trust gate, and display and reporting. Rejected images leave the flow early with a message, while all decisions are written to the audit log.

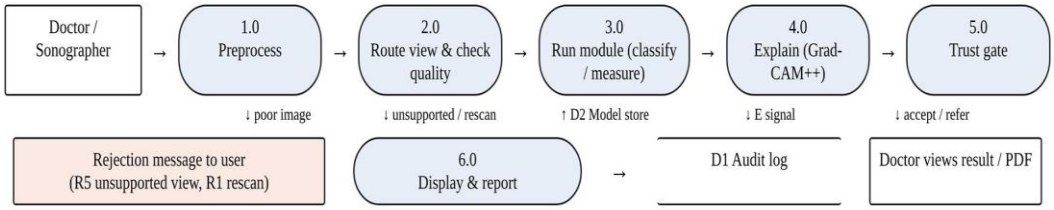


Figure 12 Level 1 data flow diagram

Deployment Architecture

Figure 4.4 shows how the system will be deployed. Models are trained and calibrated on a cloud GPU (Google Colab or Kaggle) and saved as checkpoints. The trained models, the trust-gate weights and the threshold are copied to a workstation on the hospital or laboratory network, where a Streamlit web application serves the dashboard. Doctors use the system from a browser on their own computer, and no image is sent outside the local network.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

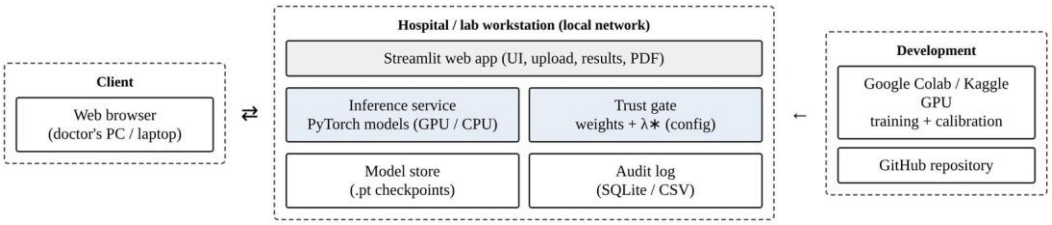


Figure 13 Planned deployment architecture

3. Process Flow / Workflow

Workflow for One Image

Figure 4.5 shows the step-by-step workflow for one image, including the two decision points: the quality check and the trust threshold.

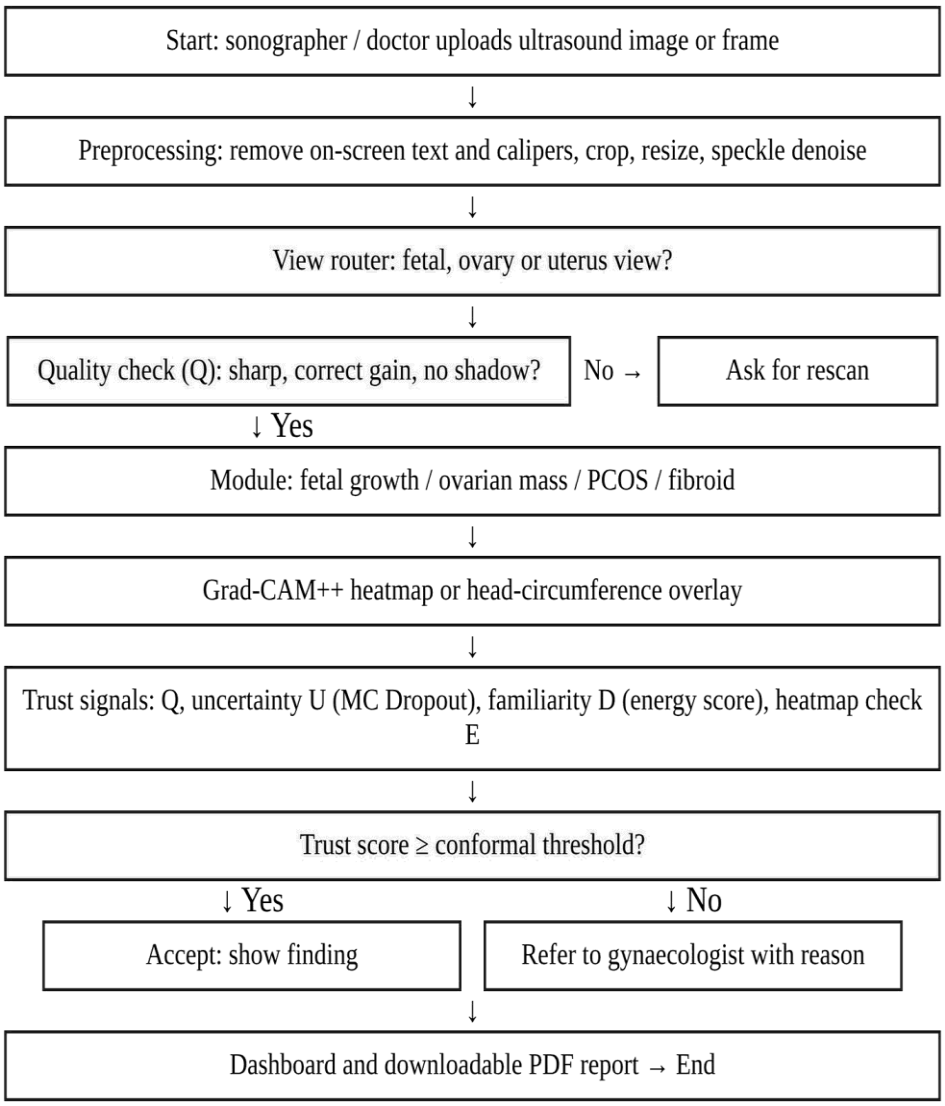


Figure 14 Proposed workflow of GynoSight AI

Sequence of Interactions

Figure 4.6 shows the sequence diagram for analyzing one image. The user uploads an image through the dashboard; the image passes through preprocessing, routing and quality checking, the selected module and the trust gate; and the decision returns to the dashboard. The alternative path shows the early rejection of unsupported or poor-quality images.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

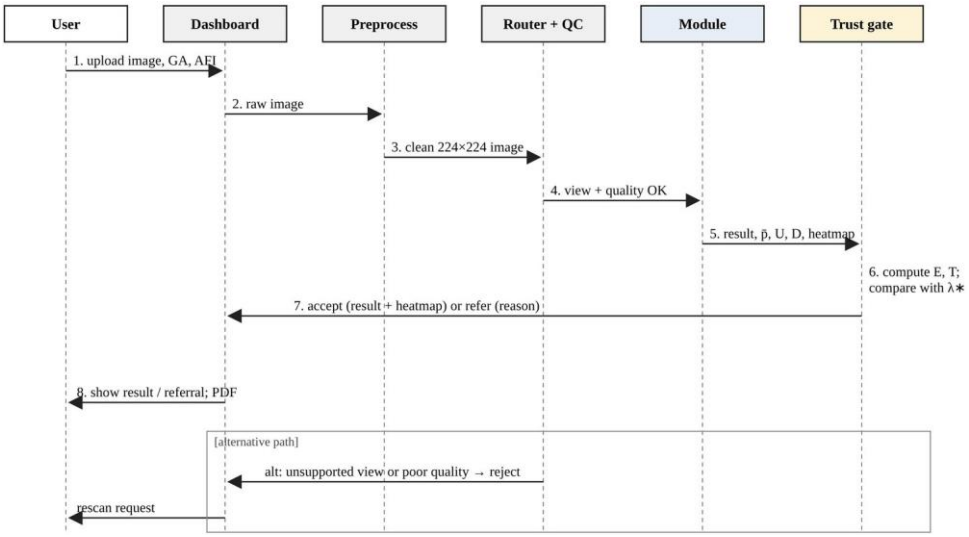


Figure 15 Sequence diagram for analyzing one image

Decision Outcomes

Every image ends in exactly one of the outcomes listed in Table 4.1.

Table 15 Possible decision outcomes for an image

Outcome	Condition	Message to the user
Rejected – view	Router predicts "unsupported" or low confidence	R5: view not supported
Rejected – quality	Quality score Q below the minimum	R1: please rescan
Referred	Trust score T below threshold λ∗	R2, R3 or R4: weakest signal explained
Accepted	Trust score T at or above λ∗	Result, heatmap and trust details

4. Module / Component Description

Data Collection and Preprocessing Module

This module organizes the datasets in Table 4.2, cleans them, splits them and prepares every image in the same way for training and inference.

Table 16 Datasets selected for each module

Module	Dataset	Size and labels	Planned use
Fetal planes	FETAL_PLANES_ DB [4]	About 12,400 images, 1,792 patients, 2 hospitals; abdomen, brain, femur, thorax, cervix, other	Plane classifier; unseen-hospital test
Fetal head size	HC18 [22]	1,334 images from 551 women with head outlines	Head outline model; HC measurement; heatmap check
Ovarian mass	MMOTU [25]	1,469 2D images with pixel-level tumour outlines; 8 tumour categories	Classifier; heatmap check against outlines

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

PCOS	PCOSGen [11]	Healthy / unhealthy frames (heavy duplication; about 225 unique per [3])	PCOS classifier after de-duplication
Fibroid	Uterine fibroid ultrasound images [24]	1,990 images, fibroid / non-fibroid, with annotation files	Fibroid classifier; heatmap check

De-duplication and splitting. Exact and near-duplicate images are removed using perceptual hashing, as recommended by Bhuiyan et al. [3]. Two images whose hashes differ by fewer than a small number of bits are treated as duplicates and only one is kept. Images are then grouped by patient wherever patient identifiers exist, and the groups are split into training, validation, calibration and test sets so that no patient appears in two sets. Splitting is done before augmentation, so augmented copies of a test image can never appear in training. Figure 4.7 shows this process and the planned proportions.

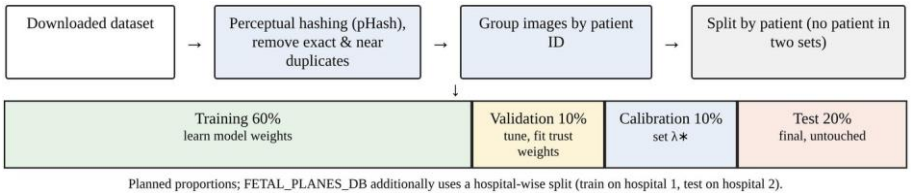


Figure 16 De-duplication and patient-level data split

Preprocessing. Figure 4.8 shows the preprocessing pipeline. Images are read (including DICOM files through pydicom) and converted to grey scale. On-screen text, calipers, logos and measurement marks are masked, because a model could otherwise learn to use them as shortcuts. The fan-shaped ultrasound sector is detected and cropped, a light median filter reduces speckle noise, and the image is resized to 224×224 pixels and normalized. The three-channel input expected by ImageNet-pretrained models is obtained by repeating the grey channel.

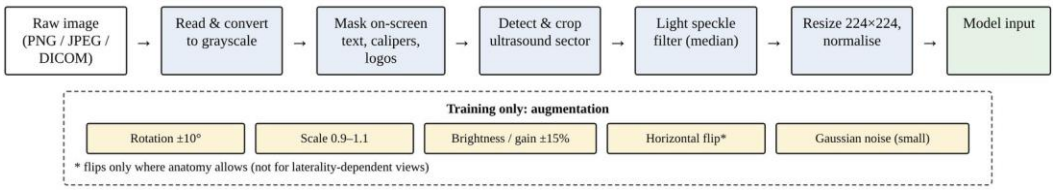


Figure 17 Preprocessing and augmentation pipeline

Augmentation. During training only, images are randomly transformed to make the models robust to normal variation between scans. Table 4.3 lists the planned augmentations.

Table 17 Planned data augmentation

Augmentation	Planned range	Purpose
Rotation	±10 degrees	Small differences in probe angle
Scaling	0.9 to 1.1	Different depth and zoom settings
Brightness and contrast	±15 percent	Different gain settings and machines
Horizontal flip	Where anatomy allows	Left–right variation in probe position
Gaussian noise	Small standard deviation	Speckle and electronic noise

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

View Router Module

The view router is a small EfficientNet-B0 classifier [20] trained to recognize fetal, ovarian and uterine views, plus an "unsupported" class trained on other ultrasound and non-ultrasound images. Fetal views come from FETAL_PLANES_DB and HC18, ovarian views from MMOTU and PCOSGen, and uterine views from the fibroid dataset. Low-confidence or unsupported images are rejected with the reason "unsupported view" (R5). Because the router only needs to separate broad view types, a light model is enough and keeps the response time low.

Quality Check Module

The quality module computes a score Q from simple measures suited to ultrasound: sharpness (variance of the Laplacian), gain (fraction of very dark and saturated pixels), acoustic shadowing (large dark regions inside the sector) and sector coverage. For fetal images the plane classifier's confidence also indicates whether a standard plane was captured. Each measure is scaled to the range 0 to 1 using its distribution on the training images and combined as shown in Equation (4.4). Images below a minimum quality are rejected with a rescan request (R1). The minimum is chosen on the validation set so that clearly unusable images are rejected without rejecting many good ones.

Table 18 Quality measures used in the quality score

Measure	How it is computed	Problem it detects
Sharpness	Variance of the Laplacian inside the sector	Blur from probe movement or poor focus
Gain	Fraction of pixels that are nearly black or saturated	Gain set too low or too high
Shadowing	Area of large connected dark regions inside the sector	Acoustic shadow from bone or gas
Coverage	Fraction of the image occupied by the scan sector	Zoomed-out or cropped images
Plane confidence	Plane classifier probability (fetal only)	Wrong or non-standard plane

Fetal Growth Module

The fetal plane classifier, based on DenseNet121 [12], recognizes the standard planes in FETAL_PLANES_DB. For trans-thalamic head views, a U-Net [17] trained on HC18 outlines the fetal skull; an ellipse is fitted to the outline and the head circumference is computed using Equation (4.3). When the gestational age is entered, the head circumference is converted into a centile using the INTERGROWTH-21st standards [15], and values below the 3rd or above the 97th centile are flagged for review.

Amniotic fluid level. The amount of amniotic fluid around the baby is an important sign of fetal well-being. Too little fluid (oligohydramnios) or too much fluid (polyhydramnios) needs closer follow-up. The sonographer measures the amniotic fluid index (AFI), the sum of the deepest vertical fluid pocket in each of the four quadrants of the uterus [16], and enters it in the dashboard. The module labels the fluid level using standard clinical cut-offs, shown in Table 4.5. A low or high level is flagged for the gynaecologist and printed in the PDF report next to the head circumference result. This check is rule-based in Phase–II; automatic measurement of the fluid pockets from the image is future work because no public annotated dataset was found.

Table 19 Fetal growth and amniotic fluid flags

Check	Rule	Output
Head circumference centile	Below 3rd centile	Flag: small head for gestational age
Head circumference centile	3rd to 97th centile	Within normal range

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Head circumference centile	Above 97th centile	Flag: large head for gestational age
Amniotic fluid index	AFI < 5 cm	Low (possible oligohydramnios)
Amniotic fluid index	5 cm ≤ AFI ≤ 24 cm	Normal
Amniotic fluid index	AFI > 24 cm	High (possible polyhydramnios)

Ovarian Mass Module

An EfficientNet-B0 [20] classifier is trained on the two-dimensional part of MMOTU to recognize the eight ovarian tumour categories provided with the dataset. For the dashboard, the categories are also grouped into a simpler benign or suspicious output, so that the gynaecologist sees both the detailed category and the clinically important grouping. Because MMOTU contains pixel-level tumour outlines, the heatmap plausibility signal can be measured directly against the tumour region. Class weights are used during training because some categories have far fewer images than others.

PCOS Module

A DenseNet121 classifier detects the polycystic ovarian pattern on the de-duplicated PCOSGen data. Because the cleaned dataset is small, five-fold cross-validation with confidence intervals will be reported, and results before and after de-duplication will be compared to show the effect of duplicates. The trust gate is expected to refer a larger share of these cases. The output is shown as "polycystic pattern present" or "absent" with a heatmap, together with a reminder that PCOS diagnosis also needs clinical and hormonal criteria [18].

Fibroid Module

An EfficientNet-B0 classifier detects uterine fibroids using the public fibroid dataset [24]. The provided annotation files are used, where suitable, to check that the heatmap falls on the fibroid region. Unlike the earlier study on the same data [23], the data is split before augmentation and duplicates are removed, so that the reported performance reflects images the model has never seen.

Trust Gate Module

The trust gate collects the four signals, combines them into a trust score T and compares it with the threshold $\lambda*$ chosen by conformal risk control. If $T \geq \lambda*$ the result is accepted; otherwise the case is referred to a gynaecologist and the weakest signal is reported as the reason, such as "image quality is low", "model is uncertain", "image is unlike the training data" or "heatmap is not on the target". Figure 4.9 shows the design.

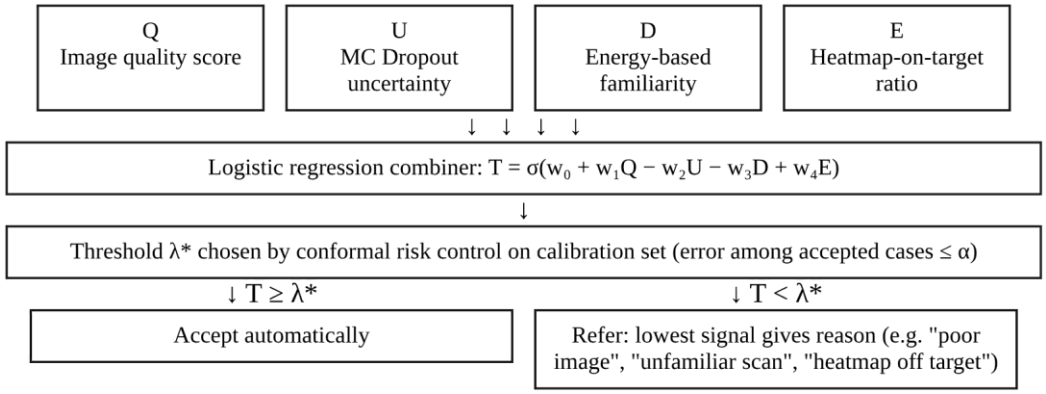


Figure 18 Design of the four-signal trust gate

The weights of the trust score are learned on the validation set by logistic regression, where the target is whether the module's prediction was correct. The threshold is then chosen on the separate calibration

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

set. Keeping these two sets separate is necessary for the conformal guarantee. To decide which signal to report as the reason, each signal is compared with its own typical range on correctly predicted validation cases, and the signal that is furthest outside its range is reported. Table 4.6 lists the reason codes.

Table 20 Referral reason codes

Code	Trigger	Message shown to the gynaecologist
R1	Low quality score Q	Image quality is too low for a reliable reading. Please rescan.
R2	High uncertainty U	The model is uncertain about this image. Please review manually.
R3	High unfamiliarity D	This image is unlike the images the model was trained on.
R4	Low heatmap plausibility E	The highlighted area is not on the expected structure.
R5	Router rejection	This view is not supported by the system.

User Interface and Report Module

A Streamlit dashboard lets the user upload an image, enter the gestational age and amniotic fluid index for fetal scans, and view the result, heatmap or head outline, trust decision and reason. A PDF report can be downloaded, and every decision is logged. Figure 4.10 shows the planned layout of the main screen.

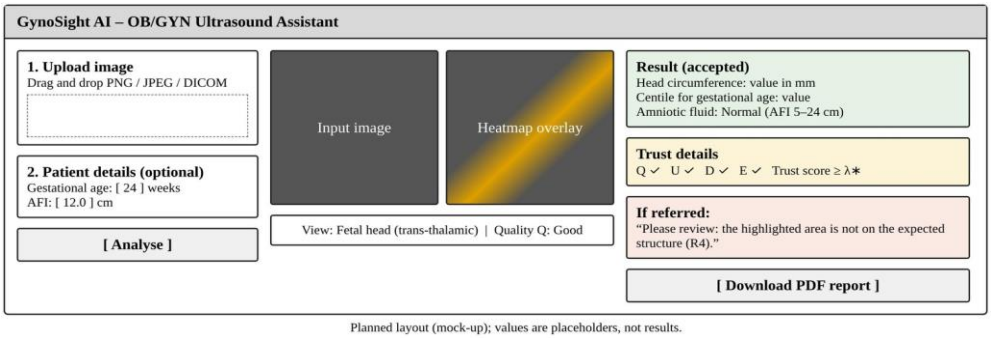


Figure 19 Planned layout of the GynoSight AI dashboard

The screen is divided into three columns: inputs on the left, the image and heatmap in the centre, and the result, trust details and actions on the right. Accepted results are shown in a green panel and referrals in a red panel, so that the status is clear at a glance. Every decision is written to an audit log with the fields listed in Table 4.7.

Table 21 Fields stored in the audit log

Field	Type	Description
case_id	Text	Unique identifier for the analysis
timestamp	Date-time	Date and time of the analysis
file_name	Text	Name of the uploaded image file
view, module	Text	Detected view and module used
prediction, probability	Text, number	Module output and its probability or measurement
Q, U, D, E, T	Number	The four trust signals and the trust score
decision, reason_code	Text	Accepted or referred, and R1–R5
model_version	Text	Version of the model and threshold used

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

5. Design Models and Algorithms

Model Architectures

Three network architectures are used. Figure 4.11 shows the building blocks of the two classification networks.

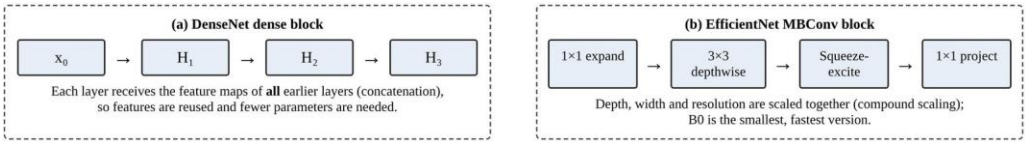


Figure 20 Building blocks of DenseNet and EfficientNet

DenseNet121 connects each layer in a dense block to every earlier layer by concatenating their feature maps [12]. This encourages feature reuse, eases gradient flow and gives strong accuracy with relatively few parameters. It is used for the fetal plane classifier and the PCOS classifier, where fine texture patterns matter. EfficientNet-B0 is built from mobile inverted bottleneck (MBConv) blocks with squeeze-and-excitation and is the smallest member of a family that scales depth, width and resolution together [20]. It is used for the view router, the ovarian classifier and the fibroid classifier because it is fast and accurate. U-Net is used to outline the fetal head. As Figure 4.12 shows, it has a contracting encoder that captures context and an expanding decoder that recovers detail, with skip connections that copy encoder features to the decoder [17]. A pretrained encoder will be used to speed up training.

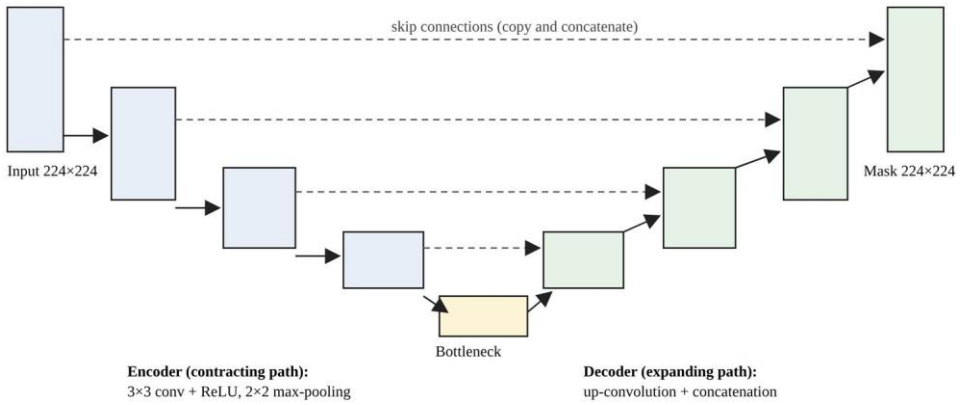


Figure 21 U-Net architecture for fetal head outlining

Table 22 Summary of planned models

Model	Architecture	Approx. parameters	Task
View router	EfficientNet-B0	5.3 million	4 classes: fetal, ovarian, uterine, unsupported
Fetal plane classifier	DenseNet121	8.0 million	Standard plane classes of FETAL_PLANES_DB
Head outline	U-Net with pretrained encoder	Depends on encoder	Pixel mask of the fetal skull
Ovarian classifier	EfficientNet-B0	5.3 million	8 tumour categories + benign / suspicious
PCOS classifier	DenseNet121	8.0 million	Polycystic pattern present / absent
Fibroid classifier	EfficientNet-B0	5.3 million	Fibroid present / absent

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Training Strategy

All classifiers start from ImageNet weights. Training proceeds in two stages: first only the new classifier head is trained with the backbone frozen, then the whole network is fine-tuned with a lower learning rate. The AdamW optimizer, a variant of Adam with decoupled weight decay, is used with a cosine learning-rate schedule and early stopping on the validation loss. Classifiers use the weighted cross-entropy loss of Equation (4.1), and the U-Net uses a combination of binary cross-entropy and the Dice loss of Equation (4.2). Table 4.9 lists the planned starting hyperparameters.

Table 23 Planned starting hyperparameters (to be tuned in Phase–II)

Hyperparameter	Planned value	Notes
Input size	224 × 224	Standard for pretrained backbones
Batch size	32 (16 for U-Net)	Fits an 8 GB GPU with mixed precision
Optimizer	AdamW	Weight decay 1e-4
Learning rate	1e-3 (head), 1e-4 (fine-tune)	Cosine decay
Epochs	Up to 50	Early stopping, patience 8
Dropout (classifier head)	0.3	Also used for MC Dropout
MC Dropout passes K	20	Trade-off between speed and stability
Allowed error α	0.05 (also 0.10 tested)	Set by the administrator

Weighted cross-entropy loss for N images, C classes, class weight w_c, true label y_{ic} and predicted probability p_{ic}:

$$L_{CE} = -(1/N) \Sigma_i \Sigma_c w_c y_{ic} \log p_{ic} \qquad (4.1)$$

Dice loss for the predicted mask P and the true mask G:

$$L_{Dice} = 1 - 2|P \cap G| / (|P| + |G|) \qquad (4.2)$$

Measurement and Trust Signal Equations

This sub-section gives the preliminary design equations for the measurement and the four trust signals.

Head circumference. For a fitted ellipse with semi-axes a and b (in millimetres, using the pixel size), Ramanujan's approximation gives:

$$HC \approx \pi [\, 3(a + b) - \sqrt{((3a + b)(a + 3b))} \,] \qquad (4.3)$$

Quality score (Q) as a weighted combination of the scaled measures in Table 4.4 (sharpness s, gain g, shadowing h, coverage v, plane confidence c), with weights β chosen on the validation set:

$$Q = \beta_1 s + \beta_2(1 - g) + \beta_3(1 - h) + \beta_4 v + \beta_5 c \qquad (4.4)$$

Uncertainty (U) using Monte Carlo Dropout. The model is run K times (planned K = 20) with dropout active:

$$\bar{p} = (1/K) \Sigma_{k=1}^{K} p_k(x), \quad U = -\Sigma_c \bar{p}_c \log \bar{p}_c \qquad (4.5)$$

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Unfamiliarity (D) using the energy score [13], for logits fc(x) and temperature τ:

$$D(x) = -\tau \cdot \log \Sigma_c \exp(f_c(x) / \tau) \quad (4.6)$$

Heatmap plausibility (E). For normalized heatmap H(p) and target mask M (tumour, head or uterus region):

$$E = \Sigma_{p \in M} H(p) / \Sigma_p H(p) \quad (4.7)$$

Combined trust score (T), learned with logistic regression on the validation set:

$$T = \sigma(w_0 + w_1 Q - w_2 U - w_3 D + w_4 E) \quad (4.8)$$

Threshold by conformal risk control [1], for n calibration cases, error rate *R̂9* n(λ) among accepted cases and loss bound B = 1:

$$\lambda* = \inf \{ \lambda : (n/(n+1)) \cdot \hat{R}_n(\lambda) + B/(n+1) \leq \alpha \} \quad (4.9)$$

If calibration and test data come from the same distribution, the expected error among accepted cases is at most α. Because this can weaken under distribution shift, the evaluation includes an unseen-hospital test.

Algorithms

The main procedures are given below as preliminary pseudo-code.

Algorithm 1: De-duplication and patient-level split

Input: image set S with patient IDs, distance limit d

- for each image x in S: h(x) ← PerceptualHash(x)
- for each pair (x, y) with Hamming(h(x), h(y)) ≤ d: keep one, drop the other
- group remaining images by patient ID (or by image if no ID)
- shuffle groups with a fixed seed
- assign groups to train 60%, validation 10%, calibration 10%, test 20%

Output: four disjoint sets with no shared patient

Algorithm 2: Training a module

Input: train and validation sets, pretrained backbone

- replace classifier head; freeze backbone; train head for a few epochs
- unfreeze backbone; train all layers with AdamW and cosine schedule
- after each epoch: compute validation loss; save best checkpoint
- stop when validation loss has not improved for 8 epochs Output: best model checkpoint

Algorithm 3: Fitting the trust gate and calibrating the threshold

Input: validation set V, calibration set C, allowed error α

- for each x in V: compute Q, U, D, E and correct(x) ∈ {0, 1}
- fit logistic regression: correct ~ Q, U, D, E → weights w
- for each x in C: compute T(x) with weights w and loss ℓ(x) = 1 − correct(x)
- for each candidate λ (from high to low): Rl (λ) ← mean loss among T(x) ≥ λ
- λ∗ ← smallest λ with (n/(n+1))·Rl (λ) + 1/(n+1) ≤ α Output: weights w and threshold λ∗

Algorithm 4: Analyzing one image (inference)

Input: ultrasound image x, gestational age GA (optional), AFI (optional), threshold λ∗

- x ← Preprocess(x)
- view, conf ← Router(x); if view = unsupported or conf is low → return Reject(R5)
- Q ← QualityScore(x); if Q < Q_min → return Reject(R1)
- module ← Module[view]; result, ph , U ← MCDropout(module, x, K)
- if view = fetal head: HC ← EllipsePerimeter(UNet(x)); centile ← INTERGROWTH(HC, GA)
- if AFI given: fluid ← Low (< 5 cm) / Normal / High (> 24 cm)
- D ← EnergyScore(module, x); H ← GradCAM++(module, x); E ← HeatmapOnTarget(H, Mask(x))
- T ← σ(w0 + w1·Q − w2·U − w3·D + w4·E)
- if T ≥ λ∗ → return Accept(result, H) else → return Refer(reason = weakest signal)

Software Class Design

The software will be written in an object-oriented style so that each part can be developed and tested separately. Figure 4.13 shows the planned class diagram. The AnalysisApp class coordinates one analysis: it calls the ViewRouter, the QualityChecker, the selected clinical module and the TrustGate, and writes the outcome to the AuditLog. All four clinical modules inherit from an abstract ClinicalModule class, which fixes a common interface for prediction, Monte Carlo Dropout prediction and Grad-CAM++ explanation. Because the TrustGate only depends on this common interface, a new module, such as endometrial thickness in future work, can be added without changing the gate.

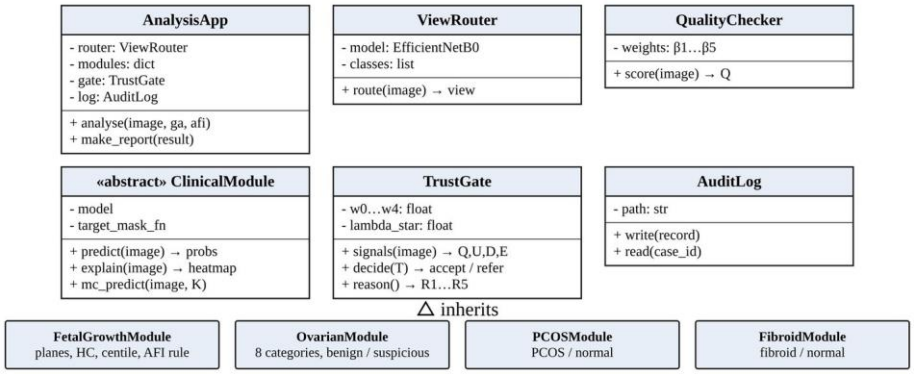


Figure 22 Planned class diagram of the GynoSight AI software

The responsibilities of the main classes are as follows.

- AnalysisApp: receives the image and the optional gestational age and AFI from the user interface, runs the pipeline in Algorithm 4 and produces the result and the PDF report.
- ViewRouter: loads the EfficientNet-B0 router and returns the view (fetal, ovarian, uterine or unsupported) with its confidence.
- QualityChecker: computes the quality score Q from the five measures in Table 4.4.
- ClinicalModule and its subclasses: load the trained network for one task, return class probabilities and heatmaps, and supply the target mask used for the plausibility score E.
- TrustGate: computes Q, U, D and E, combines them into T with the fitted weights, compares T with λ∗ and selects the reason code.
- AuditLog: appends one record per analysis and reads past records for the dashboard.

Data Storage Design

Although the prototype stores its records in a simple CSV or SQLite file, the data is organized as related entities so that it can later be moved to a hospital database. Figure 4.14 shows the planned entity–relationship diagram. Each CASE holds the upload details and the optional inputs entered by the user; each RESULT holds the module output and measurements; each TRUST_DECISION holds the four

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

signals, the trust score, the decision and the reason code; each REPORT points to the generated PDF; and MODEL_VERSION records which trained model and threshold produced the result, so that every decision can be traced.

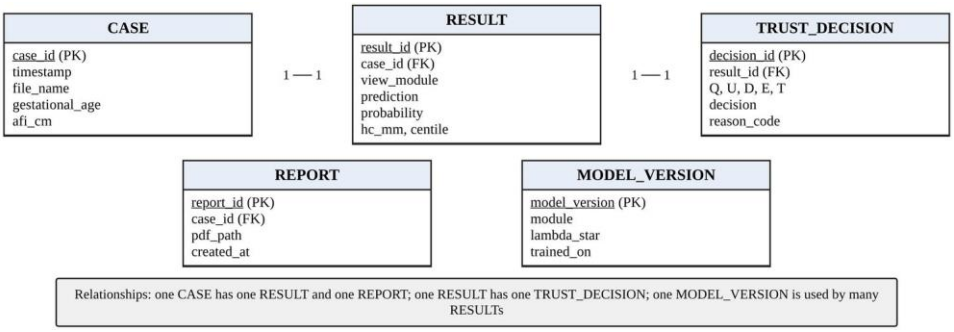


Figure 23 Planned entity–relationship diagram of the audit records

No patient name or hospital identifier is stored. The case identifier is generated by the system, and uploaded images are kept only for the session unless the user chooses to save them for review.

6. Tools, Software and Resources Required

Software Tools

Table 4.10 lists the hardware and software required and the purpose of each.

Table 24 Hardware and software requirements

Category	Tool / Resource	Purpose
Programming language	Python 3.10+	Implementation of all modules
Deep learning	PyTorch, Torchvision, timm, segmentation-models-pytorch	Classification and U-Net segmentation
Image processing	OpenCV, pydicom, NumPy, imagehash	Preprocessing, quality check, duplicate removal
Explainability	pytorch-grad-cam	Grad-CAM++ heatmaps
Statistics	scikit-learn, SciPy	Logistic regression, metrics, conformal calibration
User interface	Streamlit	Web dashboard
Reporting	ReportLab	PDF report generation
Experiment tracking	TensorBoard or CSV logs	Training curves and configuration records
Development	VS Code, Jupyter, Git and GitHub	Coding, experiments and version control
Hardware	PC with NVIDIA GPU (8 GB+), 16 GB RAM or Google Colab / Kaggle GPU	Training and testing
Datasets	FETAL_PLANES_DB, HC18, MMOTU, PCOSGen, fibroid ultrasound set	Training and evaluation

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Justification of Tool Choices

- Python and PyTorch: the most widely used tools in medical imaging research, with pretrained models and strong community support.
- timm and segmentation-models-pytorch: provide tested implementations and pretrained weights for DenseNet, EfficientNet and U-Net, saving development time.
- pytorch-grad-cam: implements Grad-CAM and Grad-CAM++ for any PyTorch CNN.
- Streamlit: builds an interactive web dashboard in pure Python, so the team does not need separate front-end development.
- Google Colab and Kaggle: provide free GPU time sufficient for transfer learning on these dataset sizes.

7. Proposed Evaluation Plan

Evaluation Metrics

The evaluation will be carried out in Phase–II; no results are reported here. With TP, TN, FP and FN the numbers of true positives, true negatives, false positives and false negatives, the main metrics are defined as follows.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \qquad (4.10)$$

$$\text{Sensitivity} = TP / (TP + FN), \qquad \text{Specificity} = TN / (TN + FP) \qquad (4.11)$$

$$\text{Precision} = TP / (TP + FP), \qquad F1 = 2 \cdot \text{Precision} \cdot \text{Sensitivity} / (\text{Precision} + \text{Sensitivity}) \qquad (4.12)$$

$$MAE_{HC} = (1/N)\, \Sigma_i \, | HC_i - HC_i^{ref} | \qquad (4.13)$$

$$\text{Coverage}(\lambda) = (1/n)\, \Sigma_i \, 1[T_i \geq \lambda], \quad \text{Risk}(\lambda) = \text{error rate among cases with } T_i \geq \lambda \qquad (4.14)$$

The area under the receiver operating characteristic curve (AUROC) summarizes classification performance across all thresholds. The expected calibration error (ECE) measures the gap between predicted probability and actual accuracy. The risk–coverage curve plots Risk(λ) against Coverage(λ) for all thresholds, and the area under it (AURC) summarizes referral quality: a lower AURC means the trust gate refers the right cases.

Planned Experiments

- Classification: AUROC, F1-score, sensitivity and specificity per class for the router, plane, ovarian, PCOS and fibroid models, compared with a plain baseline without the trust gate.
- Measurement: mean absolute error of head circumference in millimetres on the HC18 test images, and agreement of the growth flag with the reference centile.
- Localization: heatmap plausibility E against MMOTU tumour outlines, HC18 head outlines and fibroid annotations.
- Trust and referral: risk–coverage curves, AURC, error among accepted cases compared with α, and coverage; the four-signal gate is compared with a confidence-only gate.
- Robustness: train the plane classifier on one hospital of FETAL_PLANES_DB and test on the other, and report results before and after de-duplication for PCOS.

Table 25 Planned experiments

ID	Experiment	Question answered
E1	Baseline modules without trust gate	How well does each module perform on cleaned data?

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

E2	Four-signal gate vs confidence-only gate	Does the trust gate refer the right cases (lower AURC)?
E3	Ablation: remove Q, U, D or E one at a time	How much does each signal contribute?
E4	Sweep α from 0.02 to 0.20	How does coverage change with the allowed error?
E5	Unseen-hospital test (FETAL_PLANES_DB)	Does the guarantee hold under distribution shift?
E6	PCOS before vs after de-duplication	How much do duplicates inflate accuracy?

Acceptance Targets

Table 4.12 lists the acceptance targets that will be used to judge the prototype at the end of Phase–II.

Table 26 Evaluation metrics and acceptance targets

Aspect	Metric	Acceptance target (to be verified)
Classification	AUROC, F1, sensitivity, specificity	Comparable to published baselines on cleaned data
Measurement	HC mean absolute error (mm)	Comparable to the HC18 baseline
Localization	Heatmap plausibility E	Higher E for accepted than for referred cases
Safety	Error among accepted cases	At or below α (e.g. 5%) on in-distribution test data
Referral efficiency	Coverage, AURC	Better AURC than a confidence-only gate
Calibration	Expected calibration error	Reported for each module, before and after scaling
Robustness	Metrics on unseen hospital	Gap reported and explained

V. PROJECT PLANNING AND EXPECTED OUTCOMES

1. Work Breakdown and Task Allocation

Team Roles

The project is carried out by a team of four students under the guidance of the project guide. Each member has a primary area of responsibility, while all members share integration, testing and documentation. Table 5.1 lists the roles.

Table 27 Team roles and responsibilities

Member	Primary role	Main responsibilities
Aditya Prakash	Requirements and application lead	Requirements, Streamlit dashboard, PDF report, integration
Anand S	Data and fetal module lead	Datasets, de-duplication, data pipeline, view router, fetal growth module
Kiran U	Architecture and ovarian module lead	System architecture, ovarian mass module, quality check
Muhammed Swalih N	Trust and evaluation lead	PCOS and fibroid modules, Grad-CAM++, trust gate, evaluation
All members	Shared	Literature survey, testing, reports, presentations, research paper

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Task Allocation

The project is divided into the tasks in Table 5.2. Each task has a primary responsible member, while all members contribute to integration, testing and documentation.

Table 28 Work plan and task allocation

Task No.	Activity / Deliverable	Responsible Member	Planned Duration
T1	Literature survey and gap analysis	All members	Jul – Aug 2026
Task No.	Activity / Deliverable	Responsible Member	Planned Duration
T2	Requirement analysis and feasibility study	Aditya Prakash	Aug – Sep 2026
T3	Dataset study, de-duplication and data pipeline design	Anand S	Sep – Oct 2026
T4	Architecture, module and algorithm design	Kiran U	Sep – Oct 2026
T5	Trust gate and evaluation plan design	Muhammed Swalih N	Oct 2026
T6	Phase–I report and review presentations	All members	Oct – Nov 2026
T7	View router, fetal growth and ovarian modules (Phase–II)	Anand S, Kiran U	Dec 2026 – Jan 2027
T8	PCOS and fibroid modules, Grad-CAM++ and trust gate (Phase–II)	Muhammed Swalih N	Jan – Feb 2027
T9	Dashboard, PDF report and integration (Phase–II)	Aditya Prakash	Feb – Mar 2027
T10	Testing, evaluation and final report (Phase–II)	All members	Mar – Apr 2027

Work Breakdown Structure

The work breaks down into five work packages. WP1 Research covers the literature survey, gap analysis and requirements. WP2 Data covers dataset download, de-duplication, splitting and preprocessing. WP3 Models covers the router, quality check and the four modules. WP4 Trust and explanation covers Grad-CAM++, the four signals, the trust gate and calibration. WP5 Application and evaluation covers the dashboard, report, testing, evaluation, documentation and the research paper. WP1 is completed in Phase–I; WP2 to WP5 form Phase–II.

2. Project Schedule / Gantt Chart

Gantt Chart

Figure 5.1 shows the month-wise schedule. Phase–I (Semester VII) covers the survey, requirements, design, reviews and documentation; Phase–II (Semester VIII) covers implementation, testing and the final report.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Activity	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr
Topic selection and literature survey										
Problem analysis and requirements										
Dataset study and system design										
Phase I reviews, report and presentation										
Preprocessing, router and module training										
Grad-CAM++ and trust gate										
Dashboard, PDF report and integration										
Testing, evaluation and final report										

Dark: Phase I (Semester VII, 2026) Light: Phase II (Semester VIII, 2026–27)

Figure 24 Project schedule (Gantt chart) for Phase–I and Phase–II

Key Milestones

Table 5.3 lists the key milestones and their planned dates.

Table 29 Key project milestones

No.	Milestone	Phase	Planned date
M0	Phase–I report and second review completed	Phase–I	Nov 2026
M1	Clean, split datasets and working view router	Phase–II	End Dec 2026
M2	Fetal growth and ovarian modules trained	Phase–II	End Jan 2027
M3	All modules and trust gate working	Phase–II	End Feb 2027
M4	Integrated prototype fully evaluated	Phase–II	End Mar 2027
M5	Final report, demonstration and paper submitted	Phase–II	Apr 2027

3. Expected Deliverables

Phase–I Deliverables

- Literature review and gap analysis (Chapter 2).
- Requirements specification, use cases and feasibility analysis (Chapter 3).
- System architecture, data flow, workflow, module design and trust gate algorithm (Chapter 4).
- Phase–II implementation and evaluation plan (Sections 4.7 and 5.5).
- Phase–I report and review presentations.

Phase–II Deliverables

Table 30 Phase–II deliverables

Deliverable	Description
Source code	Git repository with the data pipeline, models, trust gate, dashboard and tests
Trained models	Checkpoints for the router, four modules and head outline model, with the calibrated threshold
Web prototype	Streamlit dashboard with upload, results, heatmaps, referrals and PDF report
Evaluation report	Results of experiments E1–E6 with tables and risk–coverage curves

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Final project report	Complete report following the Anna University format
Research paper	IEEE conference paper draft describing the trust gate and results
Demonstration	Live demonstration and presentation at the final review

4. Expected Outcomes and Potential Benefits

Expected Technical Outcomes

The following outcomes are projections to be verified in Phase–II, not achieved results:

- One assistant covering four common OB/GYN ultrasound checks instead of separate tools.
- Immediate rescan requests for poor images, which is especially useful for less experienced operators.
- Heatmaps and head outlines that are checked for plausibility, so doctors can see when the AI is looking at the wrong area.
- A referral mechanism with a controlled error rate and a reason for each referral.
- An honest evaluation on de-duplicated data and an unseen hospital, suitable for an IEEE publication.

Benefits to Stakeholders

The main beneficiaries are gynaecologists and sonographers in busy departments and the women they examine, assuming the system is used as decision support under medical supervision. Gynaecologists can focus their attention on the cases the system refers; sonographers receive immediate feedback that helps them improve image quality; patients may benefit from faster and more consistent reporting; and medical students can learn from examples with visual explanations.

Research Contributions

- A single OB/GYN ultrasound assistant covering fetal growth, ovarian masses, PCOS and fibroids behind a view router.
- A heatmap plausibility score that measures, rather than assumes, whether explanations point to the right structure.
- A four-signal trust gate with a conformal error guarantee and plain-language referral reasons.
- An honest evaluation protocol with de-duplication, patient-level splits and an unseen-hospital test.

5. Plan for Phase–II

Overview and Roadmap

Phase–II (Semester VIII, December 2026 to April 2027) turns the design into a working, evaluated prototype. Figure 5.2 shows the roadmap with five monthly stages and their milestones.

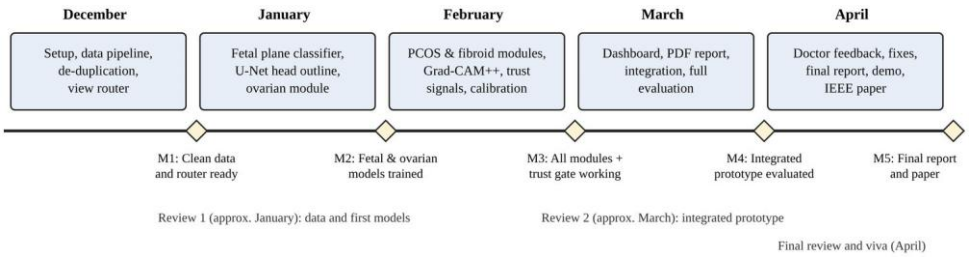


Figure 25 Phase–II implementation roadmap and milestones

Month-wise Plan

- Month 1 (December): environment setup, data pipeline with de-duplication and patient-level splits, and the view router.
- Month 2 (January): fetal plane classifier, head outline model with HC computation, and ovarian mass module.
- Month 3 (February): PCOS and fibroid modules, Grad-CAM++, the four trust signals and conformal calibration.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

- Month 4 (March): Streamlit dashboard and PDF report, integration and full evaluation including the unseen-hospital test.
- Month 5 (April): feedback from faculty and, if possible, a gynaecologist; fixes; final report, presentation and IEEE paper submission.

Week-wise Timeline

Table 5.5 gives the detailed week-wise timeline for Phase–II, with the responsible member and the expected output of each period.

Table 31 Week-wise timeline for Phase–II

Weeks	Activities	Responsible	Output
1–2	Set up Python environment, Git repository, Colab; download all datasets	Anand S	Working environment, raw data
3	Perceptual-hash de-duplication; patient-level splits; data statistics	Anand S	Clean splits, data report
4	Preprocessing pipeline (masking, sector crop, resize, filter); augmentation	Kiran U	Preprocessing code and tests
5	View router training and evaluation; quality measures	Anand S, Kiran U	Router model, Q measures
6–7	Fetal plane classifier (DenseNet121); hospital-wise split experiment	Anand S	Plane model, E5 baseline
7–8	U-Net head outline; ellipse fit; HC and INTERGROWTH centile; AFI rule	Anand S	HC module, MAE result
8–9	Ovarian mass classifier on MMOTU; benign / suspicious grouping	Kiran U	Ovarian model
10	PCOS classifier with five-fold cross-validation; before/after de-duplication	Muhammed Swalih N	PCOS model, E6 result
11	Fibroid classifier; Grad-CAM++ for all modules; heatmap plausibility E	Muhammed Swalih N	Fibroid model, heatmaps
12	MC Dropout U, energy D; logistic trust score; conformal calibration	Muhammed Swalih N	Trust gate, λ∗
13–14	Streamlit dashboard, audit log and PDF report	Aditya Prakash	Working web app
15	Integration of all modules; unit and integration testing	All members	Integrated prototype
16	Experiments E1–E6; risk–coverage curves; ablation	Muhammed Swalih N, Kiran U	Evaluation tables
17	System and acceptance testing; feedback from guide / gynaecologist	All members	Test report, feedback
18	Fixes and improvements; final model freeze	All members	Final version
19	Final report writing; IEEE paper draft	All members	Report and paper drafts
20	Final review, demonstration and viva	All members	Submission

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Technology Stack for Phase–II

Table 5.6 lists the technologies that will be used in each layer of the system and the reason for each choice.

Table 32 Technology stack for Phase–II implementation

Layer	Technology	Reason for choice
Language	Python 3.10+	Standard for AI; rich libraries
Deep learning framework	PyTorch 2.x, Torchvision	Flexible, widely used in research, GPU support
Pretrained models	timm (DenseNet121, EfficientNet-B0)	Tested implementations with ImageNet weights
Segmentation	segmentation-models-pytorch (U-Net)	U-Net with pretrained encoders
Image processing	OpenCV, scikit-image, pydicom	Preprocessing, quality measures, DICOM reading
Duplicate detection	imagehash (pHash)	Fast perceptual hashing
Explainability	pytorch-grad-cam	Grad-CAM and Grad-CAM++
Statistics and calibration	scikit-learn, NumPy, SciPy	Logistic regression, metrics, conformal threshold
Web application	Streamlit	Fast Python-only dashboard
Reporting	ReportLab	PDF report generation
Logging and storage	SQLite or CSV	Simple audit log without a server
Experiment tracking	TensorBoard, YAML configs	Reproducible experiments
Testing	pytest	Unit and integration tests
Version control	Git, GitHub	Collaboration and history
Compute	Google Colab / Kaggle GPU, lab PC	Free GPU training

Development Environment and Project Structure

The code will be kept in a single Git repository with separate folders for data, configuration files, source code (pre-processing, de-duplication, router, the four modules, explanation, trust gate and report), the Streamlit app, notebooks, tests and saved models, so that each member can work on a separate module while sharing common utilities.

Module-wise Implementation Plan

Table 5.7 summarizes the implementation plan for each module.

Table 33 Module-wise implementation plan

Module	Implementation tasks	Key technologies	Owner
Data pipeline	Download, hash, de-duplicate, split by patient, preprocess, augment	imagehash, OpenCV, pydicom	Anand S

View router	Train 4-class classifier; set confidence limit	timm EfficientNet-B0	Anand S
Quality check	Implement five measures; combine into Q; choose Q_min	OpenCV, NumPy	Kiran U
Fetal growth	Plane classifier; U-Net; ellipse fit; HC; centile; AFI rule	timm, smp, OpenCV	Anand S
Ovarian mass	Train 8-class classifier with class weights; grouping	timm EfficientNet-B0	Kiran U
PCOS	Train with 5-fold CV on de-duplicated data	timm DenseNet121	Muhammed Swalih N
Fibroid	Train classifier; use annotations for E	timm EfficientNet-B0	Muhammed Swalih N
Trust gate	MC Dropout, energy, E, logistic weights, conformal λ∗, reasons	PyTorch, scikit-learn	Muhammed Swalih N
Dashboard and report	Upload, display, overlays, log, PDF	Streamlit, ReportLab, SQLite	Aditya Prakash

Testing Strategy

Testing follows the V-model shown in Figure 5.3: each design level in Phase–I has a matching test level in Phase–II. Unit tests check individual functions such as the ellipse perimeter or the AFI rule; integration tests check that modules pass data correctly; system tests check the full application against the requirements; and acceptance testing checks that the guide and, if possible, a gynaecologist find the system useful.

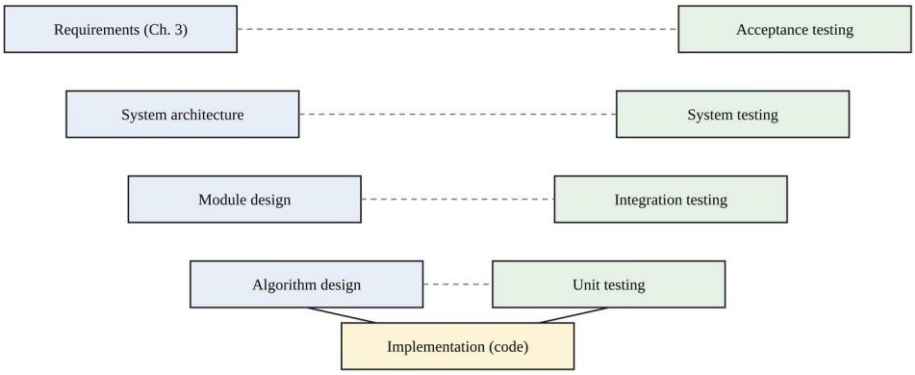


Figure 26 V-model of testing levels for Phase–II

Table 5.8 lists representative test cases that will be executed in Phase–II.

Table 34 Representative test cases for Phase–II

ID	Test case	Expected result
TC1	Upload a valid PNG fetal head image	View detected as fetal; HC computed; result shown
TC2	Upload a DICOM file	File read correctly and analyzed
TC3	Upload a non-image or corrupted file	Clear error message; application does not crash
TC4	Upload a non-ultrasound photograph	Rejected with reason R5 (unsupported view)
TC5	Upload a deliberately blurred image	Rejected with reason R1 (rescan)
TC6	Enter AFI = 4 cm	Fluid level shown as Low and flagged

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

TC7	Enter AFI = 12 cm	Fluid level shown as Normal
TC8	Enter AFI = 26 cm	Fluid level shown as High and flagged
TC9	Fetal head image without gestational age	HC shown; centile marked as not available
TC10	Ellipse perimeter unit test with known a, b	Value matches Ramanujan formula
TC11	Ovarian image from MMOTU test set	Category, group and heatmap shown
TC12	Case with trust score below λ*	Referral shown with the weakest-signal reason
TC13	Download PDF report	PDF contains image, heatmap, findings and trust details
TC14	Calibration on held-out set	Error among accepted test cases ≤ α (in-distribution)
TC15	Response time for one image on GPU	Within about 5 seconds (NFR1)

Deployment and Demonstration Plan

For the final review, the trained models and the Streamlit application will be installed on a laboratory computer, following the deployment architecture in Figure 4.4. A set of demonstration images from the test sets will be prepared to show each outcome: an accepted fetal growth result with head outline and fluid level, accepted ovarian, PCOS and fibroid results with heatmaps, a rescan request for a blurred image, an unsupported-view rejection and a referral with a reason. A README file will document installation and use so that the system can be reproduced by other students.

Risk Management during Phase–II

Table 35 Phase–II risks and contingency plans

Risk	Early warning sign	Contingency plan
Free GPU sessions are too short	Training runs interrupted	Save checkpoints every epoch; use smaller batch and mixed precision
A module performs poorly	Validation metrics far below published baselines	Try the other backbone; tune augmentation; report honestly
Conformal coverage too low	Most cases referred at α = 0.05	Report results for α = 0.10; improve signals
Integration bugs late in the schedule	Modules fail when combined	Fixed interfaces and integration tests from week 5
Dataset access problems	Download links unavailable	Keep local copies; use mirrors listed by the dataset authors
Schedule slip	Milestone missed by more than one week	Use buffer in weeks 18–19; reduce optional features

Research Publication Plan

The results of Phase–II will be written up as an IEEE conference paper. The planned outline is: introduction and motivation; related work on OB/GYN ultrasound AI and trustworthy AI; the GynoSight AI architecture and trust gate; datasets and the de-duplication and splitting protocol; experiments E1–E6; results, including risk–coverage curves and the unseen-hospital test; discussion of limitations; and conclusion. Suitable venues include IEEE conferences on biomedical imaging and engineering in medicine, as well as national IEEE conferences in India, to be selected with the guide according to deadlines.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Review Schedule

Progress will be presented to the review committee at the planned reviews of Semester VIII: a first review on the data pipeline and first models (around January), a second review on the integrated prototype (around March), and the final review with demonstration and viva (April). The exact dates will follow the department calendar.

Table 5.10 summarizes who is responsible for each major Phase–II activity and who reviews it, so that every task has a clear owner. Members are shown by short name (Aditya Prakash, Anand S, Kiran U and Muhammed Swalih N). R denotes the member responsible for doing the work, A the member accountable for its completion, and C the members consulted.

Table 36 Responsibility matrix for Phase–II activities

Activity	Aditya P.	Anand S	Kiran U	Swalih N
Data pipeline and de-duplication	C	R, A	C	C
View router	C	R, A	C	C
Image quality check	C	C	R, A	C
Fetal growth module	C	R, A	C	C
Ovarian mass module	C	C	R, A	C
PCOS and fibroid modules	C	C	C	R, A
Grad-CAM++, trust gate and calibration	C	C	C	R, A
Dashboard and PDF report	R, A	C	C	C
Testing and evaluation	R	R	R	R, A
Report, paper and presentation	R, A	R	R	R

VI. CONCLUSION AND FUTURE WORK

This chapter summarizes the work completed in Phase–I, lists the contributions of the proposed design, states how ready the project is for implementation, and describes the directions in which GynoSight AI can be extended after Phase–II.

1. Conclusion

Summary of Phase–I Work

This Phase–I report presented the design of GynoSight AI, a trustworthy explainable ultrasound assistant for the obstetrics and gynaecology department. Chapter 1 introduced the role of ultrasound in the department, the conditions addressed and the reasons why accuracy alone is not enough for safe clinical support. Chapter 2 reviewed the literature on fetal plane recognition, head circumference measurement, ovarian tumour classification, PCOS and fibroid detection, explainable AI, uncertainty estimation, out-of-distribution detection, selective prediction and conformal methods. The review showed that AI can match experts on individual OB/GYN ultrasound tasks [6] and that deferral to clinicians can improve overall reliability [8], but that existing tools handle one task each, do not check image quality, show unverified or no heatmaps, and propose referral without bounding its error.

Chapter 3 converted these gaps into a precise problem definition, identified the stakeholders and use cases, and listed thirteen functional and nine non-functional requirements together with the constraints, assumptions, risks and a feasibility analysis. Chapter 4 described the proposed approach in detail: the view router, the four clinical modules, the Grad-CAM++ explanation, the four-signal trust gate (quality Q, uncertainty U, unfamiliarity D and heatmap plausibility E), the conformal risk control threshold $\lambda*$, the reason codes, the data pipeline with de-duplication and patient-level splitting, the algorithms, the

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

tools and the evaluation plan. Chapter 5 divided the work among the team, presented the schedule and the deliverables, and laid out a week-wise plan for Phase–II with the technology stack, the testing strategy and the risk contingencies.

No implementation results are claimed in this phase. All accuracy figures quoted in this report belong to published studies, and the targets in Chapter 4 are goals to be tested in Phase–II, not achieved values.

Key Contributions of the Proposed Design

The main contributions of the GynoSight AI design are summarized below.

- One assistant for the department. A view router and four modules cover fetal growth, ovarian masses, PCOS and fibroids in a single tool, matching the way one ultrasound room handles both obstetric and gynaecological scans.
- Clinically meaningful fetal output. The fetal module goes beyond plane recognition to measure the head circumference, convert it into an INTERGROWTH-21st centile [15] and report the amniotic fluid level from the doctor-entered AFI [16].
- Image-quality check before prediction. Poor images are stopped at the start with a rescan request instead of being given a confident but unreliable answer.
- Familiarity check. An energy-based score [13] flags images that are unlike the training data, such as scans from a different machine or an unsupported view.
- Measurable explanation. Instead of judging heatmaps only by eye, the plausibility score E measures how much of the Grad-CAM++ heatmap [5] lies on the target structure and uses this value in the decision.
- Guaranteed-error referral with reasons. Conformal risk control [1] chooses the acceptance threshold so that the error rate among accepted cases stays below a chosen limit, and every referred case carries a plain-language reason code (R1–R5).
- Honest evaluation protocol. Near-duplicate removal, patient-level splits and a test on an unseen source are built into the plan to avoid the inflated results reported on duplicated datasets.

Readiness for Phase–II

At the end of Phase–I, the following are in place: the datasets have been identified and their download sources recorded; the architecture, data flow, interfaces and folder structure are fixed; the algorithms for training, calibration and inference are written in pseudo-code; the evaluation metrics and acceptance targets are defined; the week-wise schedule, task allocation and test cases are ready; and the review committee's suggestions from Phase–I will be incorporated. The project can therefore move directly into implementation at the start of Phase–II, beginning with the data pipeline and the baseline models.

Table 37 Phase–I objectives and their status

Phase–I objective	Status	Where reported
Study the clinical problem and existing AI tools	Completed	Chapters 1 and 2
Identify the research gaps	Completed	Section 2.3
Define requirements and feasibility	Completed	Chapter 3
Select public datasets for each module	Completed	Section 4.4
Design architecture, modules and trust gate	Completed	Sections 4.2–4.5
Define the evaluation plan	Completed	Section 4.7
Phase–I objective	Status	Where reported
Plan Phase–II schedule, technologies and testing	Completed	Section 5.5
Implement and evaluate the system	Planned	Phase–II

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

2. Future Work

Future work falls into two groups. The short-term work is the Phase–II implementation itself and the extensions that can reasonably be attempted within the project period. The long-term work covers directions that need more data, clinical partnerships or resources than a final-year project can provide.

Short-Term Work (Phase–II)

- Implement and evaluate all four modules, the view router and the trust gate as described in Sections 4.4, 4.5 and 4.7.
- Run the planned experiments E1–E6, including the risk–coverage analysis and the unseen-source test, and compare the results with published baselines.
- Build the Streamlit web application with the PDF report and test it with the test cases listed in Chapter 5.
- Obtain feedback from a practicing gynaecologist on the usefulness of the outputs, the heatmaps and the referral reasons.
- Prepare an IEEE conference paper describing the trust gate and the evaluation results.

Long-Term Extensions

Endometrial thickness measurement. Measuring the thickness of the uterine lining is important in abnormal bleeding, infertility and post-menopausal women. A segmentation module similar to the head circumference module can be added once an annotated public dataset becomes available.

Automatic amniotic fluid measurement. At present the AFI is entered by the doctor and checked with clinical rules. A future version can segment the fluid pockets in each quadrant and measure the deepest vertical pocket automatically, if annotated scans can be obtained from a hospital with ethics approval.
Full fetal biometry. Adding abdominal circumference, femur length and biparietal diameter would allow estimated fetal weight to be calculated and plotted on growth charts, giving a complete growth assessment.

Live video support. Extending the system from still images to live ultrasound video would allow real-time feedback to the operator, for example warning that the image is of poor quality or that the standard plane has not yet been reached.

Local hospital validation. The most important step towards clinical use is validation on images from Indian hospitals and machines, with ethics approval and gynaecologist annotation, to measure how well the trust gate holds under real conditions.

Learning from referrals and federated learning. Cases referred by the trust gate and later labelled by gynaecologists can be used to improve the models. Federated learning would allow several hospitals to train a shared model without moving patient images out of each hospital.

CPU and mobile optimization. Model compression, quantization and smaller backbones would allow the system to run on ordinary computers or tablets in low-resource clinics without a GPU.

Integration with hospital systems. Support for DICOM input and connection to hospital picture archiving systems would allow the assistant to fit into the normal reporting workflow.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 38 Future extensions and their requirements

Extension	What is needed	Expected benefit
Endometrial thickness	Annotated endometrium dataset	Supports bleeding and infertility work-up
AI amniotic fluid measurement	Annotated fluid-pocket scans from a hospital	Removes manual AFI entry
Full fetal biometry	AC, FL and BPD annotations	Estimated fetal weight and growth charts
Live video	Video datasets and faster models	Real-time guidance for the operator
Local validation	Ethics approval and hospital partner	Evidence for real-world reliability
Federated learning	Several participating hospitals	Better models without sharing images
CPU optimization	Quantization and pruning	Use in clinics without a GPU

In summary, GynoSight AI is designed as an assistant that supports, and does not replace, the gynaecologist. Its value lies in knowing when it can be trusted and saying so clearly, and the Phase–II work will test whether this design delivers that reliability on real public data.

REFERENCES

1. Angelopoulos, A.N., Bates, S., Fisch, A., Lei, L. and Schuster, T. (2024) 'Conformal risk control', Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria.
2. Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M. et al. (2017) 'SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound', IEEE Transactions on Medical Imaging, Vol.36, No.11, pp.2204–2215.
3. Bhuiyan, S.T., Rahman, R., Wasi, S., Khatun, H., Islam, A., Rahman, A.K.M.M. et al. (2026) 'PelviNeXt: A modality-agnostic hybrid network for pelvic imaging in women's health', arXiv preprint, arXiv:2608.20144.
4. Burgos-Artizzu, X.P., Coronado-Gutiérrez, D., Valenzuela-Alcaraz, B., Bonet-Carne, E., Eixarch, E., Crispi, F. et al. (2020) 'Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes', Scientific Reports, Vol.10, Article No.10200.
5. Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V.N. (2018) 'Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks', Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp.839–847.
6. Christiansen, F., Konuk, E., Ganeshan, A.R., Welch, R., Palés Huix, J., Czekierdowski, A. et al. (2025) 'International multicenter validation of AI-driven ultrasound detection of ovarian cancer', Nature Medicine, Vol.31, No.1, pp.189–196.
7. Divekar, A. and Sonawane, A. (2024) 'Leveraging AI for automatic classification of PCOS using ultrasound imaging', arXiv preprint, arXiv:2501.01984.
8. Dvijotham, K., Winkens, J., Barsbey, M., Ghaisas, S., Stanforth, R., Pawlowski, N. et al. (2023) 'Enhancing the reliability and accuracy of AI-enabled diagnosis via complementarity-driven deferral to clinicians', Nature Medicine, Vol.29, No.7, pp.1814–1820.
9. Gal, Y. and Ghahramani, Z. (2016) 'Dropout as a Bayesian approximation: Representing model uncertainty in deep learning', Proceedings of the International Conference on Machine Learning (ICML), PMLR Vol.48, pp.1050–1059.

10. Garcia-Atutxa, I., Martínez-Más, J., Bueno-Crespo, A. and Villanueva-Flores, F. (2025) 'Early-fusion hybrid CNN-transformer models for multiclass ovarian tumor ultrasound classification', Frontiers in Artificial Intelligence, Vol.8, Article No.1679310.
11. Handa, P., Saini, A., Dutta, S., Pathak, H., Choudhary, N., Goel, N. and Dhanao, J.K. (2023) 'PCOSGen-train dataset', Zenodo, doi: 10.5281/zenodo.10430727.
12. Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K.Q. (2017) 'Densely connected convolutional networks', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4700–4708.
13. Liu, W., Wang, X., Owens, J. and Li, Y. (2020) 'Energy-based out-of-distribution detection', Advances in Neural Information Processing Systems (NeurIPS), Vol.33, pp.21464–21475.
14. Munro, M.G., Critchley, H.O.D., Broder, M.S. and Fraser, I.S. (2011) 'FIGO classification system (PALM-COEIN) for causes of abnormal uterine bleeding in nongravid women of reproductive age', International Journal of Gynecology and Obstetrics, Vol.113, No.1, pp.3–13.
15. Papageorghiou, A.T., Ohuma, E.O., Altman, D.G., Todros, T., Cheikh Ismail, L., Lambert, A. et al. (2014) 'International standards for fetal growth based on serial ultrasound measurements: The Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project', The Lancet, Vol.384, No.9946, pp.869–879.
16. Phelan, J.P., Smith, C.V., Broussard, P. and Small, M. (1987) 'Amniotic fluid volume assessment with the four-quadrant technique at 36–42 weeks' gestation', Journal of Reproductive Medicine, Vol.32, No.7, pp.540–542.
17. Ronneberger, O., Fischer, P. and Brox, T. (2015) 'U-Net: Convolutional networks for biomedical image segmentation', Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI), LNCS Vol.9351, pp.234–241.
18. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group (2004) 'Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome', Fertility and Sterility, Vol.81, No.1, pp.19–25.
19. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S.Q.H., Nguyen, C.D.T. et al. (2022) 'Benchmarking saliency methods for chest X-ray interpretation', Nature Machine Intelligence, Vol.4, No.10, pp.867–878.
20. Tan, M. and Le, Q.V. (2019) 'EfficientNet: Rethinking model scaling for convolutional neural networks', Proceedings of the International Conference on Machine Learning (ICML), PMLR Vol.97, pp.6105–6114.
21. Timmerman, D., Testa, A.C., Bourne, T., Ameye, L., Jurkovic, D., Van Holsbeke, C. et al. (2008) 'Simple ultrasound-based rules for the diagnosis of ovarian cancer', Ultrasound in Obstetrics and Gynecology, Vol.31, No.6, pp.681–690.
22. van den Heuvel, T.L.A., de Bruijn, D., de Korte, C.L. and van Ginneken, B. (2018) 'Automated measurement of fetal head circumference using 2D ultrasound images', PLoS ONE, Vol.13, No.8, Article No.e0200412.
23. Xi, H. and Wang, W. (2024) 'Deep learning based uterine fibroid detection in ultrasound images', BMC Medical Imaging, Vol.24, Article No.218.
24. Yang, T. (2023) 'Uterine fibroid ultrasound images', Mendeley Data, version 2, doi: 10.17632/n2zcmcypgb.2.
25. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Cheng, G. et al. (2022) 'MMOTU: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation', arXiv preprint, arXiv:2207.06799.