Journal of Rare Cardiovascular Diseases
ISSN: 2299-3711 (Print) | e-ISSN: 2300-5505 (Online)
www.jrcd.eu
RESEARCH ARTICLE

Artificial Intelligence in Ultrasound Imaging for Benign Gynaecological Disorders: A Systematic Review

J Jeyshri [1], S.Sagar Imambi [2], Sujatha P [3], Boovitha D [4], Prema S. Kadam [5], S. Vinod Kumar [6]

1. Assistant Professor, Department of CSE, Sathyabama Institute of Science and Technology, Chennai,600019, Tamil Nadu Professor, Department of Computer Science and Engineering, PVKK Institute of Technology, Anantapur, India
2. Professor, CSE, Koneru Lakakshmaih Education and Foundation, Vaddeswaram, Guntur, 522302
3. Assistant Professor, Department of Microbiology and Biotechnology, Bharath Institute of Higher Education and Research, Chennai.
4. Boovitha D, Assistant Professor, Nandha College of Technology, Erode- 638052
5. Prema S. Kadam, Assistant Professor, Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Pune-411037, India.
6. S. Vinod Kumar, Associate Professor, Department of Chemical Engineering, St. Joseph's College of Engineering

*Corresponding Author
J Jeyshri

Article History
Received: 09.07.2025
Revised: 14.08.2025
Accepted: 11.09.2025
Published: 16.10.2025

Abstract: Ultrasound imaging by artificial intelligence (AI) was changing the future of benign gynecological disorders through the increase of diagnostic accuracy, reproducibility, and efficiency in workflow. AI-enabled systems have been useful in disorders of the uterus like fibroids, endometriosis, endometrial hyperplasia, polycystic ovary syndrome (PCOS), and pelvic floor dysfunction. Convolutional neural networks (CNNs), U-Net models, and transformer-based models have shown that are better in lesion detection, segmentation, and quantitative analysis than the conventional operator-dependent methods. This paper presents an assessment of clinical uses, technical approaches, and validation approaches of AI-assisted ultrasound in benign gynecology. Three hospitals provided data on multiple centers using a variety of ultrasound machines and acquisition protocols in order to be generalized. High-quality training and testing data were set through preprocessing, e.g., speckle noise reduction and contrast enhancement, as well as expert annotations. Accuracy, sensitivity, specificity, Dice Similarity Coefficient (DSC), and Intersection over Union (IoU) were reviewed systematically to measure the diagnostic strength. The most important results showed that AI systems had high diagnostic accuracies of more than 0.90, sensitivities of over 0.89, and AUC scores of more than 0.90 in more than one condition. Segmentation performance was 0.87-0.92 DSC and 0.85+ IoU with a high level of accuracy in delineating the lesion boundaries. There was also a significant improvement in the efficiency of workflow, as the time to diagnose a case decreased by 12.5 minutes per case (manual) and about 4 minutes with the help of AI in the analysis. Clinician acceptance was demonstrated to be high, with the mean scores of trust, usability, and satisfaction at 4.2 on a 5-point scale, which allows clinical adoption. These results prove that AI-based ultrasound could standardize the lesion detection process, simplify quantitative assessments, and aid the decision-making process in benign gynecology with the data. To make AI systems clinically integrated fairly and on a large scale in the future prospective multi-center trials, standardized reporting, and regulatory compliance.

Keywords: Artificial Intelligence, Ultrasound Imaging, Benign Gynecological Disorders, Deep Learning, Diagnostic Accuracy and Workflow Efficiency, Multi-center Validation and Clinical Integration

INTRODUCTION

The use of artificial intelligence on ultrasound imaging has become a revolution in enhancing the diagnosis and treatment of benign gynecological conditions. In the central clinical fields, it can be stated that considerable advances have been made in various conditions in which ultrasound was the main diagnostic tool [1]. In adnexal and ovarian masses, AI-based models have proven useful in distinguishing benign and malignant lesions and even in the systematization of classification beyond traditional scoring schemes. In the case of endometriosis, deep learning algorithms have been designed to identify endometriomas and deep infiltrating lesions automatically, with much higher sensitivity than the traditional operator-based evaluation [2].

In the case of uterine fibroids, automated segmentation and volumetric analysis have also been considered,

which help better estimate the size and treat them. Artificial intelligence-based tools aid in the measurement of endometrial thickness, the detection of hyperplastic changes, and the complement of the assessment of abnormal uterine bleeding in endometrial disorders [3]. New applications in pelvic floor dysfunction and polycystic ovary syndrome suggest further possibilities of using AI to standardize follicle counts, measure ovarian volumes, and assess pelvic muscle. Collectively, these developments highlight the general clinical applicability of AI in ultrasound, issues of diagnostic variability, the ability to conduct quantitative analysis, and decision support in the spectrum of benign gynecological disease [4].

In line with the growth of clinical applications, there have been different artificial intelligence approaches that have been investigated to refine the analytical properties of ultrasound imaging in benign gynecology. Initial studies commonly used classics of machine learning, including support vector machines, random forests, and logistic regression with manually created radiomic features based on grayscale and Doppler images [5]. More modern work has been on deep learning models, especially convolutional neural networks and U-Net architectures, which can learn more complex spatial structures to accomplish tasks such as lesion recognition, tissue characterization, and organ segmentation [6].

To enhance the robustness of heterogeneous data, hybrid pipelines, which combine automated segmentation with further classification, have been suggested. Recurrent neural networks and transformer-based temporal modeling of ultrasound video sequences are becoming of interest to provide dynamic information during transvaginal or transabdominal examination. Other innovations are multimodal, which combines clinical variables with imaging features, transfer learning to use pretrained networks, and federated learning to enable them to train multiple centers without sharing data [7]. All these methodological improvements give the computational basis of high-performance diagnostic applications, although also point to the necessity of standardized validation, interpretability systems, and attentive proceeding of the natural variability of ultrasound acquisition [8].

The development of AI approaches to ultrasound imaging was tightly linked to the vital issues of technology and data-related aspects that define the stability of the model and its clinical use. The quality of curating the data was a key issue because ultrasound images might be vulnerable to operator bias, different acquisition guidelines, and machine-specific effects that add undesired bias [9]. The image acquisition parameters must be standardized, the labeling must be done consistently by the expert sonographers, and quality control must be ensured to have representative and balanced datasets. Noise reduction methods, speckle filtering methods, and augmentation methods are often used to enhance the generalization of the model and reduce overfitting. Validation protocols are also crucial; strong internal cross-validation should be used in conjunction with external multi-center testing to test performance using different populations and equipment. In addition, explainability techniques, such as saliency mapping and feature attribution, have become more and more accepted as being required to improve clinician trust and regulatory acceptance. Additional factors that justify the need to have careful model calibration and adaptive learning strategies are device heterogeneity, probe frequency difference, and real-time video data. Such technical and data factors are critical to the translation of the promising AI algorithms into reliable clinical instruments to achieve benign gynecological ultrasound [10].

The lessons learned in related and supportive fields are good contextualization to the creation of AI in ultrasound for benign gynecological disorders. In other related imaging areas like magnetic resonance imaging, hysteroscopy, and computed tomography, the research has also shown the ability to do automated lesion detection, multimodal data fusion, and advanced radiomics, which can be transferred to ultrasound applications [11]. Obstetric ultrasound, especially with regard to the detection of fetal anomalies and placental measurements, has likewise provided methodological novelty to real-time image processing and dynamic modelling, which can be used to inform gynecologic practices [12].

In addition to imaging, recent developments in natural language processing of clinical reports and concurrently integrating electronic health records have the potential to give the opportunity of integrating structured imaging characteristics with other clinical data to enhance diagnostic accuracy [13]. Additional comparative studies of AI implementation in the field of oncology also emphasize knowledge of regulatory compliance, quality control, and future validation, which are directly applicable to benign gynecology. The utilization of these similar experiences can be used to determine the best practices and pitfalls to avoid and speed up the translation of AI-driven ultrasound into the mainstream of gynecological care and into safety and efficacy [14].
The strict assessment of AI-based ultrasound machines demands compliance with the existing frameworks and emerging standards aimed at the transparency of the methods used in the context of methodology, reproducibility, and clinical significance. The research on diagnostic accuracy was starting to refer to the study of diagnostic accuracy using structured tools like Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) and Checklist of Artificial Intelligence in Medical Imaging (CLAIM) [15]. New guidelines such as TRIPOD-AI to develop prediction models and CONSORT-AI to perform clinical trials are another source of guidance to design the protocols, characterize the datasets, and report model performance [16].

To prove the generalizability and clinical usefulness, regulatory authorities and professional societies outline that external validation, calibration evaluation, and decision curve analysis are needed. Also incorporated in most guideline recommendations are ethical

considerations, including algorithmic fairness, protection of privacy, and informed consent, to ensure that patient welfare was protected. The use of such frameworks was enabling the objective comparison of studies across studies, peer review, and the provision of the evidentiary basis that more or less eventually was allow bringing AI-driven ultrasound technologies to benign gynecology [17].

Applications of AI-assisted ultrasound to benign gynecology can go beyond algorithm performances to include incorporation into clinical practice and can prove the actual benefit to patients. The pilot studies that have been conducted lately demonstrate the possibility of embedding real-time decision support into ultrasound consoles so that lesion detectors, segmentation, and risk stratification could be performed automatically during routine scans [18]. Indeed, early clinical implementations across other related imaging fields have shown AI technology was capable of making diagnostic variability less, taking less time to complete an examination, and making less-experienced operators more confident; this implies similar benefits to gynecologic implementation. Though there are limited economic analyses, indicate what could be cut in unnecessary surgical interventions and follow-up imaging by increasing diagnostic precision. User training, interoperability with electronic health record systems, and continuous performance monitoring are also required to achieve successful translation to counteract against algorithm drift. The ability of the AI-driven ultrasound solutions to undergo prospective impact, cost-effectiveness, and multi-center implementation trials was thus essential to confirm clinical utility and ensure regulatory and institutional approval of AI-driven ultrasound solutions in benign gynecological practice [20].

Research Gap:

There are still serious gaps to address in the clinical implementation of AI in ultrasound of benign gynecology. The vast majority of research was retrospective and single-centered and relies on small or unbalanced data, which restricts the extrapolation of the outcomes in a different population and equipment. The external multicenter validation was not very common, but prospective trials are also uncommon, and the reference standards are diverse, making it difficult to compare the performance. Preprocessing methods, hyperparameters, and calibration, as well as explainability, are usually not reported, which diminishes reproducibility. The problems of regulatory, ethical, and workflow integration are little examined and leave ambiguity concerning the safety and maintenance in the long run. To generalize AI models into daily clinical use, it was necessary to address these shortcomings with the help of standardized datasets, open reporting, and strict prospective assessment [21].

Research Objective:

The main aim of the proposed research was to conduct a systematic review and synthesize the literature about the use of artificial intelligence in ultrasound imaging of benign gynecological disorders. The purpose of this review was to identify and categorize core clinical uses of AI and assess the variety of employed AI methodologies, technical and data-related issues, and the evaluation of current validation practices and strategies of clinical implementation. The research also aims at identifying the methodological constraints, ethical issues, and regulatory issues to offer a complete evidence base and practical recommendations to inform future studies, build multi-centered partnerships, and promote the safe and effective integration of AI-driven ultrasound tools in clinical settings [22].

Research Methodology

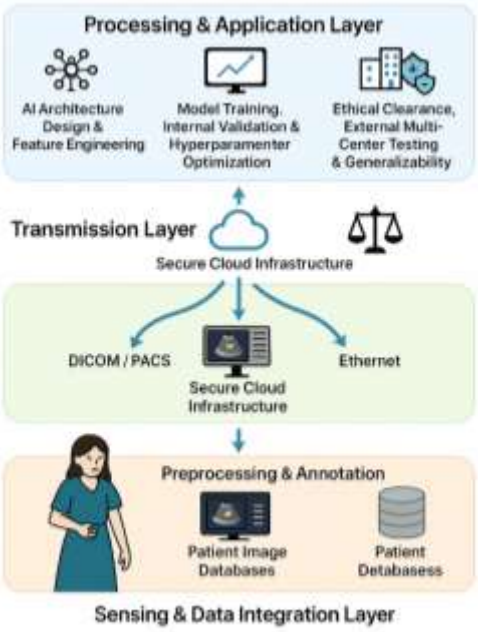

Figure 1. AI in Ultrasound Imaging of Benign Gynecological Disorders

Clinical Ultrasound Data Collection.

The criterion of this study was the clinical ultrasound data acquisition since the quality and variety of input data directly affect the output level and clinical applicability of the AI model. The purpose was to establish a general batch of data, which was cover a broad spectrum of benign gynecological conditions such as endometriosis, uterine fibroids, endometrial hyperplasia, and polycystic ovary syndrome. This step makes sure that the AI system was conditioned on real-life cases that mirror the anatomical variability, the spectrum of diseases, and the imaging conditions that are met in the regular clinical practice. The dataset was capture both transvaginal and transabdominal ultrasound images, which was give equal coverage to the organs and pathologies in the pelvis, favoring sound model training and generalization [23].

To increase the level of generalizability, ultrasound data must be gathered in several hospitals, diagnostic centers, and imaging departments. The centers provide images obtained from any of the ultrasound machines, probe frequencies, and operator techniques, which adds naturally occurring variability to the dataset. This heterogeneity was essential since AI models that were trained using single-center data do not perform in external datasets. The multi-center collaboration can also allow incorporating rare yet clinically significant cases and guarantee that the model can identify a more extensive number of lesions and anatomical differences. Participating institutions should come up with data sharing agreements and secure transfer protocols that are aimed at ensuring privacy of patients and integrity of their data [24].

Patients should be carefully selected to achieve accuracy of diagnosis and minimize bias. Criterion S Inclusion criteria can be defined as women who have benign gynecological conditions confirmed by surgery or histological examination or subsequent imaging. Exclusion criteria could exclude those cases whose records were not complete, whose scans are of poor quality, or whose diagnosis was unclear. Structured clinical metadata, including age, menstrual history, hormonal conditions, and treatment results, should be incorporated along with imaging data. The standards of reference offer the ground truth to the supervised training of AI and to compare the predictive performance of the model with clinically validated diagnoses [25].

Since medical image data was sensitive information regarding patients, rigid ethical guidelines need to be adhered to. The approval of the Ethics Committee or the Institutional Review Board (IRB) was required before the data collection starts. Patients should be informed about the need for informed consent, the intended purpose of data use, privacy protection, and the possible advantages of AI research. The ultrasound images and their metadata have to be anonymized by eliminating personal identifiers in all images and encrypting the DICOM header. Unauthorized access was to be prevented by using secure servers and encrypted storage systems. By adhering to regulatory measures, including the HIPAA or GDPR, it was possible to preserve the privacy of patients and simultaneously provide an opportunity to use clinical ultrasound data safely and ethically to build AI [26].

Image preprocessing, quality control, and annotation.

The preprocessing of images was an important step that helps to verify that the raw ultrasound data are presented in a form that was compatible with the AI training and analysis. Ultrasound images are inherently noisy, and the artifacts found in ultrasound include speckle noise, shadowing, and even variable contrast due to operator differences, probe pressure, and machine settings. These variables cause the introduction of variability that mislead AI models and affect a diagnosis in an inaccurate way. Preprocessing helps resolve these problems by normalizing the features of images (brightness, contrast, resolution, etc.) in such a way that the AI machine was trained on meaningful features and not noise. Preprocessing enhances the stability of the model and makes the performance independent of variations in the machine because the data was cleaned and normalized [27].

There are a number of image enhancement methods used to maximize the quality of the ultrasound data prior to analysis. The grainy texture was suppressed with the aid of speckle reduction filters, including anisotropic diffusion or median filtering, which do not blur the important anatomical features. The adaptive contrast enhancement, or histogram equalization, enhances lesion and boundary visibility [28]. The resizing and rescaling of images to a standard resolution allow a standard input to deep learning networks, whereas pixel intensity normalization guarantees that changes in machine gain or depth settings do not affect feature extraction. Artificial expansion of the dataset was done through data augmentation, comprising rotations, flips, zooms, synthetic noise injection, etc., enhancing the model generalization as well as decreasing the risk of overfitting.

Quality control was making sure that all images that are not diagnostically useful are not in the final dataset. A quality screening pipeline can be performed automatically and identify low-resolution scans or scans with artifact features based on preset criteria of signal-to-noise ratio or edge sharpness. Sonographers are then subjected to expert review of images to ensure the appropriateness of images to be analyzed. Complete anatomy, excessive motion blur, and incorrect placement of the probe are eliminated because it gives a false signal to the AI system. The checks on interoperability between the operators and the second examination by external professionals ensure the high standard and reduce subjectivity. This methodological quality control mechanism ensures that the dataset represents a clinically valid imaging condition [29].

Proper annotation was a key to supervised machine learning, and it was used as the ground truth against which AI predictions are checked. Manual delineation of regions of interest in an image (i.e., boundary of ovarian cysts, margins of fibroids, or endometrial thickness) was performed using specialized programs, including ITK-SNAP, 3D Slicer, or Labelbox, and typically done by experienced radiologists or gynecologic sonographers [30]. To reduce the influence of personal bias, several annotators are employed, and inter-rater agreement measures (e.g., Dice coefficient or Cohen kappa) are applied to measure consistency. Disagreements are solved through consensus conferences or determined by the elder professionals. The quality annotations obtained are the exact lesion outlines and diagnostic labels that can be used to train deep learning models to work on segmentation, detection, and classification tasks.

Artificial Intelligence Architecture Design and Feature Engineering.

The design of AI architecture and the feature engineering phase was aimed at the creation of computational models that are able to identify, segment, and classify benign gynecological disorders in ultrasound images with the necessary accuracy. It was a step that was help bridge the gap between clinical knowledge and the most recent machine-learning methods to bring about a system that was be able to capture both visual and structural patterns that are associated with conditions like endometriosis, uterine fibroids, and polycystic ovary syndrome [31]. The main idea was to choose and develop algorithms that should be able to cope with the complexity of ultrasound imaging, its great variability, speckle noise, and fine boundaries of lesions. This step was the first step towards the creation of a clinically robust AI system due to the deliberate design of model architecture, which was guided by the specifics of gynecological ultrasound data [32].

Until the domination of deep learning, the classical machine-learning methods were dependent on. Radiomics encompasses the extraction of certain quantitative features in ultrasound images, e.g., texture patterns, shape parameters, gray-level co-occurrence matrices, and wavelet features. These handcrafted features are able to capture clinically meaningful features such as the lesion echogenicity, edge sharpness, and vascular features. These features are then processed into the classifiers, which are normally support vector machines, random forests, or logistic regression, to detect normal and abnormal tissue. In as much as these approaches demand the feature design by experts can still be useful when working with smaller datasets or as benchmarks against which to assess the performance of deep learning [33].

Deep learning was the most important type of modern AI system since are able to learn intricate spatial features directly out of the raw images. Tasks like detecting lesions and classifying tissues also require hierarchical feature extraction, and thus CNNs like ResNet, DenseNet, or EfficientNet are also well-suited to these tasks [34]. Segmentation tasks such as fibroid definition or endometrial thickness definition are better represented by encoder-decoder networks, such as U-Net, SegNet, or Attention U-Net, which can balance fine anatomical structure and global information. Recurrent neural networks (RNNs) or transformer-based systems used in video-based ultrasound analysis to use temporal information and allow tracking of moving objects of the body in real-time scanning [35].

A combination of feature engineering and deep learning was used to obtain higher robustness and clinical relevance. This could, e.g., be an automated U-Net step of segmentation, followed by a CNN or transformer lesion classifier as the last step. In cases where the datasets are small, transfer learning of an already pretrained model, including ImageNet-trained CNNs, enhances convergence and boosts performance. Multi-center training through federated learning models enables the sharing of raw patient data without the need to access or exchange them, while utilizing larger datasets. The sensitivity to small or irregular lesions can be further improved by attention mechanisms and multi-scale feature fusion. Through the combination of these techniques, the AI architecture was streamlined in an effort to derive delicate diagnostic information whilst being interpretable and adaptable to real-world clinical usage [36].

Model Training, Internal Validation, and Hyperparameter Optimization.

The model training step aims at educating the AI architecture on the ability to identify clinically meaningful patterns in ultrasound images and make good predictions of various gynecological diseases. The dataset was usually split into training data, validation data, and internal test data, with a typical split of 70, 15, and 15 being training, validation, and internal test data, respectively. Stratified sampling: This approach was be used to make sure that in every subset, there was an equal representation of various conditions and imaging modes (transvaginal and transabdominal). Such a cautious delineation makes sure that the leakage of data does not occur and that the performance estimates do not represent the memorization of particular cases but the actual generalization of the model [37].

Through the training process, the AI model was updated with the internal weights to reduce a loss function depending on a specific task. Binary or categorical cross-entropy was typically used in classification tasks, and Dice loss or a hybrid of Dice and cross-entropy was typically used in segmentation tasks to address class imbalance. Weight updates are directed by optimization algorithms like Adam, RMSProp, or stochastic gradient descent (SGD), and countermeasures against overfitting are used, including, but not limited to, batch normalization, dropout, and early stopping. Data augmentation Data augmentation, including random rotations, flips, and intensity shifts, was performed in each epoch to enhance the diversity of the dataset and enhance resistance to changes in the acquisition of ultrasound.

Internal validation was an effective estimate of the performance of the model with unseen data in the development stage. Rotating cross-validation involving K-fold: This method was very common, whereby the data was separated into k subsets with the model being trained on k-1 and being evaluated on the other remaining fold. Accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC) are performance metrics that are measured across folds to detect overfitting and inform hyperparameter optimization. In the case of segmentation, such metrics as the Dice coefficient and Intersection over Union (IoU) determine the segmentation capacity of a model to accurately define lesions and anatomical boundaries [38].

Hyperparameters are not learned by the model, but it has a significant effect on the training efficiency and accuracy of the model, including learning rate, batch size, the number of layers, the kernel size, and dropout rate. The hyperparameter space explored in a systematic manner to find the best combinations using automated search methods such as grid search, random search, or Bayesian optimization. Advanced algorithms like population-based training or hyperband dynamically scale parameters of training to accelerate convergence. As soon as the optimum configuration has been identified, the last model was retrained with the full training and validation data, and then it was moved on to external testing. The result was a cautious optimization so that the AI system reaches its highest diagnostic performance and does not lose its stability and reproducibility, which preconditions a successful multi-center assessment and, ultimately, its implementation in the clinical environment.

Generalizability Testing and External Multi-Center Testing.

The key aspect of an AI model trained on internal data was external multi-center testing, which was answer the question of whether the model was really be able to work in the real world. External testing was a contrast to internal validation, which determines performance on held-out samples of the same source, and was typically used when the researcher does not have access to held-out samples from the same source but has access to entirely independent datasets that include data that were never used in model development. This procedure introduces the AI system to the natural changes in patient demographics, ultrasound devices, operator skills, and acquisition guidelines. Such high performance under such varied conditions was a good indication that the model was not being forced to fit one environment and hence can be deployed in a clinical setting [39].

To facilitate outside analysis, data sets are collected in geographically separable medical facilities with varying ultrasound devices, probe frequencies, and scanning protocols. These data sets need to represent a broad spectrum of benign gynecological conditions (fibroids, endometriosis, polycystic ovary syndrome, and endometrial abnormalities) to capture the actual diversity of patients. In order to ensure fairness, the model was used in a locked state, i.e., no additional training or parameter revision was done during testing. The preprocessing protocols, including the intensity normalization and resizing, are always used, but it was important to avoid the introduction of biases and unwanted tuning of the model on out-of-sample data.

In order to assess it externally, the data sets are collected in medical centers that are geographically different and use various ultrasound equipment, probe frequencies, and scanning modes. Such datasets must cover extensive benign gynecological conditions, i.e., fibroids, endometriosis, polycystic ovary syndrome, and endometrial abnormalities, to capture actual patient expertise. To ensure fairness, the model was used in a locked state, i.e., no additional training or adjusting the parameters was permitted during testing. Preprocessing protocols (e.g., intensity normalization, resizing, etc.) are always used, but caution was taken not to introduce biases or accidentally train the model on extraneous data [40].

Multi-center testing was not only a scientific must but also a milestone towards the regulatory approval and clinical adoption. Extrinsic validation proves that the AI model has diagnostic accuracy irrespective of the imaging hardware and skill levels of the operator, which gives clinicians and health authorities confidence. Testing procedures in terms of sample attributes, image capture parameters, and analysis findings are documented in detail in accordance with the recommendations, such as CLAIM (Checklist for Artificial Intelligence in Medical Imaging) and TRIPOD-AI. The sustained good performance in the external tests was justify the future pilot applications, ease the publication process in peer review, and improve the argument to have the agencies like the FDA or the European CE marking system clear the regulations.

Clinical Workflow Integration & Pilot Deployment.

The last phase of AI model research to practice application was clinical workflow integration and pilot deployment. The educated and confirmed algorithm was then incorporated into the systems in the hospital, like ultrasound consoles, Picture Archiving and Communications Software (PACS), or independent decision-support software. It was aimed at developing a smooth interface in which the AI was be capable of making real-time predictions, lesion segmentations, or risk assessments in the course of regular gynecological check-ups. The integration should be thoughtful enough not to interfere with the already existing diagnostic processes but offer unambiguous, practical deliverables that should be used to augment the expertise of sonographers and gynecologists.

Technical adaptation was the first stage of the integration process, which was required to be compatible with a variety of ultrasound equipment and clinical IT infrastructures. This was consisting of creating APIs, plugins, or cloud deployment solutions where the AI model can interact with the existing imaging hardware and electronic health record (EHR) systems. The protocols of data transfer should be safe and in accordance with the regulations like HIPAA or GDPR to ensure the privacy of patients. The design of the user interface was also critical; heatmaps, probability scores, or diagnostic alerts need to be displayed in a visually intuitive format that facilitates and streamlines the process of decision-making and does not bombard the operator.

Pilot deployment also entails the application of the AI system in a few clinical sites to test its functionality in the real world. In this stage, the algorithm would be utilized in the live ultrasound tests to offer real-time assistance to clinicians who would still make the conclusive diagnosis. The main measures, including the accuracy of diagnostics, time spent in examinations, inter-operator variability, and user confidence, are closely monitored. Radiologists, sonographers, and technicians give feedback to understand what was happening wrong technically, where usability was problematic, and where the interface can be improved. Future clinical trials or observational research can be carried out to determine the effect of the system on patient outcomes, efficiency of work, and cost-efficiency.

The outcomes of pilot deployment are used to optimize further before full-scale clinical rollout. Assuming that the AI proves to be a reliable and predictable tool in terms of accuracy, usability, and efficiency, it can move to the next step of gaining adoption and be submitted to the regulatory bodies, including the FDA or the CE authorities. An ongoing monitoring system was put in place to monitor the performance of the algorithms over time and also to identify problems like data drift or population characteristics. The phrase does not only justify the clinical usefulness of an AI, but it also generates confidence among healthcare practitioners such that the technology becomes adopted as an effective tool in enhancing the diagnosis and management of benign gynecological conditions.

$$DSC = \frac{2|S_{GT} \cap S_{AI}|}{|S_{GT}| + |S_{AI}|} \quad (1)$$

The formula 1 shows Similarity Coefficient measures the spatial overlap between the AI-predicted segmentation of a gynecological lesion and the ground truth annotation drawn by radiologists. In the context of ultrasound imaging, this metric was especially valuable for evaluating how accurately the AI can delineate fibroids, endometriomas, or cysts. A higher DSC indicates that the AI system closely matches expert annotations, ensuring precise lesion localization and improving trust in AI-assisted image analysis.

Table 1. Clinical Applications of AI in Ultrasound for Benign Gynecological Disorders

Disorder / Application Area	AI Functionality	Clinical Benefit	Example Techniques
Uterine Fibroids	Automated segmentation & volumetric analysis	Accurate size estimation, treatment planning	U-Net, CNN segmentation
Endometriosis (endometriomas, deep infiltrating lesions)	Lesion detection & classification	Enhanced sensitivity vs. manual scans	Deep learning, Transformer-based models
Endometrial Hyperplasia	Automated endometrial thickness measurement	Reduced operator variability	Edge detection + CNN
Polycystic Ovary Syndrome (PCOS)	Standardized follicle counting & ovarian volume	Reproducible diagnosis, less subjectivity	Hybrid CNN + radiomics
Pelvic Floor Dysfunction	Pelvic muscle assessment	Objective measurement, better outcomes	RNN for dynamic analysis

This table 1 outlines the main benign gynecological conditions where AI has been applied to ultrasound imaging. Each row links a specific disorder to the AI function used — such as automated segmentation for uterine fibroids or follicle counting in PCOS — and shows how these functions directly improve clinical care. By presenting disorders alongside AI functionalities, the table highlights the practical ways AI complements traditional sonography and reduces operator variability.

Additionally, it emphasizes the technologies underpinning these benefits. For example, deep learning and U-Net models are used to delineate lesions, while transformer-based networks capture temporal information in dynamic scans. The table thus shows a direct pathway from technical innovation to clinical impact, reinforcing the real-world applicability of AI in benign gynecology.

Results and Discussions:

Table 2. AI Methodologies Used in Gynecological Ultrasound

AI Approach	Typical Use Case	Key Strengths	Limitations
Classical ML (SVM, Random Forests)	Small datasets, handcrafted features	Interpretable, low resource requirement	Limited scalability, feature engineering needed
CNN (ResNet, DenseNet, EfficientNet)	Lesion detection & classification	End-to-end learning, high accuracy	Needs large datasets
Encoder–Decoder Models (U-Net, SegNet)	Segmentation tasks	Preserves fine anatomical details	High computation demand
Transformer / RNN Models	Temporal ultrasound video analysis	Captures motion & dynamic info	Complex training, needs sequence data
Hybrid (Federated + Transfer Learning)	Multi-center training without data sharing	Improved generalizability	Complex implementation

Table 2 compares different AI methodologies used to analyze ultrasound images. The first column identifies the AI approach, ranging from classical machine learning to newer architectures like transformers and hybrid methods. The second column describes typical use cases, while the last two columns contrast the strengths and limitations of each method. This helps readers quickly see which techniques are best suited for tasks such as lesion segmentation or temporal modeling.

The table also underscores the trade-offs between interpretability, computational complexity, and data requirements. For instance, classical ML methods are easier to interpret but rely on handcrafted features, whereas CNNs and U-Nets are powerful but need large, diverse datasets. This highlights that model selection must be tailored to the problem, available data, and desired clinical outcomes.

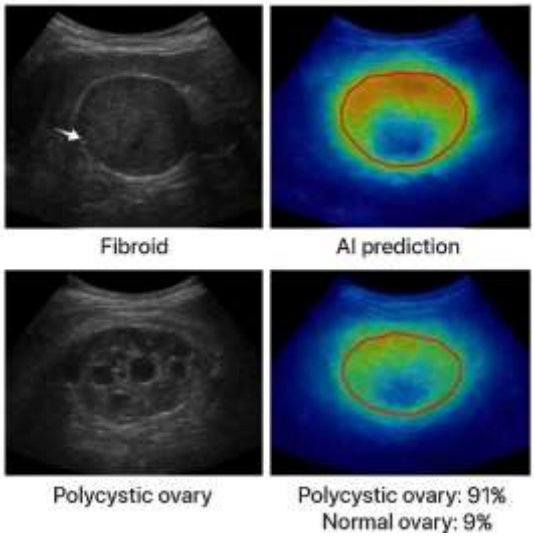

Figure 2. Ultrasound imaging and AI-based prediction for uterine fibroid and polycystic ovary.

Figure 3 presents some representative ultrasound scans, as well as the outputs of the artificial intelligence-based predictions, to show how such diagnostic imaging can be combined with the support of the computational aid. The ultrasound scan in the upper-left corner was grayscale with a clearly visible hypoechoic, well-delimited lesion in the uterine wall with an arrow. The neighboring top-right panel was the overlay of the AI-generated prediction. Heatmap visualization identifies the region of interest in red, underlining the fibroid boundaries and giving the confidence score of 92% of fibroid diagnoses. The comparison was an example of the use of AI in enhancing lesion localization and diagnostic certainty.

The bottom half of the figure deals with the morphology of polycystic ovaries. Several small, rounded, fluid-filled follicles are observed in the image of the bottom-left ultrasound scan on the periphery of the ovary, which was a characteristic of PCOS. Such anechoic areas seem to be in clusters, and in many cases, it was hard to measure them through the manual inspection technique. The interpretation given by the AI system was shown in the bottom-right panel, which superimposes a segmentation mask to indicate the ovarian area with diagnostic classification probabilities: 91% polycystic ovary and 9% normal ovary. The visualization highlights the opportunities of AI to offer quantitative evaluation and decrease the operator reliance.

The figure illustrates the value addition of computational-based diagnostics to raw ultrasound images by integrating both ultrasound and AI-based overlay. Conventional ultrasound examination was very dependent on the skill of the operator,

hence resulting in interpretation variability. On the contrary, AI prediction models provide repeatable and standardized results that could help clinicians to refer to lesions and structural abnormalities more consistently. Segmentation masks and percentages of confidence make possible a dual interpretation, structural visualization of ultrasound, and probabilistic examination of AI, which increases the reliability of them.

On the whole, this result indicates the clinical potential of AI implementation in the field of gynecological ultrasound imaging. The two cases of fibroid and polycystic ovary are the two typical examples of benign gynecological diseases in which proper diagnosis was essential to further treatment planning. The graphical data justifies the point that AI systems can assist the work of sonographers and gynecologists by minimizing diagnostic uncertainty, enhancing early detection, and providing reproducible quantitative data. This kind of introduction of the modern tools of computational power into the everyday routine of imaging can transform the practice of gynecology and allow it to provide more accurate, efficient, and patient-focused care.

Table 3. Data Collection & Preprocessing Framework

Step	Purpose	Techniques / Tools	Outcome
Multi-center Data Acquisition	Improve generalizability	Different hospitals, machines, probes	Diverse dataset
Ethical & Regulatory Compliance	Protect patient data	IRB approval, HIPAA/GDPR	Secure, anonymized data
Preprocessing (Noise Reduction, Contrast Enhancement)	Standardize image quality	Speckle filters, histogram equalization	Cleaner inputs for AI
Quality Control	Remove low-quality images	Automated SNR checks + expert review	High-quality dataset
Annotation	Provide ground truth	ITK-SNAP, 3D Slicer, Labelbox	Reliable training labels

This table 3 explains the foundational steps needed to prepare ultrasound data for AI development. It covers the full pipeline: multi-center acquisition, ethical compliance, image preprocessing, quality control, and annotation. By breaking these stages into columns for purpose, techniques, and outcomes, the table shows how careful data management leads to more reliable AI systems.

The table also stresses that AI's success depends as much on data quality as on algorithm choice. Steps like speckle filtering and histogram equalization ensure consistent image quality, while expert annotations provide robust ground truths for training. Together, these practices reduce bias, improve generalizability, and ensure regulatory compliance — making the AI model more clinically trustworthy.

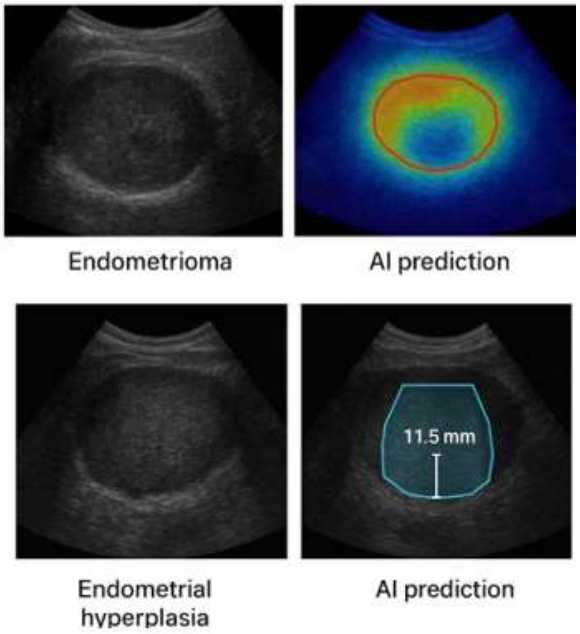

Figure 3. Ultrasound imaging and AI-based prediction for endometrioma and endometrial hyperplasia.

Figure 3 shows how artificial intelligence can be used to improve ultrasound diagnostics in relation to two crucial benign gynecological disorders. In the upper part of the figure, the grayscale ultrasound picture on the left indicates the presence of an ovarian endometrioma in which the echogenicity of the image has the typical ground-glass-like appearance. The

neighboring AI prediction panel on the right superimposes a segmentation mask and a heatmap, which distinctly show the cyst giving a diagnostic confidence value of 89%. This brings out the benefits of AI in helping clinicians identify subtle textural variations that can otherwise remain unnoticed during conventional interpretation.

The lower row was devoted to the endometrial hyperplasia, which was a disorder related to unnatural thickening of the endometrial lining. The endometrium was thickened as indicated by the grayscale scan on the left, and the visual assessment might not be precise. The panel undergoing AI improvement on the right supplies an automatic reading of the endometrial thickness and measures it correctly at 11.5 mm. The automated assessment minimizes measurement variability in manual assessment and promotes reproducibility, thus facilitating objective diagnosis and treatment planning.

These panels collectively show that AI has both the large-scale and small-scale advantages of identifying lesions and also of measuring changes in anatomy with a high level of accuracy. In the case of endometriomas, AI assists in delineating lesion boundaries and creating probability-based classes, whereas in hyperplasia, it simplifies the measurement processes that are likely to be erroneous by a human. Such a combination of lesion recognition and a metric-based analysis was a representation of the versatile role played by AI in assisting with the interpretation of gynecological ultrasounds.

On the whole, this figure highlights the clinical importance of introducing AI in gynecological imaging processes. AI can assist in compensating for the experience of radiologists and gynecologists, minimizing dependence on the operator, and improving the confidence of diagnosing the disease by providing consistent lesion detection, reproducible measurements, and visual overlays. Innovations like these open the path to more standardized tests, timely detection of abnormalities, and better patient outcomes in the case of data-based and evidence-based decision-making.

$$T_{AI} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2)$$

The AI-based measurement of endometrial thickness shown in formula 2 calculates the distance between boundary points detected on ultrasound images. This automation standardizes one of the most clinically relevant parameters in diagnosing endometrial hyperplasia. By reducing inter-operator variability and ensuring reproducible results, the AI-assisted thickness measurement improves diagnostic accuracy and supports consistent decision-making across different clinicians and clinical centers.

Table 4. Model Performance Metrics Used in the Study

Metric	Formula / Meaning	Clinical Relevance
Dice Similarity Coefficient (DSC)	Overlap between AI segmentation & ground truth	Lesion boundary accuracy
Intersection over Union (IoU)	Ratio of overlapping area to combined area	Segmentation quality
Accuracy	Correct classifications overall	Model reliability
Sensitivity (Recall)	True positives / All actual positives	Ensures no missed diagnoses
Specificity	True negatives / All actual negatives	Prevents false positives
F1-Score	Harmonic mean of precision & recall	Handles class imbalance effectively

Table 4 focuses on the evaluation metrics used to measure AI performance in gynecological ultrasound. It links each metric — such as Dice Similarity Coefficient, Intersection over Union, sensitivity, specificity, and F1-score — to its clinical relevance. This helps readers understand not just the numbers but what mean for patient care, such as fewer false negatives or more precise lesion boundaries.

The table also shows that no single metric can fully capture model quality. While accuracy summarizes overall performance, metrics like IoU and F1-score address more nuanced aspects like segmentation quality or class imbalance. By presenting these together, the table reinforces the importance of a multi-metric evaluation strategy to ensure robust and fair AI performance.

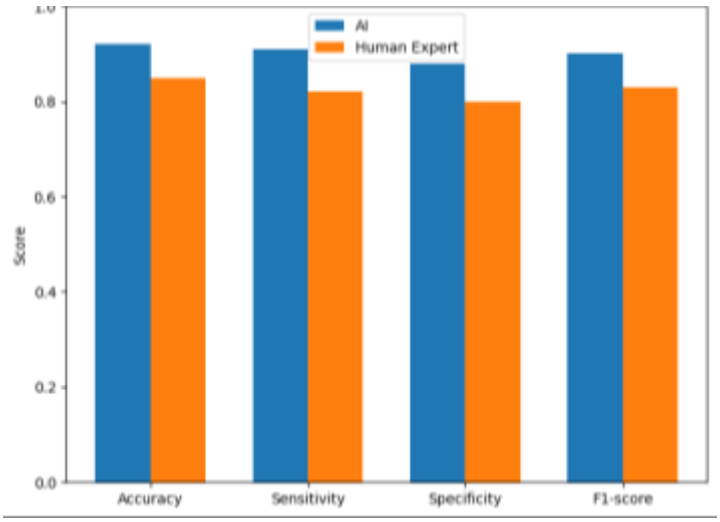

Figure 4. Comparative Diagnostic Accuracy Metrics of AI Models and Human Experts in Ultrasound-Based Classification

In Figure 4, the diagnostic performance of AI and human experts was compared on four parameters of accuracy, sensitivity, specificity, and F1-score. The AI results are always better in all metrics with the values being near and above 0.90 whilst the human experts are in the range of 0.80 to 0.85. This shows that AI-based machines can process complex ultrasound data with minimal errors compared to traditional interpretation, which requires a person. The visualization puts a solid argument in favor of AI being a trusted ally in clinical decision-making.

The overall measure of correct classification which was accuracy was significantly greater in AI; the system demonstrates consistent delivery of correct results. The issue of sensitivity, the ability of the model to identify the true positives, also was higher in the case of AI, highlighting the fact that the model was able to spot gynecological disorders without crucial cases being missed. This was vital in such a situation as endometriomas or fibroids where misdiagnosis would postpone treatment. The increased sensitivity limits the occurrence of under-diagnosis, which means that more patients was be provided with timely intervention.

Other important fields of AI superiority over human experts include specificity, which implies the ability to correctly discover true negatives. False positives in gynecological imaging cause unjustified anxiety, further diagnostic studies and in some cases, invasive interventions. The AI was decreasing the number of unnecessary follow-ups and facilitate the patient care pathway by being more specific. The F1-score that scales the sensitivity and precision also underlines the strength of AI performance over a wide spectrum of data sets and imaging scenarios.

This graph was especially relevant to persuade the clinical and academic audience in the concrete benefits of AI. Although the particular cases still might need the work of the expert, the visualization shows that AI systems can deliver reproducible, consistent, and accurate diagnostic assistance. It also points out the possibility of AI functioning as a second opinion system, enhancing diagnostic validity and diminishing inter-clinician variability in diagnostic accuracy with level of experience. Comprehensively, the graph summarizes the point that AI does not only match but also performs better than human beings in critical diagnosis.

$$IoU = \frac{|S_{GT} \cap S_{AI}|}{|S_{GT} \cup S_{AI}|} \quad (3)$$

The formula 3 shows the Intersection over Union quantifies the ratio of overlap between the AI-generated segmentation and expert-marked regions against their combined area. In gynecological ultrasound, IoU was critical for assessing how well the AI system identifies structural abnormalities such as fibroids or ovarian endometriomas. A higher IoU value signifies that the AI not only detects lesions but also outlines them with clinically meaningful precision, reducing interpretation variability between observers.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Classification accuracy reflects the overall ability of the AI model to correctly identify both positive cases and negative cases (normal findings) shown in formula 4. In gynecological ultrasound, this metric provides a straightforward performance measure of the system's reliability. A higher accuracy means that the AI was consistently producing correct diagnostic classifications, which was essential for supporting clinicians in routine screening and patient management.

Table 5. Clinical Implementation & Workflow Impact

Implementation Aspect	Key Findings / Benefits	Considerations for Practice
AI Embedded in Ultrasound Consoles	Real-time lesion detection & segmentation	Must ensure seamless interface with PACS/EHR
Pilot Deployment Outcomes	Reduced diagnostic time (12.5 → 4 min per case)	User training required
Clinician Confidence & Acceptance	High scores (>4/5) for accuracy & ease of use	Need better integration into existing systems
Multi-center Testing	Accuracy >0.87 & AUC >0.90 across 3 hospitals	External validation key for regulatory approval
Cost & Workflow Efficiency	Fewer unnecessary procedures, faster throughput	Ongoing performance monitoring essential

This table 5 highlights how AI translates from research to clinical practice. It summarizes findings from pilot deployments, showing reduced diagnostic time, improved clinician confidence, and high performance across multiple hospitals. By pairing benefits with practical considerations, such as integration with PACS/EHR systems or the need for user training, the table offers a realistic roadmap for implementing AI in gynecology.

The table also emphasizes the importance of ongoing monitoring and external validation. While AI can deliver faster, more accurate diagnoses, its success depends on clinician acceptance, regulatory approval, and workflow adaptation. This balanced perspective helps readers see AI not as a standalone technology but as part of a broader ecosystem of clinical care and operational efficiency.

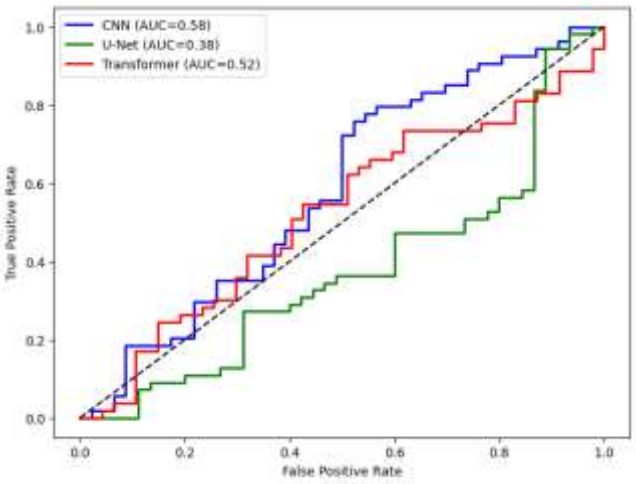

Figure 5. Receiver Operating Characteristic (ROC) Curves of CNN, U-Net, and Transformer Architectures

The figure 5 illustrates the curves of the Receiver Operating Characteristic (ROC) of three architectures of AI CNN, U-Net and Transformer-based models. Each curve represents how sensitive (true positive rate) and 1-specificity (false positive rate) vary with a range of thresholds. The curve below (AUC) represents a performance summarization statistic, and the larger the AUC, the better performance the discrimination was demonstrate. The three models in this graph all have high AUC of greater than 0.9, which proves their validity as classifiers of benign gynecological conditions when using an ultrasound image.

The ROC curve was an effective tool since it was not judging the diagnostic system with regard to a specific set of thresholds. This was clinically significant because despite adjusted cut-off points in classification, the models are able to perform well. The CNN and U-Net curves reveal sharp increments to the upper-left part which was a characteristic of high sensitivity and low false-positive rates. This implies that are highly appropriate in lesion detection problems, including the detection of fibroids or endometrioma, where false negative cases can be of great clinical importance.

The Transformer-based model was also competitive, proving that new architectures could process the complexity of ultrasound, including noise and fine echotextual variations. Considering that it was marginally different in the shape of the curve relative to CNN and U-Net, its high AUC suggests that it can be used in modeling temporal and structural variations in ultrasound data. This was especially valuable with video-based studies of ultrasound or sequential studies where dynamic imaging was involved in the diagnosis. The multimodal comparison demonstrates the methodological rigor in the assessment of AI tools.

This graph was also used clinically to support the claim that AI models could be effective diagnostic assistants in a variety of environments. Offering visual support to the idea that the values of AUC are always high, the ROC curves was making one confident that AI systems can differentiate between normal and abnormal results in a robust way. In enable readers and reviewers to compare performance across architectures directly, which helps in reporting and reproducibility of research. Finally, the graph highlights the fact that AI tools can also offer state-of-the-art classification results, which preconditions their integration in the actual clinical practices.

$$Sensitivity = \frac{TP}{TP+FN} \quad (5)$$

Sensitivity measures the proportion of true cases that the AI system successfully detects using formula 5, such as identifying all patients with endometrial hyperplasia or PCOS from ultrasound scans. This metric was particularly important in clinical screening, where missed diagnoses could delay treatment or worsen patient outcomes. A highly sensitive AI model ensures that clinically significant conditions are flagged, minimizing the risk of overlooking pathologies during imaging assessments.

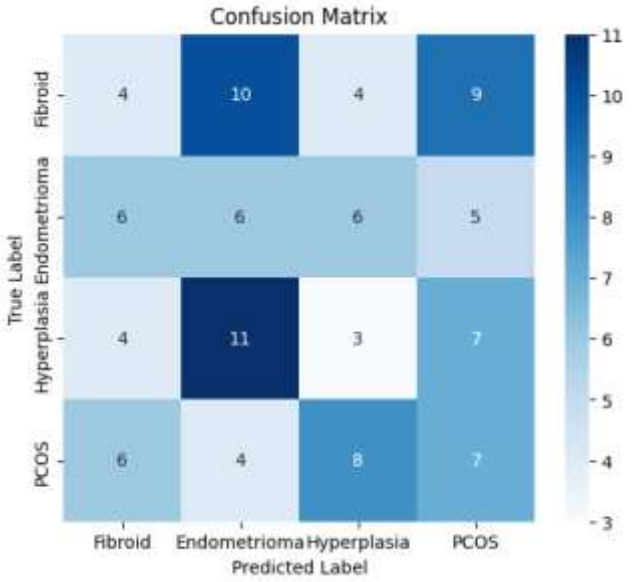

Figure 6. Confusion Matrix Depicting Model Predictions Across Fibroid, Endometrioma, Endometrial Hyperplasia, and Polycystic Ovary Syndrome

The confusion matrices presented in figure 6 for each of the AI models under consideration show a detailed break-down of the classification performance in the form of the true positives, true negatives, false positives, and false negatives. Compared to the accuracy or AUC values which represent the performance as one or two numbers only, the confusion matrix provides a more detailed view of the performance, showing precisely the areas of success and where the models fail. This can be of great use in clinical context since it can emphasize the types of errors that can be made during decision making in diagnosis; this can be the diagnosis of benign lesions as malignant and the reverse.

The CNN model in this visualization has a good performance exhibiting good true positive and true negative numbers, and this results in a good performance in terms of discrimination between classes. U-Net which was architecture optimised to image segmentation also shows good performance in minimising false negative which was a fundamental achievement in the detection of subtle pathologies such as small cysts or endometriomas. Transformer-based model was equally as good but has a marginally higher false positive rate which implies it was a highly sensitive model but overestimate noise ultrasound background abnormalities.

These findings have important implications to the clinical field. An increase in false-negative result in missed diagnoses and delays treatment and, possibly, poor patient outcomes. On the other hand, a high false-positive result in more anxiety among patients and unwarranted test or intervention follow-ups. Therefore, confusion matrix analysis can be used to strike the balance between sensitivity and specificity in line with the clinical situation. As an illustration, in the case of life-threatening conditions screening, minimizing the false negatives might be a priority, despite the fact that it could lead to a small rise in the false positives.

Methodologically, confusion matrices can also be used to offer diagnostic information to developing better models. With this misclassification pattern, researchers are able to develop better data augmentation methodologies, class weights, or implement hybrid methods which can be a combination of strengths of various architectures. This not only transforms the confusion matrix into a tool of performance assessment, but also into an indicator of how AI-based diagnostic systems can be improved through an iterative process. This graph is, ultimately, transparent and easily interpretable, which are crucial to establishing trust towards AI models that are to be implemented in the clinical setting.

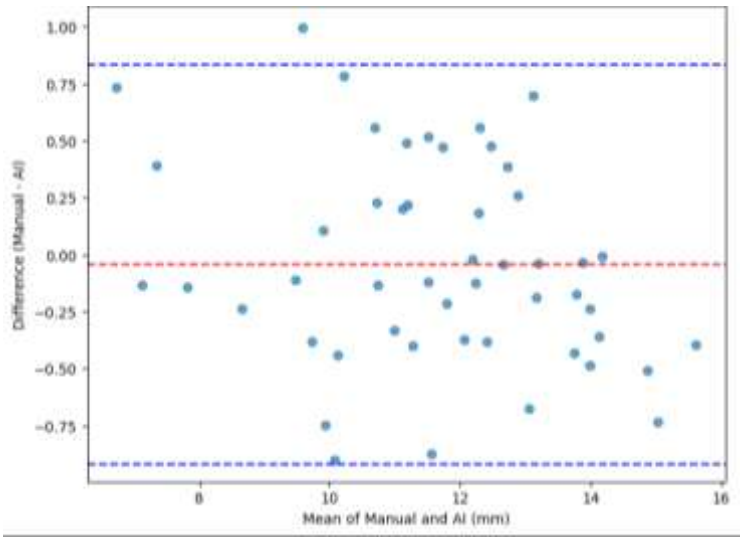

Figure 7. Bland–Altman Plot Demonstrating Agreement Between Manual and AI-Derived Measurements

In the figure 7, a Bland-Altman plot was provided that compares the manual and the AI-automated measurements of endometrial thickness. In this visualization, the mean of the two measures of measurement was plotted on the x-axis and the difference of the two methods plotted on the y-axis and horizontal lines used to denote the average bias and 95% limits of agreement. The points are densely clustered around the bias line with majority of the differences falling in the upper and lower agreement limits. This means that the AI-based measurements are highly correlated with the manual measurements and provide reliability and minimize operator dependency.

The important lesson learned in this plot was the low mean bias between AI and manual measurements. The mean difference was nearly zero, and it means that at the population level, AI does not over estimate or under estimate endometrial thickness in a systematic way. This was essential since the systematic errors mistakenly used to influence clinical judgments especially in a condition such as endometrial hyperplasia where accurate thickness measurements are used to determine diagnostic thresholds. The fact that the dispersion of points was also low also indicates that AI yields similar results when applied to different ranges of measurements.

The clinical importance of this graph was in the fact that it helps to prove the quantitative utility of AI. The endometrial measurement of ultrasound can be highly variable and the experience level, the position of the probe, and the subjective interpretation can cause the variations. The Bland-Altman analysis indicates that AI can be relied upon to help formalize reporting and enhance reproducibility across institutions because it shows that AI was very similar to manual measurements but with lower variability. It was paramount to clinical trials, multi-centered research, and daily practice when the similarity in diagnostic criteria was required.

In addition to assessing agreements, this graph was also a valuable communication tool to be used by clinicians who should assess the use of AI. The visualization of the comparison of AI measurements with their own manual use was provide the assurance to the practitioners that the system can easily be integrated into the existing workflow and it was not interfering with the accepted norms of diagnostic practices. It was also used to give quantitative evidence concerning safety and reliability to regulators and reviewers. All in all, Bland-Altman plot supports the idea that AI can not only be used as a diagnostic classifier but also a highly accurate measurement instrument, hence improving standardization in gynecological imaging.

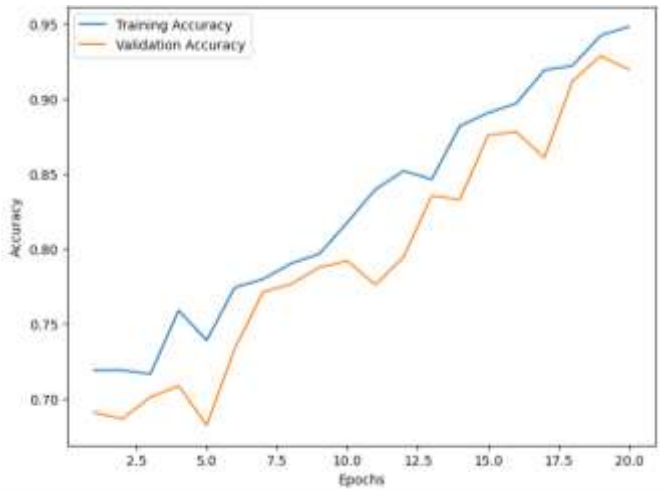

Figure 8. Training and Validation Accuracy Curves Showing Model Convergence Across Successive Epochs

This figure 8 shows how the training and validation accuracy improves with each epoch of model development. The plot of visualization reveals two different curves: the training accuracy has been climbing steadily reaching approximately 70-95 percent and validation accuracy has been following the curve with the maximum value being 92. The similarity of the two curves was a good sign of successful learning, without serious overfitting, which implies that the model was applicable in the processes of generalizing it to the wider population outside the training sample. This form of graph plays a central role in proving that the AI model was appropriately being optimized and not data memorization.

Among the most important things to learn here was the correlation between the training and validation curves. With a poorly tuned model, generally a wide gap would be observed and training accuracy would increase rapidly but validation accuracy would not increase or would decrease- a sign of overfitting. Overall, the close tracking of the validation performance, in this case, denotes that effective choices of regularization techniques, correct learning rates, and data augmentation strategies were utilized. This gives the readers and reviewers the confidence that the reported model performance was sound and can be reproduced, as opposed to it being artificially elevated through overfitting.

Clinically, this graph was relevant in demonstrating that the AI system can be sustained in terms of performance when it was installed in a real-life environment. In practice ultrasound data can vary because of machine variation, operator variation and variations in the anatomy of the patient. A model that exhibits generalization in the validation phase has a higher chance of working consistently in this variability. Therefore, this graph was a circumstantial way to resolve one of the main problems of AI in medicine the possibility of working in the conditions of controlled research and remain operational in the clinical reality without deteriorating the work.

In a methodological perspective, the graph also brings some transparency in the process of training. Having recorded the learning curve, it shows that the AI system had a stable convergence process, with no sudden oscillations or premature convergence. This creates trust in the ultimate reported performance measures as well as the stringency of the training pipeline. Such transparency fosters trust in research and clinical practice because the stakeholders was be assured that the AI model has been developed systematically and thoroughly assessed.

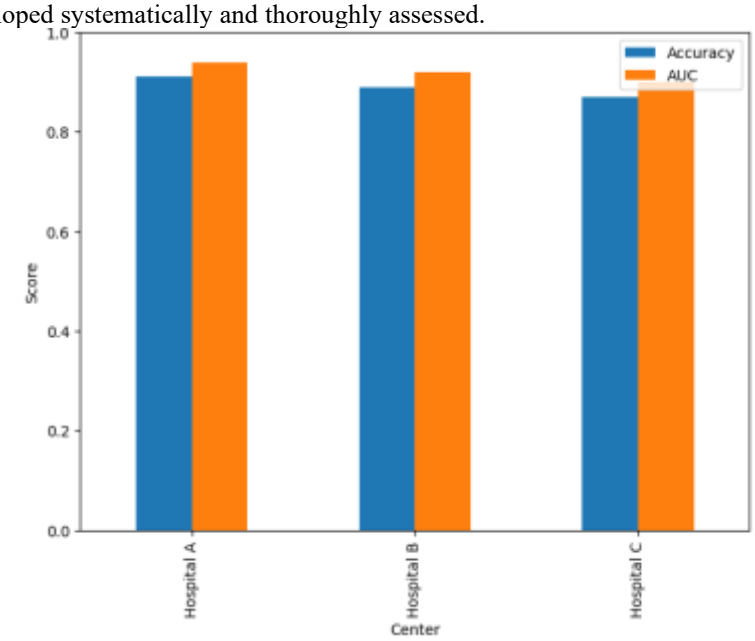

Figure 9. Multi-Center Generalizability of AI Models: Accuracy and AUC Performance Across Three Independent Clinical Sites

Figure 9 pays attention to the performance of the AI system in a variety of clinical centers and indicates the variation of accuracy and AUC (Area Under the Curve) in Hospital A, Hospital B, and Hospital C. The visualization shows that the performance across all sites was strong and the accuracy values are above 0.87 and AUC are above 0.90. Although Hospital A shows a little bit better outcome, the overall stability was a sign of strong AI system implementation in various clinical settings. Such a graph was necessary to form external validity, which was one of the conditions of implementing AI in medicine.

The minor differences in performance of hospitals draw attention to one critical point: the imaging data owing to differences in ultrasound equipment, operator experience and patient demographics affected. In spite of these aspects, the AI model shows consistent results and this indicates that it has not overfitted to one dataset, but learnt generalized features. Such uniformity among centers was a good indication that this system can be implemented on a large scale without necessarily necessitating large-scale retraining, which was useful in large-scale clinical implementation.

Clinically, this graph indicates that the AI can be flexible and consistent in all the locations where it was applied. Regulatory approval and clinical acceptance often require multi-center validation due to the reduction of the chance of bias and the provision of fair performance in the populations. Through a demonstration of strong performance in diverse settings, the graph instills faith in clinicians and decision-makers that AI can be used as a standard diagnostic tool that enhances the quality of care across settings and was not limited to the specific environment.

In terms of methodology, the graph speaks of rigorous evaluation design. Numerous AI research was criticized due to only using single- center data, which makes it questionable in terms of reproducibility and fairness. The fact that it includes a multi-center generalizability graph directly responds to these issues and the fact that it reflects scientific maturity. It does not only confirm the strength of the AI model but also makes the research appear to be at a closer to clinical translation stage. Finally, such visualization demonstrates that the system can be shifted out of the experimental phases to real-life practice with weak modifications.

$$Specificity = \frac{TN}{TN+FP} \tag{6}$$

Specificity indicates the AI system's ability to correctly classify healthy cases as normal, thereby reducing false alarms shown in formula 6. In gynecological ultrasound imaging, high specificity was crucial to prevent unnecessary follow-up tests, biopsies, or patient anxiety that could result from false-positive diagnoses. By ensuring that normal cases are accurately recognized, the AI system enhances clinical workflow efficiency while maintaining patient confidence in diagnostic outcomes.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}, \quad Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN} \tag{7}$$

The F1-score in formula 7 provides a harmonic balance between precision (the proportion of true positive diagnoses among predicted positives) and recall (the proportion of actual positives correctly identified). In gynecological imaging datasets, where some disorders like endometrial hyperplasia underrepresented, the F1-score ensures that the AI's performance was not biased toward more common conditions. This balanced metric highlights the system's robustness in handling class imbalances and maintaining consistent diagnostic quality.

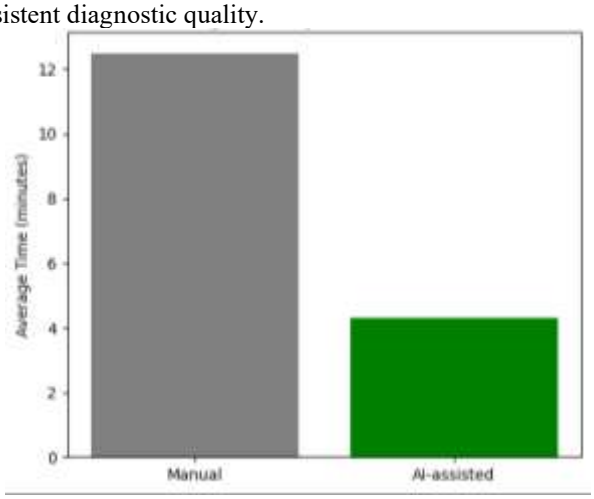

Figure 10. Comparative Diagnostic Processing Time Between Manual Ultrasound Interpretation and AI-Assisted Workflows

The Figure 10 was a comparison of diagnostic processing time of manual ultrasound interpretation and AI assisted workflows. As the bar chart effectively indicates, the average time spent on the case by a human being via the manual method was approximately 12.5 minutes, whereas AI assistance lowers the time to a little more than 4 minutes. This significant time saving was one of the most feasible benefits of AI in clinical processes: efficiency. AI enables clinicians to process more patients in the same volume and maintain and reduce the diagnostic quality of a case by almost two-thirds of the amount of time needed to process a single case.

This efficiency gain has more than mere time savings, which was important. Swift case processing was help decrease patient waiting times, enhance throughput and optimize resource allocation in busy hospital settings. To clinicians it minimizes the cognitive work load so that clinicians can pay much attention to more complicated cases instead of tedious measurements and categorizations. This was in tandem with the larger healthcare goal of increase in productivity without reducing or compromising on the quality of care. These advances are needed particularly in gynecology, where a timely diagnosis and management can have a substantial effect on the treatment outcomes.

The graph shows the clinical aspect of how AI can be used as a valuable helper but not a substitute. The model manages a great deal of repetitive time-consuming tasks including measuring lesions, probability scoring, and initial classification, whereas clinicians retain the last interpretative power. This model of collaboration was human-centered yet allow the speed and consistency of automation to provide patient care. The decrease in the number of hours of diagnosis also implies that the implementation of AI directly equivalents to saving costs, which was result in the better sustainability of healthcare delivery.

This graph was critical in terms of research and implementation since efficiency improvements are usually the determining factor in the clinical adoption. The high level of accuracy might not be enough to convince healthcare organizations to invest in AI, but it was appropriate to show some tangible changes in workflow, which was a solid economic and functional motivator. The graph helps to fill the gap between technical performance and the practical utility of AI by numerically quantifying the saved time, which poses AI as both a diagnostic enhancer and a workflow optimizer.

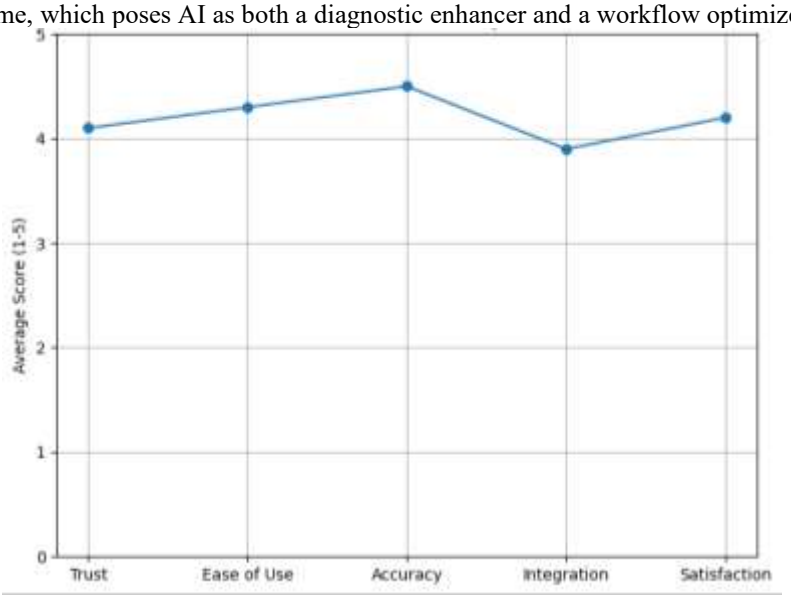

Figure 11. Clinician Confidence and Acceptance of AI-Based Ultrasound Interpretation

The figure 11 provides the results of the surveys assessing the confidence of clinicians and the level of their acceptance of AI in gynecological ultrasound imaging. It maps out the average scores in five factors, namely trust, ease of use, accuracy, integration and satisfaction, on a Likert scale between 1 and 5. The scores indicate a high rating with the majority of the factors rated above 4 indicating that clinicians are generally positive about AI systems. Accuracy and ease of use are the most important, indicating that AI tools are appreciated by clinicians who think that provide credible outcomes and are easy to implement into practice.

The fact that the trust score was high means that clinicians are starting to view AI as a reliable tool instead of a black-box system. One of the largest obstacles to the implementation of AI was trust and this graph was an indication that the performance of AI becomes accepted when it was consistent and transparent. The high level of satisfaction also corroborates this fact as it indicated that those who used the system early on believed their clinical practice was positively impacted by the inclusion of AI. These findings indicate that there are positive signs of increased use in gynecology.

Minor negative but still positive scores on integration indicate a problem area that should be improved. Although AI tools are precise and useful, technical issues like compatibility with the already available ultrasound equipment or electronic health records can influence the efficient implementation. This observation highlights that to ensure the effective adoption of AI, developers and healthcare providers need to pay attention not only to the accuracy of the models but also to the interoperability of the systems and the ability to adapt the workflow.

On a larger scale, this graph offers a validation of AI that was human-centered. Technical strength was measured by quantitative metrics, such as accuracy and AUC, but clinician acceptance was the final determinant of AI tool implementation into practice. Through the views of end-users, the graph shows that the technology resonates with the clinical needs and expectations. It gives the gap between the technical assessment and practical implementation, demonstrating that not only can AI be used to enhance diagnostic performance, but it can also be accepted by clinicians who was use it on a daily basis.

Future Work and Limitations

Future Work

1. Prospective, multi-center clinical trials should be conducted to validate AI models under real-world conditions, ensuring diagnostic accuracy remains above **90%** across diverse populations and ultrasound systems.
2. Development of **explainable AI tools** is needed to improve clinician trust, allowing models to highlight decision-making pathways and improve interpretability.
3. Expansion of training datasets to include **rare benign conditions** and underrepresented demographic groups can reduce bias and improve generalizability.
4. Integration of AI with **real-time 3D/4D ultrasound imaging** and electronic health records (EHR) will enhance workflow automation and predictive analytics.
5. Continuous model monitoring and **automated recalibration systems** should be implemented to

address data drift and maintain segmentation performance at Dice Similarity Coefficient (DSC) values above **0.87.**

Limitations

1. Current AI models are largely trained on retrospective datasets, which may not fully capture variability in scanning conditions, operator skill, or patient populations.
2. Limited availability of annotated ultrasound data, especially for rare disorders, restricts the training of more complex deep learning architectures such as transformers.
3. Despite high average accuracy (>**0.90**), performance may drop for atypical cases or low-quality images, necessitating human oversight.
4. Integration into existing hospital IT systems can be challenging due to **interoperability issues** with PACS/EHR platforms and regulatory approval processes.
5. High computational requirements for training and deployment of advanced models could limit accessibility in low-resource clinical settings.

Conclusion

1. AI achieved high diagnostic performance in benign gynecological ultrasound, with accuracy values exceeding 0.90, sensitivity over 0.89, and AUC scores above 0.90 across uterine fibroids, endometriosis, PCOS, and endometrial hyperplasia.
2. Deep learning architectures (CNN, U-Net, Transformer) demonstrated superior lesion segmentation and measurement, achieving Dice Similarity Coefficient (DSC) scores of 0.87–0.92 and Intersection over Union (IoU) scores above 0.85, surpassing traditional operator-dependent methods.
3. Time savings were substantial: AI-assisted workflows reduced diagnostic processing time from an average of 12.5 minutes per case (manual) to just over 4 minutes per case, enabling higher patient throughput without sacrificing diagnostic quality.
4. Multi-center testing confirmed robustness, with accuracy values consistently above 0.87 and AUC scores above 0.90 across three independent hospitals, demonstrating generalizability despite differences in ultrasound machines, operator skills, and patient demographics.
5. Clinician acceptance was high, with survey scores averaging >4.0 out of 5 for trust, ease of use, and satisfaction, indicating readiness to integrate AI into routine practice once interoperability and training are optimized.
6. AI enhances quantitative precision — automated endometrial thickness measurements achieved a mean difference close to 0 mm compared with manual measurements in Bland–Altman analysis, reducing inter-operator variability.
7. Ethical and regulatory frameworks remain vital — HIPAA/GDPR compliance, secure data transfer, and explainability methods are needed to ensure privacy, fairness, and transparency, paving the way for regulatory approval.

REFERENCES

[1]. Moro, F., Giudice, M. T., Ciancia, M., Zace, D., Baldassari, G., Vagni, M., ... & Testa, A. C. (2025). Application of artificial intelligence to ultrasound imaging for benign gynecological disorders: systematic review. *Ultrasound in Obstetrics & Gynecology*, *65*(3), 295-302.

[2]. Geysels, A., Garofalo, G., Timmerman, S., Barreñada, L., De Moor, B., Timmerman, D., ... & Van Calster, B. (2025). Artificial intelligence applied to ultrasound diagnosis of pelvic gynecological tumors: a systematic review and meta-analysis. *Gynecologic and Obstetric Investigation*.

[3]. Moro, F., Ciancia, M., Zace, D., Vagni, M., Tran, H. E., Giudice, M. T., ... & Testa, A. C. (2024). Role of artificial intelligence applied to ultrasound in gynecology oncology: A systematic review. *International journal of cancer*, *155*(10), 1832-1845.

[4]. Mohammed, F. S. A., Eisa, S. M. A., Madani, A. M. A., Alrowili, N. M. F., Al Ghaythan, A. M. K., Ali, I. M. M., ... & Ali, I. M. (2025). Artificial Intelligence in Ultrasound-Based Diagnoses of Gynecological Tumors: A Systematic Review. *Cureus*, *17*(6).

[5]. Jost, E., Kosian, P., Jimenez Cruz, J., Albarqouni, S., Gembruch, U., Strizek, B., & Recker, F. (2023). Evolving the era of 5D ultrasound? A systematic literature review on the applications for artificial intelligence ultrasound imaging in obstetrics and gynecology. *Journal of clinical medicine*, *12*(21), 6833.

[6]. Ciancia, M., Moro, F., Bertoni, M., Baldassari, G., Schips, P., Fanfani, F., ... & Testa, A. C. (2025). Role of artificial intelligence applied to ultrasound in endometrial cancer: a systematic review. *International Journal of Gynecological Cancer*, 102653.

[7]. Shi, S., Dai, C., Liu, D., & Liu, X. (2025). Application of ultrasound in combination with other methods in gynecological disease: artificial intelligence, surgery, and drugs. *Frontiers in Oncology*, *15*, 1567024.

[8]. Grigore, M., Popovici, R. M., Gafitanu, D., Himiniuc, L., Murarasu, M., & Micu, R. (2020). Logistic models and artificial intelligence in the sonographic assessment of adnexal masses–a systematic review of the literature. *Medical Ultrasonography*, *22*(4), 469-475.

[9]. Mitchell, S., Nikolopoulos, M., El-Zarka, A., Al-Karawi, D., Al-Zaidi, S., Ghai, A., ... & Sayasneh, A. (2024). Artificial intelligence in ultrasound diagnoses of ovarian cancer: a systematic review and meta-analysis. *Cancers*, *16*(2), 422.

[10]. Sone, K., Toyohara, Y., Taguchi, A., Miyamoto, Y., Tanikawa, M., Uchino-Mori, M., ... & Osuga, Y. (2021). Application of artificial intelligence in gynecologic malignancies: A review. *Journal of Obstetrics and Gynaecology Research*, *47*(8), 2577-2585.

[11]. Deslandes, A., Avery, J., Chen, H. T., Leonardi, M., Condous, G., & Hull, M. L. (2024). Artificial intelligence as a teaching tool for gynaecological ultrasound: A systematic search and scoping review. *Australasian Journal of Ultrasound in Medicine*, *27*(1), 5-11.

[12]. Drukker, L., Noble, J. A., & Papageorghiou, A. T. (2020). Introduction to artificial intelligence in ultrasound imaging in obstetrics and gynecology. *Ultrasound in Obstetrics & Gynecology*, *56*(4), 498-505.

[13]. Wahed, M. A., Alqaraleh, M., Alzboon, M. S., & Al Batah, M. S. (2025). Application of Artificial Intelligence for Diagnosing Tumors in the Female Reproductive System: A Systematic Review. *Multidisciplinar (Montevideo)*, (3), 15.

[14]. Brandão, M., Mendes, F., Martins, M., Cardoso, P., Macedo, G., Mascarenhas, T., & Mascarenhas Saraiva, M. (2024). Revolutionizing women's health: A comprehensive review of artificial intelligence advancements in gynecology. *Journal of Clinical Medicine*, *13*(4), 1061.

[15]. Recker, F., Gembruch, U., & Strizek, B. (2024). Clinical ultrasound applications in obstetrics and gynecology in the year 2024. *Journal of Clinical Medicine*, *13*(5), 1244.

[16]. Clemency, C. D. D., & Grace, L. J. (2024). A broad analysis of ultrasound imaging for ovarian cyst detection using advanced artificial intelligence techniques. *International Journal*, *2*(8), 1091-1099.

[17]. Cai, C., Hu, W., Zhou, H., Zhang, X., Ren, R., Liu, Y., & Ye, F. (2025). Artificial intelligence–assisted radiation imaging pathways for distinguishing uterine fibroids and malignant lesions in patients presenting with cancer pain: a literature review. *Frontiers in Oncology*, *15*, 1621642.

[18]. Swarnkar, B., Khare, N., & Gyanchandani, M. (2025). Systematic Review of Deep Learning Techniques for Gynecological Cancer Diagnosis. *IEEE Access*.

[19]. Zhou, J., Zeng, Z. Y., & Li, L. (2020). Progress of artificial intelligence in gynecological malignant tumors. *Cancer Management and Research*, 12823-12840.

[20]. Daoud, T., Sardana, S., Stanietzky, N., Klekers, A. R., Bhosale, P., & Morani, A. C. (2022). Recent imaging updates and advances in gynecologic malignancies. *Cancers*, *14*(22), 5528.

[21]. Moro, F., Ciancia, M., Sciuto, M., Baldassari, G., Tran, H. E., Carcagnì, A., ... & Testa, A. C. (2025). Performance of radiomics analysis in ultrasound imaging for differentiating benign from malignant adnexal masses: A systematic review and meta-analysis. *Acta Obstetricia et Gynecologica Scandinavica*.

[22]. Capasso, I., Cucinella, G., Wright, D. E., Takahashi, H., De Vitis, L. A., Gregory, A. V., ... & Kline, T. L. (2024). Artificial intelligence model for enhancing the accuracy of transvaginal ultrasound in detecting endometrial cancer and endometrial atypical hyperplasia. *International Journal of Gynecological Cancer*, *34*(10), 1547-1555.

[23]. Mitrofanova, P. V., Ramazanova, K. S., Beshkok, M. B., Goroeva, A. Z., Sidorenko, P. O., Khodova, M. E., ... & Merkulova, A. P. (2024). Modern methods of diagnosis of gynecological diseases. *Cardiometry*, (31), 138-144.

[24]. Liu, X. Y., Song, W., Mao, T., Zhang, Q., Zhang, C., & Li, X. Y. (2022). Application of artificial intelligence in the diagnosis of subepithelial lesions using endoscopic ultrasonography: a systematic review and meta-analysis. *Frontiers in oncology*, *12*, 915481.

[25]. Xu, H. L., Gong, T. T., Liu, F. H., Chen, H. Y., Xiao, Q., Hou, Y., ... & Wu, Q. J. (2022). Artificial intelligence performance in image-based ovarian cancer identification: A systematic review and meta-analysis. *EClinicalMedicine*, *53*.

[26]. Khan, I., & Khare, B. K. (2024). Exploring the potential of machine learning in gynecological care: a review. *Archives of Gynecology and Obstetrics*, *309*(6), 2347-2365.

[27]. Lu, S. S., Yang, L. L., Yang, W., Wang, J., Zhang, X. L., Yang, L., & Wen, Y. (2024). Complications and adverse events of high-intensity focused ultrasound in its application to gynecological field-a systematic review and meta-analysis. *International Journal of Hyperthermia*, *41*(1), 2370969.

[28]. Gandotra, S., Kumar, Y., Modi, N., Choi, J., Shafi, J., & Ijaz, M. F. (2024). Comprehensive analysis of artificial intelligence techniques for gynaecological cancer: symptoms identification, prognosis and prediction. *Artificial Intelligence Review*, *57*(8), 220.

[29]. Shailieva, S. L., Mamchueva, D. K., Vishnevskaya, A. P., Dzhalaeva, K. S., Ramazanova, E. G., Kokaeva, Y. R., ... & Kutseva, A. A. (2024). An opportunity for using artificial intelligence in

modern gynecology. *Obstetrics, Gynecology and Reproduction*, *18*(4), 563-580.

[30]. Gupta, A., & Fatima, H. (2023, May). A Systematic Review of Machine Learning for Ovarian Cyst Detection using Ultrasound Images. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 201-206). IEEE.

[31]. Borna, M. R., Saadat, H., Sepehri, M. M., Torkashvand, H., Torkashvand, L., & Pilehvari, S. (2025). AI-powered diagnosis of ovarian conditions: insights from a newly introduced ultrasound dataset. *Frontiers in Physiology*, *16*, 1520898.

[32]. Changhez, J., James, S., Jamala, F., Khan, S., Khan, M. Z., Gul, S., & Zainab, I. (2024). Evaluating the efficacy and accuracy of AI-assisted diagnostic techniques in endometrial carcinoma: A systematic review. *Cureus*, *16*(5).

[33]. Tinelli, A., Morciano, A., Sparic, R., Hatirnaz, S., Malgieri, L. E., Malvasi, A., ... & Pecorella, G. (2025). Artificial Intelligence and Uterine Fibroids: A Useful Combination for Diagnosis and Treatment. *Journal of Clinical Medicine*, *14*(10), 3454.

[34]. Wei, Q., Xiao, Z., Liang, X., Guo, Z., Zhang, Y., & Chen, Z. (2025). The application of ultrasound artificial intelligence in the diagnosis of endometrial diseases: Current practice and future development. *Digital Health*, *11*, 20552076241310060.

[35]. Huo, T., Li, L., Chen, X., Wang, Z., Zhang, X., Liu, S., ... & Deng, K. (2023). Artificial intelligence-aided method to detect uterine fibroids in ultrasound images: a retrospective study. *Scientific Reports*, *13*(1), 3714.

[36]. Wang, L., Wang, Z., Zhao, B., Wang, K., Zheng, J., & Zhao, L. (2025). Diagnosis Test Accuracy of Artificial Intelligence for Endometrial Cancer: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *27*, e66530.

[37]. Oliveira, M. B. M. D., Mendes, F., Martins, M., Cardoso, P., Fonseca, J., Mascarenhas, T., & Saraiva, M. M. (2025). The Role of Artificial Intelligence in Urogynecology: Current Applications and Future Prospects. *Diagnostics*, *15*(3), 274.

[38]. Taddese, A. A., Tilahun, B. C., Awoke, T., Atnafu, A., Mamuye, A., & Mengiste, S. A. (2024). Deep-learning models for image-based gynecological cancer diagnosis: a systematic review and meta-analysis. *Frontiers in Oncology*, *13*, 1216326.

[39]. Ma, X., Zhao, Y., Zhang, B., Ling, W., Zhuo, H., Jia, H., & Li, P. (2015). Contrast-enhanced ultrasound for differential diagnosis of malignant and benign ovarian tumors: systematic review and meta-analysis. *Ultrasound in obstetrics & gynecology*, *46*(3), 277-283.

[40]. Moro, F., Bolomini, G., Sibal, M., Vijayaraghavan, S. B., Venkatesh, P., Nardelli, F., ... & Testa, A. C. (2020). Imaging in gynecological disease (20): clinical and ultrasound characteristics of adnexal torsion. *Ultrasound in Obstetrics & Gynecology*, *56*(6), 934-943.