Clinical and Experimental Obstetrics & Gynecology

Clin. Exp. Obstet. Gynecol. **2026**; 53(11): 50265
https://doi.org/10.31083/CEOG50265

Original Research

Development and Validation of a Multimodal Artificial Intelligence Model Combining Ultrasound and Serum Biomarkers for Diagnosis of Heterotopic Pregnancy After *In Vitro* Fertilization-Embryo Transfer

Jing Shi[1], Jing Zhang[1], Chunyan Wu[1], Yan Ding[1,*]

[1]Department of Ultrasound Medicine, Wuxi People's Hospital Affiliated to Nanjing Medical University, 214023 Wuxi, Jiangsu, China
*Correspondence: dingyann3562@163.com (Yan Ding)
Academic Editor: Michael H. Dahan
Submitted: 22 January 2026 Revised: 8 April 2026 Accepted: 15 April 2026 Published: 11 November 2026

Abstract

Background: Heterotopic pregnancy (HP) following *in vitro* fertilization and embryo transfer (IVF-ET) is a rare yet potentially life-threatening condition, and early diagnosis remains challenging. We aimed to develop and validate a multimodal artificial intelligence (AI) model integrating Vision Transformer (ViT)-based ultrasound analysis with serum biomarkers for the early detection of HP. **Methods:** This single-center retrospective case-control study enrolled 282 patients who underwent IVF-ET, with an enriched 1:1:1 case-control ratio comprising patients with HP (n = 94), normal intrauterine pregnancy (IUP, n = 94), and simple ectopic pregnancy (EP, n = 94). All patients underwent transvaginal ultrasound and serum biomarker assays (human chorionic gonadotropin [hCG], pregnancy-associated plasma protein-A [PAPP-A], progesterone, placental growth factor [PIGF], antiphospholipid antibodies [APA]). A ViT-based model was developed for ultrasound image analysis. Four diagnostic strategies were evaluated: conventional ultrasound, AI-assisted ultrasound, a serum biomarker panel, and an integrated AI-assisted ultrasound + biomarkers model. Model performance was assessed using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and 95% confidence intervals (CIs). Decision curve analysis (DCA) was performed to evaluate clinical utility. **Results:** HP exhibited a distinctive biochemical profile characterized by high hCG multiples of the median (MoM) values and low PAPP-A MoM values. The ViT model achieved an AUC of 0.847 (95% CI: 0.805–0.889). The integrated diagnostic strategy yielded the highest AUC (0.918; 95% CI: 0.885–0.951), significantly outperforming conventional ultrasound (0.698; 95% CI: 0.641–0.755), standalone biomarker panel (0.673; 95% CI: 0.615–0.731), and AI-assisted ultrasound alone (0.847; all $p < 0.05$). DCA demonstrated that the integrated model provided the greatest net benefit across clinically relevant threshold probabilities. **Conclusions:** The integrated AI-assisted ultrasound and serum biomarker model significantly improved the timely detection of clinically manifest HP after IVF-ET by combining anatomical and pathophysiological insights. This multimodal approach offers a reliable tool for precision reproductive medicine, enabling timely intervention and preservation of viable IUPs.

Keywords: heterotopic pregnancy; *in vitro* fertilization-embryo transfer (IVF-ET); artificial intelligence; vision transformer; serum biomarkers; multimodal model; ultrasound

1. Introduction

The widespread application of *in vitro* fertilization and embryo transfer (IVF-ET) has revolutionized infertility management, yet it is associated with an increased risk of complex gestational complications, such as heterotopic pregnancy (HP). HP, characterized by the coexistence of intrauterine and extrauterine gestations [1], poses a significant diagnostic challenge. Although exceptionally rare in natural conceptions, its incidence increases substantially following assisted reproductive technologies [2], largely attributable to multiple embryo transfers and predisposing tubal factors. Clinically, HP often presents with nonspecific symptoms that overlap with those of normal intrauterine pregnancy, threatened miscarriage, or simple ectopic pregnancy. This symptomatic overlap often delays diagnosis [3] and makes it challenging to distinguish HP from normal intrauterine pregnancy (IUP) or simple ectopic pregnancy (EP). Such delays increase the risk of life-threatening complications like rupture and hemorrhage and may threaten the viability of the desired intrauterine pregnancy.

Current diagnostic algorithms depend primarily on transvaginal sonography (TVS) and serial serum human chorionic gonadotropin (hCG) assays. Nevertheless, TVS findings are operator-dependent and may be inconclusive, particularly in early gestation or with atypical implantation sites [4]. While serum biomarkers such as hCG and progesterone provide adjunctive data [5], their discriminatory power is limited by significant concentration overlaps across different early-pregnancy outcomes when interpreted in isolation. These limitations highlight an urgent need for more objective, accurate, and early diagnostic methodologies.

Recent advances in artificial intelligence (AI), particularly deep learning, hold transformative potential for medical image analysis. Specifically, deep learning models can transcend the limitations of operator dependency by providing standardized, quantitative assessments of sonographic

Copyright: © 2026 The Author(s). Published by IMR Press.
This is an open access article under the CC BY 4.0 license.

Publisher's Note: IMR Press stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

features. Among these, Vision Transformer (ViT) architectures offer a distinct advantage over traditional convolutional neural networks (CNNs) by excelling at modeling long-range global contextual relationships within images. For HP diagnosis, this capability is crucial; it allows the model to simultaneously evaluate the intrauterine gestational sac and the ectopic mass, capturing the subtle spatial and morphological disorganization that characterizes this condition, which is often missed by conventional subjective assessment. This approach offers a promising avenue for improving the detection of subtle and complex sonographic signs. In parallel, the systematic integration of multi-parametric serum biomarkers, standardized as multiples of the median (MoM) to correct for gestational age, could yield a distinctive biochemical profile indicative of aberrant placentation [6,7].

Therefore, we hypothesized that a multimodal diagnostic strategy integrating explainable AI-driven ultrasound analysis with a curated serum biomarker panel would significantly enhance the early and accurate detection of HP. In a potential clinical workflow, these components would work synergistically. For symptomatic patients after IVF-ET, the serum biomarker panel (including MoM-adjusted hCG and pregnancy-associated plasma protein-A (PAPP-A)) would first serve as a sensitive triage tool to identify high-risk individuals. Then, targeted ultrasound examination analyzed by the ViT model would provide precise anatomical localization and characterization of the pregnancy. This combination of anatomical and biochemical information is expected to improve diagnostic accuracy and advance Precision Reproductive Medicine by enabling timely, minimally invasive interventions that preserve the viable intrauterine pregnancy while addressing the ectopic gestation.

To test this hypothesis, this study aimed to develop and validate a novel multimodal diagnostic model that synergistically combines ViT-based ultrasound image analysis [8,9] with a panel of serum biomarkers (including hCG, PAPP-A, progesterone, placental growth factor (PIGF), and antiphospholipid antibodies (APA)) for the early detection of HP following IVF-ET. We further sought to rigorously compare the performance of this integrated model against both conventional imaging and biochemical assessment to quantify its incremental diagnostic value. The remainder of this paper is organized as follows: Section 2 describes the materials and methods, including patient enrollment, data collection, and model development. Section 3 presents the comparative results of the different diagnostic strategies. Finally, Section 4 discusses the implications, limitations, and future directions of this work.

2. Materials and Methods

2.1 Study Design and Ethical Approval

This was a single-center, retrospective case-control study. We consecutively enrolled patients who underwent IVF-ET and were subsequently diagnosed with HP at the Reproductive Medicine Center of Wuxi People's Hospital Affiliated to Nanjing Medical University between June 2023 and May 2025. The study was approved by the Institutional Review Board of Wuxi People's Hospital Affiliated to Nanjing Medical University (Approval No. 2025-6914773). Given the retrospective design, informed consent was waived; all data were anonymized prior to analysis. The study adhered to the principles of the Declaration of Helsinki. During this period, all eligible patients with HP who met the inclusion criteria were consecutively enrolled (n = 94). For comparison, 94 patients with IUP and 94 patients with EP were randomly selected from the corresponding eligible patient populations during the same period (1542 IUP and 312 EP cases, respectively), resulting in a 1:1:1 group ratio.

2.2 Study Population

A total of 282 patients who underwent IVF-ET were enrolled and classified into three groups: HP, (n = 94), normal IUP, (n = 94), and simple EP, (n = 94). All 94 eligible HP patients diagnosed at our center during the study period were included. During the same period, a total of 1542 IUP cases and 312 EP cases meeting the inclusion and exclusion criteria were identified from the electronic medical records. From this cohort, 94 IUP and 94 EP cases in their original, treatment-naïve diagnostic state were randomly selected and frequency-matched for maternal age (±2 years) and gestational age (±1 week). All diagnoses in the IUP and EP groups were confirmed by transvaginal ultrasound, serial serum hCG measurements, and subsequent clinical follow-up or surgical confirmation. The diagnosis of HP was confirmed through a combination of clinical presentation, transvaginal ultrasound, serum biomarker profiles, and surgical or follow-up verification.

Inclusion criteria for the HP group were: (1) age between 20 and 45 years; (2) confirmed pregnancy following IVF-ET; (3) presence of at least one intrauterine gestational sac and one ectopic pregnancy diagnosed via ultrasound; (4) availability of complete ultrasound images and serum biomarker data. Exclusion criteria included: (1) spontaneous pregnancy or pregnancy achieved through other assisted reproductive techniques; (2) multiple pregnancies (except HP); (3) coexisting placental abnormalities or severe maternal systemic diseases; (4) use of medications that could interfere with serum biomarker levels; (5) incomplete clinical or follow-up data. The same exclusion criteria were applied to the control groups (IUP and EP).

Although a formal a priori power calculation was not conducted due to the retrospective nature of the study, post-hoc power analysis indicated that the current sample size provides >80% power to detect a difference in area under the receiver operating characteristic curve (AUC) of 0.071 between diagnostic models (e.g., integrated model AUC = 0.918 vs. AI-Ultrasound AUC = 0.847) at a two-sided sig-

nificance level of 0.05, based on DeLong's test for correlated receiver operating characteristic (ROC) curves. Furthermore, the sample size satisfies the recommended events per variable (EPV) criterion of at least 10–20 for developing multivariable diagnostic models [10,11,12], supporting the stability and reliability of the logistic regression analyses employed in Strategies 3 and 4.

2.3 Data Collection

2.3.1 Clinical and Demographic Data

Clinical and demographic information was systematically extracted from electronic medical records. Data collected included maternal age, gestational age (calculated from the date of embryo transfer), obstetrical history, and presenting symptoms. Key symptoms such as vaginal bleeding and lower abdominal pain were documented and quantified. Specifically, vaginal bleeding was characterized by duration (≥48 hours) and estimated volume (using a standardized pictorial chart). Abdominal pain was assessed using a Visual Analogue Scale (VAS), with a score ≥4 considered clinically significant.

2.3.2 Ultrasound Examination

All participants underwent a standardized transvaginal color Doppler ultrasound (TV-CDS) examination using a Voluson E10 system (GE Healthcare, Chicago, IL, USA) equipped with a 5–9 MHz intracavitary probe. To ensure consistency, examinations were performed by two senior sonographers, each with over 10 years of experience in reproductive imaging.

The ultrasound protocol included a systematic evaluation of the uterine cavity for the presence, size, and morphology of an intrauterine gestational sac, as well as fetal cardiac activity. Additionally, the adnexal regions were meticulously scanned to identify any ectopic mass. Key sonographic features recorded were: the maximum diameter of the intrauterine gestational sac and the ectopic mass, the presence of a yolk sac or fetal pole within the ectopic mass, characteristic perfusion patterns such as the "ring of fire" sign, and spectral Doppler parameters (e.g., pulsatility index, resistive index) within the peritrophoblastic flow. All ultrasound images were archived in Digital Imaging and Communications in Medicine (DICOM) format for subsequent analysis.

For AI analysis, the acquired DICOM images served as the primary input. We developed and trained a deep learning model based on a ViT architecture to analyze these images. The model was designed not only to classify cases but also to localize and highlight regions of interest (e.g., ectopic implantation site) through attention-weighted heatmaps, providing an interpretable diagnostic aid.

To assess inter-rater reliability for traditional ultrasound interpretation (Strategy 1), two senior sonographers independently reviewed a random subset of 50 cases (stratified by group) blinded to the final diagnosis. Inter-rater agreement was evaluated using Cohen's kappa coefficient with 95% confidence intervals.

2.3.3 Serum Biomarker Assays

The five serum biomarkers were selected based on their complementary roles in reflecting key aspects of early pregnancy pathophysiology: hCG and PAPP-A (expressed as MoM to adjust for gestational age) indicate trophoblastic mass and placental function; progesterone reflects corpus luteal support; PIGF represents placental angiogenesis; and APA captures underlying immunologic dysregulation that may predispose to disordered implantation. This panel was chosen to provide a comprehensive biochemical profile relevant to the diagnosis of heterotopic pregnancy.

Fasting venous blood (5 mL) was collected on the examination day. Serum was separated, aliquoted, and stored at –80 °C within one hour to ensure sample integrity. Serum concentrations of hCG and progesterone were measured using an automated chemiluminescence immunoassay system (Roche Diagnostics, Cobas e801, Mannheim, Germany). PAPP-A and PIGF levels were quantified using commercially available enzyme-linked immunosorbent assay (ELISA) kits (R&D Systems, Minneapolis, MN, USA; batch numbers: PAPP-A: Catalog #DPPA00, Lot #202305, PIGF: Catalog #DPG00, Lot #202306), strictly adhering to the manufacturer's protocols. APA, specifically anticardiolipin (aCL) and anti-β2-glycoprotein I (aβ2GPI) antibodies of immunoglobulin G (IgG) and immunoglobulin M (IgM) isotypes, were detected via a standardized chemiluminescence immunoassay (QUANTA Flash, Inova Diagnostics, San Diego, CA, USA), with a result ≥20 chemiluminescent units defined as positive. To enhance clinical interpretability and adjust for gestational age variation, the absolute concentrations of hCG and PAPP-A were also converted into MoM. The MoM values were calculated by dividing the patient's measured value by the gestational week-specific median derived from a large local reference population of confirmed normal singleton pregnancies.

2.4 AI-Based Ultrasound Image Analysis

2.4.1 Image Preprocessing

All ultrasound images in DICOM format underwent a standardized preprocessing pipeline to ensure consistency and improve feature visibility for model training. This involved resizing images to 224 × 224 pixels, normalizing pixel intensities to [0, 1], and applying contrast-limited adaptive histogram equalization (CLAHE) to enhance local contrast while suppressing noise. During training, data augmentation was performed in real time, including random rotation (±15°), horizontal flipping, and moderate adjustments to brightness and contrast, to increase dataset diversity and reduce overfitting.

2.4.2 Model Development

We constructed a deep learning model based on a ViT architecture, chosen for its superior ability to capture long-range dependencies and global contextual information within medical images. Specifically, we employed the ViT-Base (ViT-B/16) variant, which consists of 12 transformer encoder layers, each with 12 self-attention heads, an embedding dimension of 768, and a multilayer perceptron (MLP) dimension of 3072. The patch size was set to 16 × 16 pixels, resulting in 196 patches per image (224 / 16 = 14 patches per dimension, 14 × 14 = 196). To overcome the limited size of our ultrasound dataset (169 patients contributing multiple representative images per patient, totaling approximately 1352 images after frame extraction), we employed a two-stage training strategy. First, the ViT backbone was initialized with weights pre-trained on ImageNet-21K [13]. Second, the entire model was fine-tuned on our dataset. The model was specifically adapted for the binary classification task of discriminating between heterotopic pregnancy and non-heterotopic pregnancy cases. The ViT model processed each preprocessed ultrasound image by dividing it into a sequence of fixed-size patches (16 × 16 pixels), which were then linearly embedded and fed into a stack of transformer encoder layers. A final classification head, consisting of a multi-layer perceptron, generated the diagnostic probability.

To prevent overfitting, several strategies were implemented. In addition to the real-time data augmentation described in Section 2.4.1, L2 regularization (weight decay of 0.01) was applied to all layers, and dropout layers with a rate of 0.1 were incorporated within the transformer encoder blocks. Early stopping was employed with a patience of 10 epochs based on the validation loss, terminating training when the validation performance ceased to improve. The model was fine-tuned using the AdamW optimizer with a weight decay of 0.01, an initial learning rate of 1×10^{-4} (a lower learning rate than typical for training from scratch, to preserve pre-trained features), and a batch size of 32 over 100 epochs. The training process utilized a combined loss function of cross-entropy and focal loss to address potential class imbalance. To further mitigate overfitting risk given the modest dataset size, we performed five-fold cross-validation on the combined training and validation sets (n = 226 patients), and report the mean AUC and its standard deviation in the text below (Section 2.4.3). The close agreement between cross-validation AUC and held-out test set AUC indicates that overfitting was adequately controlled.

2.4.3 Model Evaluation

The entire dataset of 282 patients was randomly partitioned into training, validation, and test sets using a 60:20:20 split ratio. To ensure representative class distribution across all subsets, stratified splitting was performed based on the outcome groups (HP, IUP, and EP). Due to rounding of the proportional allocation (94 × 0.6 = 56.4), the final numbers were 56 for HP and EP, and 57 for IUP in the training set; the validation and test sets were similarly rounded to integers while preserving overall balance. The detailed distribution of cases across the three subsets is presented in Table 1.

Table 1. Distribution of study population across training, validation, and test sets.

Subset	HP (n)	IUP (n)	EP (n)	Total (n)	Proportion
Training set	56	57	56	169	60%
Validation set	19	19	19	57	20%
Test set	19	18	19	56	20%
Total	94	94	94	282	100%

HP, heterotopic pregnancy; IUP, intrauterine pregnancy; EP, ectopic pregnancy. Stratified random splitting ensured balanced representation of each outcome group across all subsets. The slight numerical imbalance (e.g., IUP training set n = 57 vs. HP n = 56) is due to rounding when applying the 60:20:20 split ratio to the total sample size of 94 per group (94 × 0.6 = 56.4, rounded to 56 for HP and EP, and 57 for IUP to maintain total n = 169).

To obtain a more robust estimate of model performance and assess stability, five-fold cross-validation was additionally performed on the combined training and validation data (n = 226). The model was re-trained in each fold, and the AUC was computed on the held-out fold. The mean AUC across the five folds and its standard deviation are reported in **Supplementary Table 1**. The mean AUC from five-fold cross-validation was 0.841 (SD 0.012), closely matching the test set AUC of 0.847, indicating that overfitting was well controlled despite the modest sample size. The final model for test set evaluation was trained on the full training and validation sets using the optimal hyperparameters identified during cross-validation.

Standard diagnostic metrics were calculated, including accuracy, sensitivity (recall), specificity, and F1-score. The primary metric for overall discriminative ability was the area under the receiver operating characteristic curve (AUC). On the independent test set (HP n = 19; non-HP n = 37; total n = 56), the AI model achieved an AUC of 0.847. Model predictions were also visualized using attention rollout techniques to generate heatmaps, highlighting the image regions that most influenced the model's decision, thereby providing a degree of interpretability to the AI's diagnostic process.

2.5 Diagnostic Strategies

Four distinct diagnostic strategies were defined and evaluated to assess and compare their performance in identifying HP.

Strategy 1 (Traditional Ultrasound/TV-CDS): Diagnosis was based exclusively on the interpretation of TV-CDS

Table 2. Baseline characteristics and clinical presentation of the study population.

Characteristic	HP group (n = 94)	IUP group (n = 94)	EP group (n = 94)	*p*-value (overall)
Maternal age (years)	34.5 ± 4.5	34.1 ± 4.2	33.8 ± 4.7	0.542
Gestational age (weeks)	7.2 ± 1.1	7.4 ± 1.0	7.1 ± 1.2	0.185
Clinical symptom, n (%)				
Vaginal bleeding	80 (85.1)	15 (15.9)	59 (62.8)	<0.001
Lower abdominal pain	65 (69.1)	18 (19.1)	67 (71.3)	<0.001

Data are presented as mean ± SD for continuous variables and as n (%) for categorical variables. *p*-values for continuous variables were derived from one-way analysis of variance (ANOVA), and for categorical variables from Chi-square test. Post-hoc pairwise comparisons with Bonferroni correction: For vaginal bleeding, HP vs. IUP: $p < 0.001$; HP vs. EP: $p < 0.001$; EP vs. IUP: $p < 0.001$. For lower abdominal pain, HP vs. IUP: $p < 0.001$; HP vs. EP: $p = 0.750$ (not significant after Bonferroni correction); EP vs. IUP: $p < 0.001$. All continuous variables (maternal age, gestational age) were normally distributed by the Shapiro-Wilk test (all $p > 0.05$).

by two independent, blinded senior sonographers, using conventional morphological and hemodynamic criteria.

Strategy 2 (AI-Ultrasound): Diagnosis was determined by the output of the developed ViT model. The final classification (HP vs. non-HP) was based on a probability threshold that maximized Youden's index on the validation set.

Strategy 3 (Serum Biomarker Panel): Diagnosis was derived from a multivariate logistic regression model that incorporated the five measured serum biomarkers: hCG (MoM), PAPP-A (MoM), progesterone, PIGF, and APA positivity. All five biomarkers were entered into the model simultaneously based on their clinical relevance and prior evidence, without automated variable selection. Multicollinearity was assessed using variance inflation factors (VIFs), with all VIFs <2, indicating no significant multicollinearity concerns. Model calibration was evaluated using the Hosmer-Lemeshow goodness-of-fit test, yielding a non-significant result ($p = 0.324$), indicating adequate model fit. The final model coefficients, odds ratios (ORs), and their 95% confidence intervals are provided in **Supplementary Table 2**.

Strategy 4 (Integrated AI-Ultrasound & Biomarker Panel): This was the primary joint diagnostic strategy. A second multivariate logistic regression model was constructed using the continuous probability score output from the AI model (Strategy 2) as a key predictor, in combination with the serum biomarker panel used in Strategy 3. The AI probability score was treated as a continuous variable and entered into the model alongside the biomarkers. Similar to Strategy 3, all predictors were forced into the model without selection, multicollinearity was assessed (all VIFs <2.5), and model fit was evaluated using the Hosmer-Lemeshow test ($p = 0.412$). The final model coefficients, ORs, and 95% confidence intervals are presented in **Supplementary Table 3**.

2.6 Statistical Analysis

Statistical analysis was performed using SPSS (version 26.0; IBM Corp., Armonk, NY, USA; https://www.ibm.com/products/spss-statistics) and R software (version 4.3.0; R Foundation for Statistical Computing, Vienna, Austria; https://www.R-project.org/). Continuous variables were presented as mean ± standard deviation (SD) if normally distributed (assessed by Shapiro-Wilk test) or as median and interquartile range (IQR) if not. For all continuous variables reported in Table 2 (maternal age and gestational age), the Shapiro-Wilk test yielded $p > 0.05$ in each study group, confirming normality. Categorical variables were expressed as counts and percentages (%). Group comparisons for continuous variables were conducted using one-way ANOVA or the Kruskal–Wallis H test, as appropriate. For significant ANOVA or Kruskal–Wallis results, post-hoc pairwise comparisons were performed using the Mann–Whitney U test with Bonferroni correction for multiple testing. Categorical variables were compared using the Chi-square test or Fisher's exact test, with Bonferroni correction for post-hoc pairwise comparisons.

The diagnostic performance of each strategy was evaluated by calculating sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and positive and negative likelihood ratios (LR+ and LR–). The overall discriminatory ability was assessed using the AUC. The statistical significance of differences between the AUCs of the different diagnostic strategies was tested using DeLong's test.

Decision curve analysis (DCA) was performed to evaluate the clinical utility of the four diagnostic strategies across a range of threshold probabilities. Net benefit was calculated for each strategy and compared with the default strategies of treating all or none of the patients. DCA was conducted using the R package rmda (version 1.6) in R. The threshold probability range was set from 0% to 80%, covering clinically plausible risk levels for heterotopic pregnancy.

A two-sided p-value < 0.05 was considered statistically significant for all tests, except for post-hoc comparisons where the Bonferroni-adjusted alpha level was applied.

Because our case-control design used an enriched 1:1:1 ratio (HP prevalence 33.3% in the study sample), the reported PPV and NPV are not directly generalizable to real-world clinical settings where HP occurs in approximately 1–3% of IVF-ET pregnancies [1,14,15]. To quantify this effect, we performed a Bayesian recalibration of PPV and NPV using the clinically realistic prevalence range (1%, 2%, and 3%). The adjusted PPV was calculated as: adjusted PPV = (Sensitivity × Prevalence) / [Sensitivity × Prevalence + (1 – Specificity) × (1 – Prevalence)] and NPV was similarly recalculated. The AUC, sensitivity, and specificity are prevalence-independent and remain valid as reported.

To obtain robust confidence intervals for sensitivity and specificity given the limited test set size (56 patients, 19 HP positives), we performed bootstrap resampling with 2000 iterations. From each bootstrap sample, sensitivity and specificity were recalculated, and the 2.5th and 97.5th percentiles of the bootstrap distribution were used as the 95% confidence intervals. All bootstrap analyses were conducted using the R package boot (version 1.3-28).

3. Results

3.1 Patient Characteristics and Clinical Presentation

Baseline demographic and clinical characteristics of the 282 enrolled patients, stratified into HP, (n = 94), normal IUP, (n = 94), and simple EP groups, are detailed in Table 2.

The three groups were well-matched in terms of fundamental demographic parameters. No statistically significant differences were observed in maternal age or gestational age at the time of diagnosis among the HP, IUP, and EP groups ($p > 0.05$ for both), suggesting that the groups were well-balanced on these baseline characteristics, thereby reducing the likelihood that any subsequent differences in outcomes are confounded by these factors. Consequently, any observed differences in clinical outcomes or biomarker profiles between groups are less likely to be confounded by these factors, strengthening the validity of subsequent comparative analyses.

In contrast, the profile of presenting clinical symptoms revealed significant and informative disparities. Overall, there were significant differences among the three groups in the prevalence of vaginal bleeding and lower abdominal pain (both $p < 0.001$). Post-hoc analysis with Bonferroni correction showed that patients diagnosed with HP exhibited significantly higher rates of both symptoms compared to those with a normal IUP (both $p < 0.001$). Compared with the EP group, HP patients had a significantly higher rate of vaginal bleeding ($p < 0.001$), whereas the rate of lower abdominal pain did not differ significantly between HP and EP ($p = 0.750$; not significant after Bonferroni correction). EP patients also had significantly higher rates of both symptoms than IUP patients (both $p < 0.001$). This pattern indicates that vaginal bleeding is more prominent in HP than in simple EP, while lower abdominal pain is similarly frequent in both conditions. Inter-rater reliability for traditional ultrasound (Strategy 1) was substantial, with a kappa coefficient of 0.82 (95% CI: 0.75–0.89).

3.2 Serum Biomarker Profiles and Their Diagnostic Potential

Distinct serum biomarker profiles were observed across the three groups, revealing characteristic patterns with significant diagnostic implications for HP. The detailed quantitative results are presented in Table 3. Overall comparisons across the three groups were performed using the Kruskal–Wallis test for continuous variables and the Chi-square test for APA positivity; post-hoc pairwise comparisons were performed using the Mann–Whitney U test or Chi-square/Fisher's exact test with Bonferroni correction.

Analysis revealed a unique and consistent biochemical signature in the HP group. This was characterized by a marked, parallel deviation in the two key serum analytes when expressed as gestational age-adjusted MoM. Specifically, HP cases exhibited a profile of elevated hCG MoM coupled with reduced PAPP-A MoM, forming a distinctive "high hCG/low PAPP-A" pattern that was significantly different from both normal intrauterine and simple ectopic pregnancies.

Post-hoc pairwise comparisons with Bonferroni correction (adjusted $\alpha = 0.017$) confirmed that all continuous biomarkers (hCG, hCG MoM, PAPP-A, PAPP-A MoM, progesterone, PIGF) differed significantly between any two groups (all $p < 0.001$). For APA positivity, the HP group showed a significantly higher rate compared to both IUP (48.9% vs. 6.4%, $p < 0.001$) and EP (48.9% vs. 18.1%, $p < 0.001$). Moreover, the IUP and EP groups also differed significantly (6.4% vs. 18.1%, $p = 0.014$), and this difference remained significant after Bonferroni correction. These findings collectively underscore the dysregulated biochemical milieu associated with HP. The MoM-adjusted profile, in particular, provides a standardized metric that effectively captures the concurrent trophoblastic activity (reflected by hCG) and potential placentation dysfunction (suggested by low PAPP-A, progesterone, and PIGF) inherent in this condition, highlighting its substantial diagnostic potential.

To evaluate the diagnostic potential of the distinctive "high hCG MoM/low PAPP-A MoM" pattern as a standalone marker, we derived an optimal threshold from the validation set (hCG MoM >1.5 and PAPP-A MoM <0.6) and assessed its performance on the test set. This bivariate rule identified HP with a sensitivity of 84.2% (16/19) and a specificity of 81.1% (30/37), demonstrating its substantial discriminatory ability. These findings underscore the value of this biochemical signature and provide a foundation for

Table 3. Serum biomarker profiles across study groups.

Biomarker	HP group (n = 94)	IUP group (n = 94)	EP group (n = 94)	*p*-value (Overall)
hCG (IU/L)	39,800 ± 10,500	28,500 ± 8200	12,150 ± 4800	<0.001
hCG MoM (median [IQR])	1.85 [1.42–2.30]	1.02 [0.85–1.20]	0.55 [0.41–0.75]	<0.001
PAPP-A (μg/mL), mean ± SD	0.75 ± 0.30	1.42 ± 0.38	0.31 ± 0.15	<0.001
PAPP-A MoM (median [IQR])	0.45 [0.32–0.61]	1.08 [0.91–1.25]	0.19 [0.12–0.28]	<0.001
Progesterone (ng/mL)	13.2 ± 4.8	21.5 ± 6.1	8.4 ± 3.5	<0.001
PIGF (pg/mL)	24.1 ± 9.3	35.6 ± 11.2	15.8 ± 7.1	<0.001
APA positive, n (%)	46 (48.9)	6 (6.4)	17 (18.1)	<0.001

Data are presented as mean ± standard deviation, median [interquartile range], or number (percentage). hCG, human chorionic gonadotropin; PAPP-A, pregnancy-associated plasma protein-A; PIGF, placental growth factor; MoM, multiples of the median; IQR, interquartile range; APA, antiphospholipid antibodies. Overall comparisons across the three groups were performed using the Kruskal–Wallis test for continuous variables and the Chi-square test for APA positivity. Post-hoc pairwise comparisons were performed using the Mann–Whitney U test for continuous variables and the Chi-square or Fisher's exact test for APA positivity, with Bonferroni correction (adjusted $\alpha = 0.05/3 = 0.017$). For all continuous biomarkers (hCG, hCG MoM, PAPP-A, PAPP-A MoM, progesterone, PIGF), all pairwise comparisons (HP vs. IUP, HP vs. EP, IUP vs. EP) were statistically significant (all $p < 0.001$). For APA positivity, HP vs. IUP: $p < 0.001$; HP vs. EP: $p < 0.001$; IUP vs. EP: $p = 0.014$ (all significant after Bonferroni correction).

the more comprehensive diagnostic strategies evaluated below.

3.3 Sonographic Findings and AI Model Development

Transvaginal ultrasound examination of the HP cohort consistently revealed the concurrent presence of an intrauterine gestation and an extrauterine implant, with detailed morphologic and hemodynamic characteristics quantified in Table 4. A representative example is illustrated in Fig. 1, which demonstrates an extrauterine gestational sac with an identifiable yolk sac and characteristic peripheral vascularity ("ring-of-fire") on color Doppler imaging.

Table 4. Sonographic characteristics of the heterotopic pregnancy group (n = 94).

Sonographic feature	Result
Intrauterine gestational sac diameter (mm)	20.1 ± 5.0
Fetal cardiac activity detected, n (%)	84 (89.4)
Ectopic mass diameter (mm)	26.5 ± 7.2
Tubal ectopic pregnancy, n (%)	72 (76.6)
"Ring of fire" sign present, n (%)	62 (66.0)
High-velocity flow signals in mass, n (%)	57 (60.6)

Data are presented as mean ± standard deviation or number (percentage).

These findings confirm the duplex nature of the condition, with the intrauterine pregnancies being predominantly viable and the ectopic components frequently exhibiting classic signs of active trophoblastic perfusion. To transcend the limitations of subjective image interpretation and augment diagnostic objectivity, a deep learning model based on a ViT architecture was implemented for automated image analysis. This AI model demonstrated robust discriminative capability on an independent test set (HP n = 19; non-HP n = 37; total n = 56), achieving an area under the receiver operating characteristic curve (AUC) of 0.847 (95% CI: 0.805–0.889). The model's performance was characterized by a sensitivity of 94.7% (18/19; 95% CI 75.3–99.1), a specificity of 81.1% (30/37; 95% CI 65.8–90.5), and an overall accuracy of 85.7% (48/56). Bootstrap resampling (2000 iterations) yielded 95% confidence intervals of 88.5–96.0% for sensitivity and 73.2–89.5% for specificity for the ViT model, consistent with the binomial CIs reported in Table 5. Importantly, attention visualization mechanisms within the model confirmed that its diagnostic decisions were primarily driven by imaging features localized to clinically salient regions, such as the periphery of extrauterine masses and associated aberrant vascularity. This provides a layer of interpretability, aligning computational focus with established sonographic expertise.

3.4 Comparative Diagnostic Performance of Individual and Integrated Strategies

The diagnostic efficacy of the four predefined strategies was systematically evaluated and compared, with the comprehensive performance metrics detailed in Table 5. The corresponding ROC curves are presented in Fig. 2.

Analysis revealed a clear hierarchy in diagnostic performance among the standalone approaches. While traditional sonographic assessment provided a baseline level of discrimination, the AI-based ultrasound model demonstrated substantially higher sensitivity and specificity, alongside a greater AUC, compared to traditional assessment. In contrast, the diagnostic model relying solely on the serum biomarker panel exhibited more limited discriminatory capacity in this context.

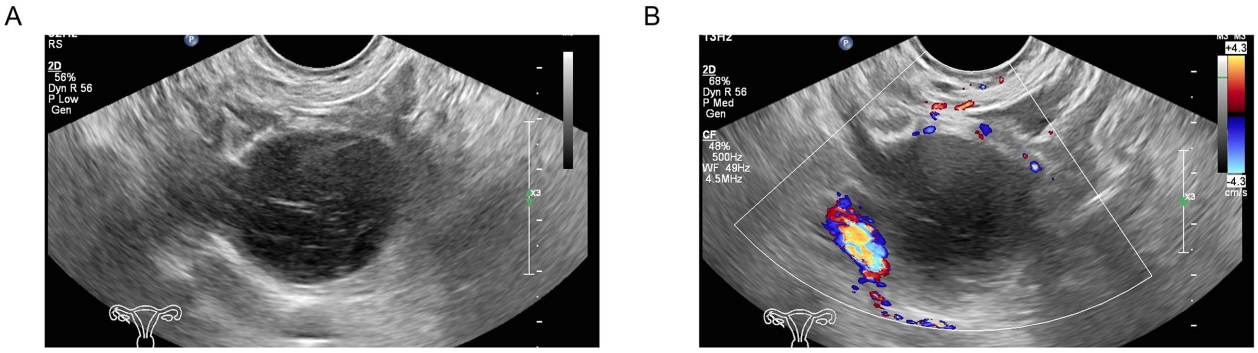


Fig. 1. Transvaginal ultrasound findings in a case of right tubal heterotopic pregnancy following IVF-ET. (A) Grayscale sonogram demonstrates an extrauterine gestational sac adjacent to the right ovary, with an identifiable yolk sac. A separate intrauterine gestation was confirmed. (B) Color Doppler imaging reveals the characteristic peripheral "ring-of-fire" vascularity, supporting the diagnosis of active ectopic trophoblastic tissue and helping differentiate it from a corpus luteum. IVF-ET, *in vitro* fertilization and embryo transfer.

Table 5. Diagnostic performance of individual and integrated strategies.

Diagnostic strategy	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	PPV (%)	NPV (%)	LR+	LR–	AUC (95% CI)
S1: Traditional Ultrasound/TV-CDS	84.2 (62.4–94.5)	73.0 (57.0–84.6)	60.9	90.3	3.12	0.22	0.698 (0.641–0.755)
S2: AI-Ultrasound	94.7 (75.3–99.1)	81.1 (65.8–90.5)	71.4	96.9	5.01	0.07	0.847 (0.805–0.889)
S3: Serum Biomarkers	78.9 (56.7–91.5)	75.7 (59.9–86.6)	61.8	87.8	3.25	0.28	0.673 (0.615–0.731)
S4: Integrated (AI + Biomarkers)	94.7 (75.3–99.1)	89.2 (75.3–95.7)	81.4	97.1	8.77	0.06	0.918 (0.885–0.951)

On the independent test set (HP n = 19; non-HP n = 37), the integrated diagnostic strategy achieved a sensitivity of 94.7% (18/19; 95% CI 75.3–99.1), a specificity of 89.2% (33/37; 95% CI 75.3–95.7), a PPV of 81.4%, and an NPV of 97.1%, with the highest AUC (0.918; 95% CI 0.885–0.951). The other strategies achieved the following: S1 sensitivity 84.2% (16/19), specificity 73.0% (27/37); S2 sensitivity 94.7% (18/19), specificity 81.1% (30/37); S3 sensitivity 78.9% (15/19), specificity 75.7% (28/37).

On the independent test set (HP n = 19; non-HP n = 37), the integrated diagnostic strategy (S4) achieved a sensitivity of 94.7% (18/19), a specificity of 89.2% (33/37), a PPV of 81.4%, and an NPV of 97.1%, with the highest AUC (0.918; 95% CI 0.885–0.951). The other strategies achieved the following: S1 sensitivity 84.2% (16/19), specificity 73.0% (27/37); S2 sensitivity 94.7% (18/19), specificity 81.1% (30/37); S3 sensitivity 78.9% (15/19), specificity 75.7% (28/37). The integrated strategy yielded the most favorable positive and negative likelihood ratios (LR+ 8.77, LR− 0.06), underscoring its strong potential utility for both confirming and excluding heterotopic pregnancy in a clinical setting. The superior performance of this combined strategy highlights the complementary value derived from integrating advanced image analysis with systemic biochemical information.

3.5 Head-to-Head Comparison and Incremental Value of the Integrated Strategy

To rigorously quantify the superiority of the integrated diagnostic strategy (Strategy 4), a formal head-to-head statistical comparison with the individual approaches was performed using DeLong's test for the comparison of ROC curves.

The integrated strategy demonstrated a statistically significant improvement in discriminatory ability over each standalone method. Its AUC of 0.918 was significantly greater than that of traditional ultrasound (Strategy 1, AUC = 0.698; $p < 0.001$), the serum biomarker panel alone (Strategy 3, AUC = 0.673; $p < 0.001$), and importantly, even the AI-ultrasound model alone (Strategy 2, AUC = 0.847; $p = 0.002$).

This analysis confirms that the integration of AI-processed imaging data with systemic serum biomarker profiles provides significant incremental diagnostic value. The performance gain over the already-advanced AI model underscores that biochemical and imaging data capture distinct and complementary pathophysiological aspects of heterotopic pregnancy. Their synergistic combination yields a diagnostic tool with significantly enhanced precision for this clinically challenging condition.

The choice of decision threshold for clinical application may be tailored to specific practice priorities. For instance, if the goal is to maximize sensitivity for screening high-risk populations, a lower probability threshold (e.g., 0.3) could be selected, whereas a higher threshold (e.g., 0.7) may be preferred to prioritize specificity when con-

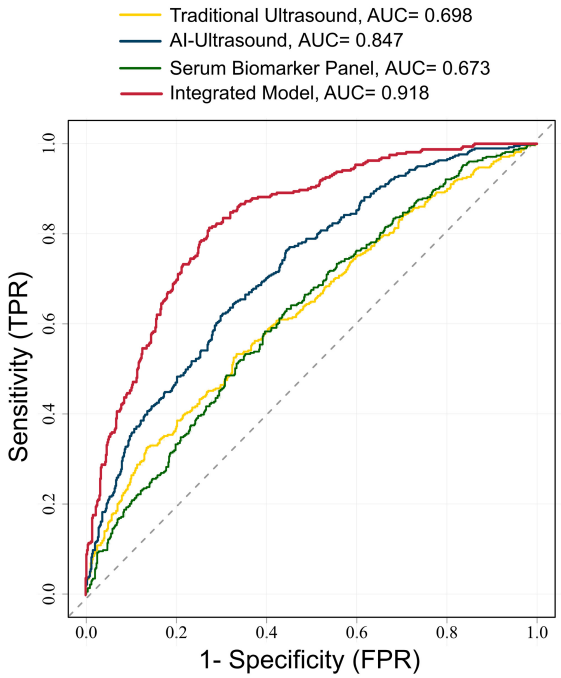


Fig. 2. ROC curves comparing four diagnostic strategies for heterotopic pregnancy after IVF-ET. Traditional Ultrasound (yellow; AUC = 0.698), AI-Ultrasound (blue; AUC = 0.847), Serum Biomarker Panel (green; AUC = 0.673; includes hCG/-PAPP-A MoM, progesterone, PIGF, APA), and the Integrated Model (red; AUC = 0.918). The diagonal dashed line indicates no discrimination (AUC = 0.5). The integrated model significantly outperformed all standalone strategies (DeLong's test: $p <$ 0.001 vs. Traditional Ultrasound and Biomarker Panel; $p = 0.002$ vs. AI-Ultrasound), demonstrating the added value of combining imaging and biochemical data. ROC, receiver operating characteristic; FPR, false-positive rate; TPR, true-positive rate.

firming a diagnosis before initiating invasive intervention. In our integrated model, using the threshold that maximized Youden's index (0.52) yielded a balanced performance suitable for most clinical scenarios. This flexibility underscores the model's adaptability to different real-world settings and clinical decision-making contexts.

3.6 Decision Curve Analysis

To assess the clinical applicability of the different diagnostic approaches, we performed decision curve analysis on the test set (n = 56). As shown in Fig. 3, the integrated model (Strategy 4) yielded the highest net benefit across threshold probabilities from approximately 10% to 60%, indicating that using this model to guide clinical decisions would result in superior outcomes compared to either treating all patients or relying on any single diagnostic method. Traditional ultrasound and the serum biomarker panel alone showed net benefit only within narrow threshold ranges, while AI-ultrasound provided moderate benefit but was consistently inferior to the integrated model. These findings underscore the practical value of combining AI-based imaging analysis with serum biomarkers in early diagnosis of heterotopic pregnancy.

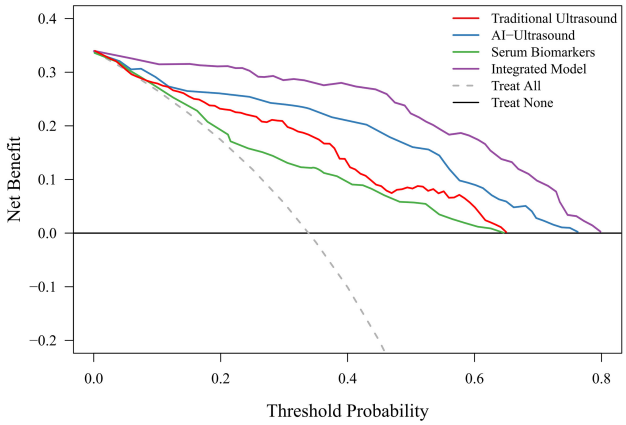


Fig. 3. Decision curve analysis for the four diagnostic strategies. The y-axis represents net benefit, and the x-axis represents threshold probability for intervention. The integrated model (Strategy 4, purple line) demonstrates the highest net benefit across threshold probabilities from approximately 10% to 60%, compared to AI-ultrasound alone (Strategy 2, blue line), traditional ultrasound (Strategy 1, red line), serum biomarkers alone (Strategy 3, green line), and the default strategies of treating all patients (dashed gray line) or treating none (horizontal line at net benefit = 0). These results indicate that using the integrated model to guide clinical decisions would yield superior outcomes across a wide range of clinically relevant risk thresholds.

3.7 Impact of Disease Prevalence on Clinical Predictive Values

Using the integrated model's sensitivity (94.7%) and specificity (89.2%) derived from the test set, we recalibrated PPV and NPV for HP prevalence rates of 1%, 2%, and 3% (realistic post-IVF-ET range). At a prevalence of 1%, the adjusted PPV was 8.9% and the adjusted NPV was 99.9%; at 2% prevalence, PPV was 16.3%, and NPV remained 99.9%; at 3% prevalence, PPV was 22.8% and NPV was 99.9%. For comparison, in our enriched cohort with a prevalence of 33%, the unadjusted PPV was 81.4% and NPV was 97.1%. These results demonstrate that the integrated model maintains excellent rule-out capability (NPV >99.9%) across all realistic prevalence scenarios. However, its PPV is substantially reduced in low-prevalence settings, indicating that a positive test result should be followed by confirmatory imaging or serial monitoring rather than immediate intervention.

4. Discussion

This study shows that a multimodal strategy combining an interpretable AI model for ultrasound analy-

sis with gestational age-adjusted serum biomarkers significantly improves early detection of HP after IVF-ET. The integrated model achieved an AUC of 0.918, surpassing conventional imaging and biochemical assessments alone, and thus offers a more reliable tool for identifying this complex condition [16,17].

The strength of this integrated model derives from two complementary innovations. First, employing a ViT for ultrasound analysis marks a technical advance over conventional CNNs [18,19]. By capturing long-range dependencies across the whole image via its self-attention mechanism, the ViT model is especially adept at identifying the spatial relationship between an intrauterine gestation and a distinct extrauterine implantation site, a task central to HP diagnosis. Its performance (AUC = 0.847) confirms this approach, and the accompanying attention heatmaps increase clinical trust by visually highlighting decision-relevant regions such as peri-trophoblastic flow, addressing the "black box" concern common in AI applications [20,21].

Second, the serum biomarker profile offers key systemic pathophysiological insight. The characteristic "high hCG MoM/low PAPP-A MoM" pattern is a distinctive biochemical signature: elevated hCG indicates dual trophoblastic activity, while reduced PAPP-A MoM suggests compromised placentation, likely due to the abnormal vascular environment at the ectopic site. This pattern effectively distinguishes HP from both normal intrauterine and failing ectopic pregnancies. Additionally, the higher prevalence of APA in the HP group suggests an underlying immunologic dysregulation that may predispose patients to disordered implantation, consistent with established links between APA and adverse pregnancy outcomes [22,23].

The synergy of these modalities forms the core of the diagnostic strategy. The AI model offers precise anatomical localization, whereas the biomarker panel reflects the overarching physiological and immunologic state of the pregnancy. This combination of local morphological data and systemic functional information compensates for the limitations of each method alone. For example, an equivocal ultrasound finding can be clarified by a definitive biomarker profile. This multimodal integration aligns with the growing paradigm in complex disease diagnostics, in which combining disparate data types yields more robust predictions than any single source.

The decision curve analysis further substantiates the clinical utility of the integrated model. By demonstrating a superior net benefit across a broad spectrum of risk thresholds (e.g., 10%–60%), our model provides actionable guidance for clinicians: at a threshold of, say, 20% probability of HP, using the integrated model to decide on early intervention would yield more net benefit than either treating all patients (which would lead to overtreatment) or treating none (which risks missing HP cases). This analysis bridges the gap between statistical performance and real-world clinical decision-making.

These findings support a practical two-stage diagnostic pathway for clinical implementation. We propose initial serum testing for hCG and PAPP-A with immediate MoM calculation for all post-IVF-ET patients at risk, followed by targeted transvaginal ultrasound with AI-based image analysis for those exhibiting the characteristic biochemical pattern. This streamlined triage optimizes resource use by prioritizing high-risk individuals. The principal clinical benefit of early and accurate diagnosis is the preservation of the viable intrauterine pregnancy alongside timely, minimally invasive management of the ectopic gestation, thereby reducing maternal morbidity and improving overall reproductive outcomes [24,25].

Methodologically, this study introduces several innovations. To our knowledge, it is the first to implement and validate a Vision Transformer-based model for ultrasound diagnosis of HP, capitalizing on its ability to model global image context [26]. Moreover, we have introduced and validated the "hCG/PAPP-A MoM" profile as a sensitive and specific biochemical discriminator. The successful integration of these advanced image analytics with a multiparametric serum panel into a high-performance diagnostic model represents a key achievement. The robustness of our conclusions is supported by a rigorous study design, including a well-defined retrospective case-control framework, use of an independent test set, and application of advanced statistical comparisons (e.g., DeLong's test) with appropriate correction for multiple testing.

Limitations

This study has several limitations. First, its retrospective, single-center design may introduce selection bias and limit generalizability. Because all HP cases were surgically confirmed, only clinically overt disease was captured, not subclinical or early-stage HP; thus, our findings apply only to clinically manifest disease. Clinical symptoms overlapped substantially between HP and EP (Table 2), so our proposed workflow does not rely on them for risk stratification. Second, the case-control design enriched HP prevalence to 33%, which greatly inflates PPV compared to the real-world setting (1–3%). To quantify this, we performed Bayesian recalibration: in real-world prevalence, NPV remains >99.9%, but PPV drops to 8.9–22.8% (Results 3.7). Hence, the model is best used as a triage screening tool, with confirmatory testing for positives. Third, although we used Bonferroni and DeLong's tests, Type I error inflation cannot be fully excluded. Moreover, the AI model was trained on a single device by two sonographers, introducing device- and operator-dependent biases; external validation across diverse settings is needed [27,28]. Finally, the test set was small (56 patients, only 19 HP events), yielding wide confidence intervals and violating the EPV criterion for logistic regression validation; thus, our findings are exploratory. DCA was performed on the same small test set, below the recommended sample size for stable net benefit

estimates; therefore, our DCA results should be interpreted as exploratory, and validation in larger cohorts is needed. Additionally, despite using pre-trained weights and regularization, the ViT model was fine-tuned on a modest dataset (~1352 images), and overfitting to single-center data cannot be fully excluded; multi-center external validation is essential. Future work should also aim to develop a user-friendly clinical decision support system for routine care [29,30].

5. Conclusions

In conclusion, this study demonstrates that an AI-enhanced ultrasound and serum biomarker (MoM-based) strategy significantly improves timely detection of clinically manifest HP after IVF-ET. By integrating anatomical and physiological insights, this multimodal approach outperforms conventional methods and represents a tangible step toward precision medicine in reproductive care, with clear potential to optimize clinical management and patient outcomes.

Availability of Data and Materials

All data associated with this study and the custom code developed for the AI model are available upon reasonable request from the corresponding author.

Author Contributions

YD conceived and designed the study. JS, JZ, and CW performed the data collection, including ultrasound examinations and serum biomarker assays. JS and JZ conducted the AI model development and statistical analysis. CW contributed to the clinical data interpretation and patient follow-up. YD provided overall supervision, administrative support, and critical revision of the manuscript for important intellectual content. The first draft of the manuscript was written by JS and JZ, and all authors commented on previous versions. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Ethics Approval and Consent to Participate

This study was carried out in accordance with the guidelines of the Declaration of Helsinki. Ethical oversight for this research was managed by the Medical Ethics Committee at Wuxi People's Hospital Affiliated to Nanjing Medical University (Approval No. 2025-6914773). Given the retrospective nature of the study, which involved the analysis of anonymized data from electronic medical records, the requirement for written informed consent was waived by the ethics committee. All patient data were anonymized prior to analysis to ensure confidentiality.

Acknowledgment

Not applicable.

Funding

This study was supported by the Wuxi Municipal Double Hundred Young and Middle-Aged Reserve Top Talents Program in Medical and Health Fields (Grant No. HB2023001) and the Jiangsu Provincial Health Commission Scientific Research Fund Project (Grant No. X202336).

Conflicts of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT-3.5 in order to check spelling and grammar. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/CEOG50265.

References

[1] Maleki A, Khalid N, Rajesh Patel C, El-Mahdi E. The rising incidence of heterotopic pregnancy: Current perspectives and associations with in-vitro fertilization. European Journal of Obstetrics, Gynecology, and Reproductive Biology. 2021; 266: 138–144. https://doi.org/10.1016/j.ejogrb.2021.09.031

[2] Nasrollahi F, Cao K, Hohbein J, Eilbert W. Spontaneously Conceived Ruptured Heterotopic Pregnancy Presenting with Chest Pain and Dyspnea: A Case Report. Clinical Practice and Cases in Emergency Medicine. 2024; 8: 64–67. https://doi.org/10.5811/cpcem.1671

[3] Utomo A, Aditiyono A, Marlina D, Disastra YP, Susilo AFP, Adinda Adriansyah PN. Diagnostic Challenges in Ectopic Pregnancy: Identifying Pseudo Sacs in Suspected Heterotopic Cases. The American Journal of Case Reports. 2025; 26: e944924. https://doi.org/10.12659/AJCR.944924

[4] Liu Z, Jing C. A Case Report of Retroperitoneal Ectopic Pregnancy After in vitro Fertilization-Embryo Transfer and Literature Review. International Journal of Women's Health. 2023; 15: 679–693. https://doi.org/10.2147/IJWH.S408319

[5] Li W, Liu Z, Ma Y, Zhao T, Liu Y, Qin L, et al. Diagnostic value of colour Doppler ultrasound, serum β-hCG concentration, and progesterone concentration in ectopic pregnancy. Technology and Health Care: Official Journal of the European Society for Engineering and Medicine. 2025; 33: 789–796. https://doi.org/10.1177/09287329241296379

[6] Aviram A, Mei-Dan E, Cassir G, Figueiro-Filho E, Liauw J, Jain V, et al. Technical Update No. 462: Abnormal maternal serum markers and adverse pregnancy outcomes. Journal of Obstetrics and Gynaecology Canada: JOGC = Journal D'obstetrique et Gynecologie du Canada: JOGC. 2025; 47: 103060. https://doi.org/10.1016/j.jogc.2025.103060

[7] Ceballos Medina A, Gómez-Acebo I, Gallego de Largy CC, Alonso-Molero J, Vilares Calvo S, Odriozola Feu JM, et al. Study of first-trimester serum levels of β-hCG and PAPP-A as a screening test for fetal development of intrauterine growth restriction. BMC Pregnancy and Childbirth. 2025; 25: 655. https://doi.org/10.1186/s12884-025-07787-7

[8] Ji X, Liu C, Hu J, Li S, Wang L, Cheng X, et al. A multicenter deep learning framework integrating radiomics and vision transformers for comprehensive ovarian tumor analysis from ultrasound imaging. European Journal of Medical Research. 2026; 31: 252. https://doi.org/10.1186/s40001-025-03395-5

[9] Yang J, Zhong X, Luo Z, Wen Y. Impact of differential privacy on breast ultrasound image classification performance using vision transformer. Quantitative Imaging in Medicine and Surgery. 2025; 15: 12279–12289. https://doi.org/10.21037/qims-2025-1064

[10] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology. 1996; 49: 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3

[11] Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. Journal of Clinical Epidemiology. 2016; 76: 175–182. https://doi.org/10.1016/j.jclinepi.2016.02.031

[12] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics in Medicine. 2019; 38: 1276–1296. https://doi.org/10.1002/sim.7992

[13] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv Preprint. 2020. https://doi.org/10.48550/arxiv.2010.11929 (preprint)

[14] Wang LL, Chen X, Ye DS, Liu YD, He YX, Guo W, et al. Misdiagnosis and delayed diagnosis for ectopic and heterotopic pregnancies after in vitro fertilization and embryo transfer. Journal of Huazhong University of Science and Technology. 2014; 34: 103–107. https://doi.org/10.1007/s11596-014-1239-7

[15] Vavrušová R, Hudeček R, Kuřecová B, Ventruba P. Diagnosticko-terapeutický algoritmus heterotopické gravidity. Praktická Gynekologie. 2007.

[16] Wang K, Deng XF, Zheng Z, Huang ZQ, Liao SQ, Liu SJ, et al. Development and validation of a multimodal radiomics-serum biomarker model for diagnosing solid pulmonary nodules via machine learning. Journal of Thoracic Disease. 2025; 17: 9721–9734. https://doi.org/10.21037/jtd-2025-1214

[17] Zeng X, Chen J, Zeng X, Tang X, Peng J. Integrating multiparametric MRI radiomics and clinical models to assess sensitivity to neoadjuvant chemotherapy in breast cancer: A multicenter study. Journal of Applied Clinical Medical Physics. 2025; 26: e70347. https://doi.org/10.1002/acm2.70347

[18] Kim JW, Khan AU, Banerjee I. Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis. Journal of Imaging Informatics in Medicine. 2025; 38: 3248–3262. https://doi.org/10.1007/s10278-024-01322-4

[19] Aburass S, Dorgham O, Al Shaqsi J, Abu Rumman M, Al-Kadi O. Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases. Journal of Imaging Informatics in Medicine. 2025; 38: 3928–3971. https://doi.org/10.1007/s10278-025-01481-y

[20] Hettikankanamage N, Shafiabady N, Chatteur F, Wu RMX, Ud Din F, Zhou J. eXplainable Artificial Intelligence (XAI): A Systematic Review for Unveiling the Black Box Models and Their Relevance to Biomedical Imaging and Sensing. Sensors (Basel, Switzerland). 2025; 25: 6649. https://doi.org/10.3390/s25216649

[21] Hoghooghi Esfahani H, Toyonaga S, Oyibo K. The application of explainable artificial intelligence in the prediction, diagnoses, treatment, and management of chronic diseases: A systematic review. Digital Health. 2025; 11: 20552076251355669. https://doi.org/10.1177/20552076251355669

[22] Fischer AC, Georges HM, Jones DM, Chamley LW, Abrahams VM. Nicotinamide Riboside Mitigates Antiphospholipid Antibody-Induced Dysfunction in Human Trophoblast Cells. American Journal of Reproductive Immunology (New York, N.Y.: 1989). 2026; 95: e70205. https://doi.org/10.1111/aji.70205

[23] Wang Q, Feng W, Tan Y, Qiao J, Liu H, Liu Q, et al. Antiphospholipid antibodies inhibit the migration and invasion of trophoblast cells by suppressing the JNK/C-Jun/MMP1 signaling pathway. Journal of Translational Medicine. 2025; 23: 581. https://doi.org/10.1186/s12967-025-06596-y

[24] Li S, Cao M, Liu H, He Y, Liu J. Management of 14 patients with cornual heterotopic pregnancy following embryo transfer: experience from the past decade. Reproductive Biology and Endocrinology: RB&E. 2021; 19: 152. https://doi.org/10.1186/s12958-021-00834-w

[25] Jiang Y, Chen J, Zhou H, Zheng M, Han K, Ling J, et al. Management and obstetric outcomes of 17 heterotopic interstitial pregnancies. BMC Pregnancy and Childbirth. 2018; 18: 78. https://doi.org/10.1186/s12884-018-1700-x

[26] Tehsin S, Alshaya H, Bouchelligua W, Nasir IM. Hybrid State-Space and Vision Transformer Framework for Fetal Ultrasound Plane Classification in Prenatal Diagnostics. Diagnostics (Basel, Switzerland). 2025; 15: 2879. https://doi.org/10.3390/diagnostics15222879

[27] Qu Y, Lian J, Liu T, Xu Y, Wang Z, Weng F, et al. Binary and Ternary Classification Prediction for Breast Cancer and Breast Sclerosing Adenosis With Interpretable Artificial Intelligence From Clinical and Imaging Features: A Retrospective, Diagnostic Accuracy Cohort Study. Cancer Innovation. 2026; 5: e70049. https://doi.org/10.1002/cai2.70049

[28] Xue Z, Feng Y, Liu N. Hybrid receptive field U-Net for lesion segmentation in breast ultrasound images. Medical Physics. 2026; 53: e70370. https://doi.org/10.1002/mp.70370

[29] Neff MC, Schaaf J, Ahmadi N, Zoch M, Schütze D, Storf H. Experiences With Integrating Medical Terminologies Into User Interfaces for a Decision Support System for Primary Care: Conceptual and Development Study. JMIR Medical Informatics. 2026; 14: e74934. https://doi.org/10.2196/74934

[30] Peek N, Capurro D, Rozova V, van der Veer SN. Bridging the Gap: Challenges and Strategies for the Implementation of Artificial Intelligence-based Clinical Decision Support Systems in Clinical Practice. Yearbook of Medical Informatics. 2024; 33: 103–114. https://doi.org/10.1055/s-0044-1800729