Clinical Investigation / Araştırma

DOI: 10.4274/tjod.galenos.2026.10663

Turk J Obstet Gynecol

Machine learning prediction of IVF outcomes in unexplained infertility

Açıklanamayan infertilite tedavisinde IVF sonuçlarının makine öğrenimi ile tahmini

Yeliz Kaya[1], Yunus Aydın[2], Coşkun Kaya[3], Tuğba Tahta[4], Özer Çelik[5], Şermin Özkaya Karabıçak[2]

[1]Eskişehir Osmangazi University Faculty of Health Sciences, Department of Obstetrics and Gynecology Nursing, Eskişehir, Türkiye
[2]Denizli Özel Sağlık Hospital IVF Center, Denizli, Türkiye
[3]University of Health Sciences Türkiye, Eskişehir City Hospital, Department of Urology, Eskişehir, Türkiye
[4]Ankara Medipol University Faculty of Health Sciences, Department of Midwifery, Ankara, Türkiye
[5]Vocational School of Information Technologies, Anadolu University, Eskişehir, Türkiye

Abstract

Objective: To develop and evaluate machine learning (ML) models to predict clinical pregnancy (CP) outcomes in couples with unexplained infertility (UI) undergoing in vitro fertilization (IVF), thereby enhancing early treatment decision-making and patient counseling.

Materials and Methods: A retrospective analysis was conducted on 145 couples diagnosed with UI and treated at a single-center IVF clinic. Live birth data were not available; therefore, CP was used as the primary outcome. Demographic, clinical, and laboratory parameters were collected, and five ML algorithms—light gradient boosting machine (LGBM), random forest, logistic regression, extra trees classifier, and eXtreme gradient boosting—were applied to predict CP outcomes. Model performance was assessed using the area under the receiver operating characteristic curve (ROC-AUC), accuracy, sensitivity, and specificity. Permutation-based feature importance analysis was performed to identify key predictive variables.

Results: Among the evaluated models, the LGBM model achieved the highest test-set ROC-AUC (0.712) and an accuracy of 65.2%. Given the small test sample and the lack of external validation, these findings should be interpreted as exploratory. Variables receiving the highest permutation feature-importance scores included total gonadotropin dose, initial gonadotropin dose, total sperm count, progressive motile sperm count, and infertility duration. These findings should be interpreted as exploratory and do not establish that these variables are independent predictors of CP.

Conclusion: This study presents the first ML-based approach tailored to predicting CP outcomes in couples with UI undergoing IVF. The results highlight the potential of artificial intelligence to inform personalized reproductive care. These exploratory findings demonstrate the feasibility of applying ML methods in UI. Larger prospective multicenter studies are required to determine whether clinically useful prediction models can be developed.

Keywords: Artificial intelligence, IVF, machine learning, predictive model, pregnancy outcomes, unexplained infertility

Öz

Amaç: Açıklanamayan infertilitesi (UI) olan ve in vitro fertilizasyon (IVF) uygulanan çiftlerde klinik gebelik (KG) sonuçlarını öngörmek amacıyla makine öğrenmesi (ML) modelleri geliştirmek ve değerlendirmek; erken tedavi karar verme ve hasta danışmanlığını iyileştirmek.

Gereç ve Yöntemler: Tek merkezli bir IVF kliniğinde tedavi gören ve UI tanısı alan 145 çift üzerinde retrospektif analiz yapılmıştır. Canlı doğum verileri mevcut olmadığından, birincil sonuç olarak KG kullanılmıştır. Demografik, klinik ve laboratuvar parametreleri toplanmış; KG sonuçlarını tahmin etmek için beş ML algoritması—*light gradient boosting machine* (LGBM), random forest, lojistik regresyon, *extra trees classifier* ve *eXtreme gradient boosting*—

PRECIS: Machine learning models moderately predicted clinical pregnancy outcomes in unexplained infertility in vitro fertilization patients, with light gradient boosting machine showing the best performance and identifying key clinical and laboratory predictors.

Corresponding Author/Sorumlu Yazar: Tuğba Tahta PhD,
Ankara Medipol University Faculty of Health Sciences, Department of Midwifery, Ankara, Türkiye
E-mail: tahtatugba@gmail.com **ORCID ID:** orcid.org/0000-0003-0190-977X

Received/Geliş Tarihi: 30.04.2026 **Accepted/Kabul Tarihi:** 20.10.2026 **Epub:** 02.11.2026

Cite this article as: Kaya Y, Aydın Y, Kaya C, Tahta T, Çelik Ö, Özkaya Karabıçak Ş. Machine learning prediction of IVF outcomes in unexplained infertility. Turk J Obstet Gynecol. [Epub Ahead of Print]

Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of Turkish Society of Obstetrics and Gynecology.
This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

uygulanmıştır. Model performansı, alıcı işletim karakteristiği eğrisi altında kalan alan (ROC-AUC), doğruluk, duyarlılık ve özgüllük ile değerlendirilmiştir. Önemli öngörücü değişkenleri belirlemek için permütasyon temelli özellik önem analizi yapılmıştır.

Bulgular: Değerlendirilen modeller arasında LGBM, test kümesinde en yüksek gözlemlenen ROC-AUC değerine (0,712) ve %65,2 doğruluğa ulaştı. Küçük test örneklemi ve dış doğrulama bulunmaması nedeniyle bu bulgular keşifsel olarak yorumlanmalıdır. En yüksek permütasyon özellik önem puanlarını alan değişkenler arasında toplam gonadotropin dozu, başlangıç gonadotropin dozu, toplam sperm sayısı, ilerleyici hareketli sperm sayısı ve kısırlık süresi yer almaktadır. Bu bulgular keşifsel olarak yorumlanmalı ve bu değişkenleri klinik gebeliğin bağımsız öngörücüleri olarak belirlememektedir.

Sonuç: Bu çalışma, UI'sı olan ve IVF uygulanan çiftlerde KG sonuçlarını tahmin etmeye yönelik ilk ML temelli yaklaşımı sunmaktadır. Bulgular, yapay zekanın kişiselleştirilmiş üreme sağlığı bakımına katkı potansiyelini göstermektedir. Bu ön bulgular, açıklanamayan kısırlıkta makine öğrenme yöntemlerinin uygulanabilirliğini göstermektedir. Klinik olarak faydalı tahmin modellerinin geliştirilip geliştirilemeyeceğini belirlemek için daha büyük, prospektif, çok merkezli çalışmalara ihtiyaç duyulmaktadır.

Anahtar Kelimeler: Yapay zeka, IVF, makine öğrenmesi, öngörü modeli, gebelik sonuçları, açıklanamayan infertilite

Introduction

Unexplained infertility (UI) is diagnosed in couples who fail to conceive despite engaging in regular, unprotected intercourse for at least 12 months, and in whom standard infertility evaluations reveal no identifiable abnormalities in either partner[1]. Among couples with UI, treatment options vary widely and may include expectant management, intrauterine insemination (IUI) with ovarian stimulation (OS + IUI), or in vitro fertilization (IVF), and depends on factors such as age, duration of infertility, and evaluation outcomes[2]. However, significant uncertainty remains regarding the optimal treatment strategy for UI couples as no single approach has been universally recognized as the first-line treatment option[2-5]. IVF is not always the first-line treatment for UI and is typically considered after failure of expectant management or IUI.

Recent reviews have highlighted the challenges in comparing the success rates of assisted reproductive technologies (ART) for UI. These challenges arise from several factors, including the limited availability of comparative studies[3,4], inconsistencies in the definition of UI across the literature[6], and variability in outcome measures, such as live birth rates, clinical pregnancy (CP) rates, and ongoing pregnancy rates[7]. Current challenges in UI treatment include a lack of reliable predictors and the overuse of invasive and costly treatments, such as IVF, in patients with uncertain prognoses. This lack of standardization emphasizes the urgent need for robust and unified criteria to evaluate the efficacy of ART in UI. Predictive models could assist clinicians in personalizing treatment strategies, improving success rates, and alleviating the physical, emotional, and financial burden on couples; studies in this area have already begun. While several statistical models have been developed to predict ART success rates[8,9], these models frequently rely on linear assumptions and often fail to capture the complex, non-linear interactions inherent in biological processes such as fertilization and embryo development. Furthermore, many of these models are limited by small sample sizes and heterogeneous patient populations, which reduce their generalizability and practical applicability.

In recent years, artificial intelligence (AI) has emerged as a transformative tool across various scientific disciplines, including reproductive medicine. One of the primary advantages of AI, particularly machine learning (ML) algorithms, is their ability to analyze high-dimensional data and uncover hidden patterns that traditional methods may overlook. AI applications in infertility management primarily focus on predicting ART outcomes, selecting optimal embryos for transfer, and classifying sperm quality[10,11]. Despite its growing adoption, most AI-driven studies in infertility have not accounted for the type of infertility in their analyses. To date, only one study has specifically focused on couples with UI. In this study, Bachelot et al.[12] utilized an ML model to classify fertile couples and couples with idiopathic infertility, providing a preliminary exploration of the potential of AI in this domain. However, this study did not address the prediction of treatment success, leaving a critical gap in the application of ML models for forecasting ART outcomes.

Materials and Methods

Data Source and Study Workflow

This retrospective study was conducted with ethical approval. Ethical approval was obtained from Ankara Medipol University Non-Interventional Clinical Research Ethics Committee (approval no: 15, date: 22.01.2024), all participants provided consent for publication of their data in an anonymized form. The medical records of infertile couples who underwent IVF at Denizli Özel Sağlık Hospital's IVF Center (Denizli, Türkiye) between January 2022 and January 2023 were reviewed. Given the study's aim and design, only couples with UI were included. UI was defined as the inability to conceive after 12 months of regular, unprotected intercourse despite normal findings in standard infertility evaluations, including ovulatory function, tubal patency (assessed via hysterosalpingography or laparoscopy), uterine anatomy (evaluated by ultrasound or hysteroscopy), and normal semen analysis as defined by the 6th edition of the World Health Organization's (WHO) criteria[13]. Couples who underwent IVF treatment with a standard OS protocol, defined as controlled OS using exogenous gonadotropins (recombinant or urinary follicle-stimulating hormone ± luteinizing hormone), in combination with either a gonadotropin releasing hormone (GnRH) antagonist or long GnRH agonist protocol for pituitary suppression, along with serial transvaginal ultrasonographic and/or hormonal monitoring, were included in the study. Couples classified as having UI were evaluated based on the inclusion and exclusion criteria outlined below to ensure a homogeneous study population.

Inclusion Criteria

• Couples diagnosed with UI, defined as the inability to conceive after 12 months of regular, unprotected intercourse despite normal findings in standard infertility evaluations (e.g., normal ovulation, patent fallopian tubes, normal uterine anatomy, and semen analysis meeting WHO criteria).
• Women aged ≤35 years at the time of IVF treatment.
• Couples who previously underwent expectant management or IUI without achieving pregnancy.
• Normal ovarian reserve, determined by serum anti-Mullerian hormone (AMH) levels and antral follicle count (AFC).
• Couples who underwent IVF treatment with standard OS protocols, including GnRH antagonist or long GnRH agonist protocols.
• Couples undergoing their first IVF cycle in which a fresh embryo transfer was performed on day 5.

Exclusion Criteria

• Women with a history of endometriosis (Stage III or IV), polycystic ovary syndrome, or diminished ovarian reserve (e.g., AMH <1.1 ng/mL or AFC <5).
• Women with uterine abnormalities (e.g., septum, submucosal fibroids, endometrial polyps) that could affect implantation or pregnancy outcomes.
• Women with reproductive tract infections or chronic endometritis
• Men with abnormal semen analysis results (e.g., oligozoospermia, asthenozoospermia, teratozoospermia, or azoospermia).
• Couples with systemic diseases, significant lifestyle factors [e.g., smoking, excessive alcohol consumption, or body mass index (BMI) >35 kg/m^2], or genetic abnormalities affecting fertility.
• Couples who underwent non-standard IVF protocols, such as donor oocyte cycles or frozen embryo transfers.

Data Collection

The data of the couples included in the study were collected using standardized forms prepared by the researchers. The following key variables were recorded and categorized for the study:

1. General couple data:
• Duration of infertility (years),
• Presence of previous OS + IUI treatment (yes/no).
2. Female data:
• Age (years),
• BMI, kg/m^2,
• AMH level (ng/mL),
• AFC 3rd day.
3. Male data:
• Age (years),
• Basal sperm count (million/mL),
• Basal sperm progressive motility rate (%),
• Sperm morphology (% normal forms).
4. IVF cycle features:
• Total gonadotropin dose (IU),
• Cycle type (agonist/antagonist protocol),
• Day 5 estradiol level (pg/mL).
5. End-of-cycle characteristics:
• Number of oocytes collected,
• Number of Day 5 blastocysts,
• Implantation rate.

Data Preprocessing

A positive beta human chorionic gonadotropin level (β-hCG), indicating successful implantation, was defined as a serum β-hCG level >25 mIU/mL measured approximately 14-17 days after oocyte retrieval. CP was selected as the primary outcome for the model because live birth data were not available for all couples. CP was defined as the presence of an intrauterine gestational sac with fetal cardiac activity confirmed by ultrasound by the 7th week of gestation. The couples were categorized into two groups: those without CP after the IVF cycle (Group 1) and those with CP (Group 2).

Statistical Analysis

As this study was exploratory and focused on developing and evaluating ML models for predictive performance, no formal power analysis was performed. The sample size was determined based on data availability within the retrospective cohort. The Shapiro-Wilk test was used to assess the normality of variables included in the analysis. Continuous variables were expressed as mean ± standard deviation, and categorical variables as frequency (percentage), and median (minimum-maximum). Comparisons between groups of continuous variables that were normally distributed were performed using the independent samples t-test, whereas comparisons between groups of continuous variables that were not normally distributed were performed using the Mann-Whitney U test. The relationship between categorical variables across the groups was assessed using the chi-square test. In all analyses, a two-sided p-value <0.05 was considered statistically significant.

ML Models

In this study, several ML algorithms were employed to predict CP outcomes in couples with UI undergoing IVF treatment. The following models were employed: random forest classifier, logistic regression, light gradient boosting machine (LGBM classifier), extra trees classifier, and eXtreme gradient boosting (XGBoost). These algorithms were chosen based on their ability to handle complex nonlinear relationships between variables, their effectiveness in dealing with missing or imbalanced data, and their robustness in providing accurate predictions.

The random forest algorithm is an ensemble learning method that builds multiple decision trees during training and merges their outputs to improve prediction accuracy. It is particularly effective for handling large datasets with high dimensionality and can manage both regression and classification tasks. The random forest also provides insights into feature importance,

making it useful for understanding which variables contribute most to predictions.

Logistic regression is a statistical model used for binary classification. It uses a logistic function to predict the probability of an outcome based on one or more independent variables. Although simpler than other algorithms, Logistic Regression is interpretable and serves as a baseline model against which more complex methods can be compared.

LGBM is a highly efficient gradient- boosting framework designed to handle large datasets with low computational cost. It employs histogram-based decision-tree learning to optimize both speed and accuracy. LGBM is well-suited for handling categorical and numerical variables, and its ability to focus on relevant data points makes it particularly powerful for predictive tasks.

Extra trees, also referred to as extremely randomized trees, is another ensemble learning technique similar to random forest. However, it introduces randomness by selecting split points during tree construction, which can improve generalization. Extra trees is computationally efficient and less prone to overfitting compared to traditional decision trees.

XGB is an optimized implementation of gradient boosting, designed to improve computational speed and model performance. It employs regularization techniques to reduce overfitting and efficiently handles large, sparse datasets. XGB is widely used in ML approaches and recognized for its superior predictive accuracy.

Performance Metrics

The dataset was randomly split into a training set (70%) and a testing set (30%) to ensure that both subsets were representative of the overall population. Random splitting was performed to avoid introducing any bias in the allocation of data points between the training and testing sets and to ensure that the training set captures the diversity of the dataset for model development, while the testing set provides an unbiased evaluation of the model's performance. Although an initial 70/30 split was performed, the final test set for performance evaluation consisted of 23 cases after preprocessing. This reduction should be considered when interpreting model performance, as it may increase variability and limit generalizability. The random splitting process was repeated across different runs to verify the stability and consistency of the results. The accuracy rate, area under the curve (AUC) from receiver operating characteristic (ROC) curve analysis, and confusion matrix metrics were selected to identify the optimal ML algorithm.

A 5-fold cross-validation approach was employed to evaluate the performance of the ML models. The complete dataset was randomly divided into five equal parts, and in each iteration, one part was used as the validation set while the remaining four parts were used for training. This method ensured that every data point was used just once for validation, providing a comprehensive evaluation of the models' predictive capabilities. The 5-fold cross-validation was conducted on the training data to ensure a robust evaluation, while the performance on the test set was used to assess the models' generalizability to unseen data.

Feature Selection

Out of the 23 variables listed in Table 1, 18 that affect CP were selected using the permutation feature importance method applied to the complete dataset. This approach enabled a comprehensive evaluation of all available data to identify the most influential factors associated with CP.

The Permutation Feature Importance technique evaluates the importance of each feature individually by measuring the change in prediction quality, quantified as the decrease in the coefficient of determination score, after permuting a single feature. The rate of reduction of the coefficient of determination indicates the importance of the feature. The permutation feature importance plot is shown in Figure 1.

Statistical analyses and ML algorithms were carried out using Python software (Python Software Foundation. Python Language Reference, version 3.5. Available at http://www.python.org) and the Scikit-Learn library. Feature selection was performed using permutation importance on the complete dataset. However, this approach may introduce information leakage, as feature selection ideally should be restricted to the training data or performed within cross-validation folds.

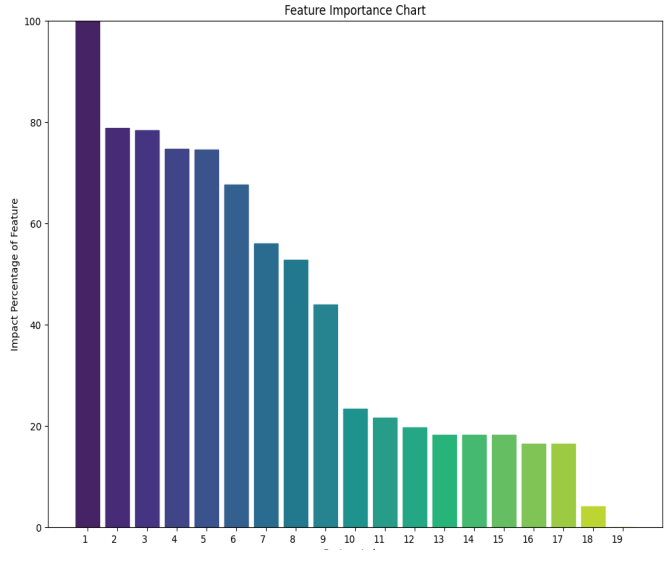


Figure 1. The permutation feature importance plot. 1: Total gonadotropin dose; 2: Starting gonadotropin dose; 3: Total sperm count; 4: Total progressive motile sperm count; 5: Duration of infertility 6: Smoking (male) 7: Age (male), 8: Antral follicle count per cycle, 9: Total gonadotropin usage time; 10: Smoking (female); 11: Number of antagonist cycles; 12: History of abdominal surgery; 13: Alcohol (male); 14: Age (female); 15: Alcohol (female); 16: History of varicocelectomy; 17: The history of intrauterine insemination with ovarian stimulation; 18: Normal sperm morphology rate

Results

The study analyzed data from 145 couples who were divided into two groups: Group 1 (no CP, 40.7%) and Group 2 (CP achieved, 59.3%). No significant differences in women's ages (p=0.811), BMI (p=0.071), or lifestyle factors such as smoking and alcohol consumption were observed between the groups (Table 1). Similarly, men's ages, total sperm count, and total progressive motile sperm count (TPMSC) were comparable across both groups. Additionally, duration of infertility, gonadotropin use, and estradiol levels during treatment did not differ significantly between the groups. These findings indicate that demographic, clinical, and laboratory characteristics were evenly distributed, resulting in a homogeneous study population. Eighteen of the 23 parameters listed in Table 1 were identified as important features, as shown in Figure 1. The following variables were included in the model: duration of infertility, history of IUI with OS, total gonadotropin dose, starting gonadotropin dose, AFC per cycle, total gonadotropin usage time, number of antagonist cycles, history of abdominal surgery, age (female), smoking (female), alcohol consumption (female), total sperm count, TPMSC, normal sperm morphology rate, alcohol consumption (male), smoking (male), age (male), and history of varicocelectomy. Among the evaluated variables, total gonadotropin dose, starting gonadotropin dose, total sperm count, TPMSC, and duration of infertility had the highest

Table 1. The demographic and laboratory findings of the patients included in the study

	Parameters		Group 1 59 (40.7%)	Group 2 86 (59.3%)	p
Women	Age [median (Q1-Q3)]		31.50 (24-34)	31.00 (24-35)	0.811[b]
	BMI (mean ± SD)		26.04±5.57	23.16± 3.78	0.071[a]
	Smoking (n, %)	+	10 (17)	14 (16)	0.915[c]
		-	49 (83)	72 (84)	
	Alcohol (n, %)	+	4 (7)	8 (9)	0.588[c]
		-	55 (93)	78 (91)	
	History of abdominal surgery (n, %)	+	16 (27)	23 (27)	0.960[c]
		-	43 (73)	63 (73)	
	The history of OS + IUI (n, %)	+	45 (76)	73 (85)	0.191[c]
		-	14 (24)	13 (15)	
	AMH [median (Q1-Q3)]		2.07 (1.62-3.72)	1.14 (0.74-3.21)	0.118[b]
	AFC per cycle [median (Q1-Q3)]		5.50 (4-19)	9.00 (3-14)	0.103[b]
Male	Age (mean ± SD)		34.10±4.77	34.69±4.73	0.203[a]
	Smoking (n, %)	+	22 (37.2)	37 (40.3)	0.641[c]
		-	37 (62.8)	49 (59.7)	
	Alcohol (n, %)	+	10 (17)	15 (18)	0.938[c]
		-	49 (83)	71 (82)	
	History of varicocelectomy (n, %)	+	7 (12)	10 (11)	0.965[c]
		-	52 (88)	76 (89)	
	Post-wash sperm concentration ($\times 10^6$/mL), median (Q1-Q3)		5.75 (5-8)	6.5 (5-18)	0.442[b]
	TPMSC (a + b) (10^6) [median (Q1-Q3)]		1.04 (0.8-1)	1.03 (0.80-1.20)	0.067[b]
	Normal sperm morphology rate (mean ± SD)		0.02±0.01	0.03±0.01	0.781[a]
IVF cycle features	Duration of infertility (month) [median (Q1-Q3)]		30 (12-120)	24 (12-96)	0.165[b]
	Initial gonadotropin dose [median (Q1-Q3)]		300 (225-375)	300 (187.50-300)	0.166[b]
	Total gonadotropin usage time (days)		11.8±1.75	11.23±1.73	0.576[a]
	Total gonadotropin dose		3150±474	2759.6±396.8	0.574[a]
	Number of antagonist injections per cycle [median (Q1-Q3)]		2 (1-2)	1 (1-2)	0.189[b]
	Number of agonist injections per cycle [median (Q1-Q3)]		1 (1-2)	2 (1-2)	0.188[b]
	Estradiol level at 5th day [median (Q1-Q3)]		336.50 (94-888)	310 (85-1150)	0.328[b]
	Estradiol level at β-hCG administration day		2283±989.90	2569±990.29	0.755[a]

[a]: Student's t-test, [b]: Mann-Whitney U, [c]: Chi-square, BMI: Body mass index, OS + IUI: Intrauterine insemination with ovarian stimulation, AMH: Anti-Mullerian hormone, AFC: Antral follicle count, TPMSC: Total progressive motile sperm count, IVF: In vitro fertilization, β-hCG: Beta human chorionic gonadotropin level, SD: Standard deviation

permutation feature importance scores. Given the exploratory nature of the study and the lack of statistically significant between-group differences, these variables should not be interpreted as independent determinants of CP.

The performance of the ML models in predicting CP outcomes is presented in Tables 2 and 3. Among the evaluated models, the LGBM classifier achieved the highest observed test-set ROC-AUC value (0.712), with an accuracy of 65.2%, a sensitivity of 72.7%, and a specificity of 58.3%. Logistic regression, random forest, extra trees classifier, and XGB achieved ROC-AUC values of 0.424, 0.311, 0.375, and 0.182, respectively (Figure 2).

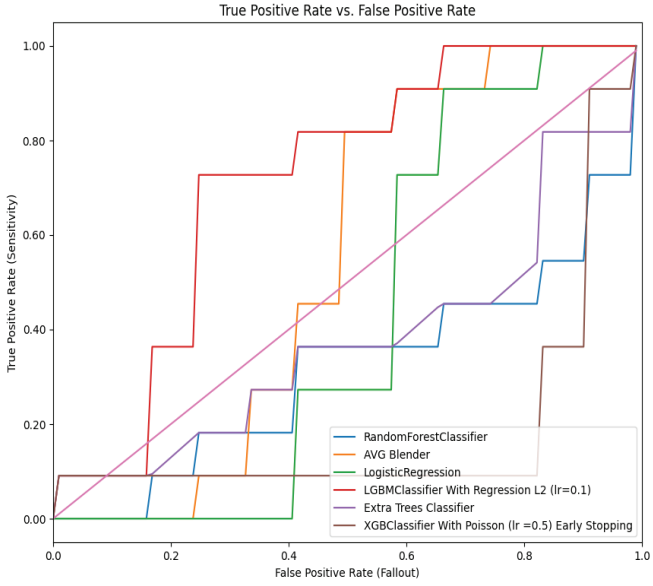


Figure 2. Receiver operating characteristic curve graphs

LGBM: Light gradient boosting machine, XGB: eXtreme gradient boosting

Discussion

Main Findings

In the literature, statistical methods have been used to predict outcomes in ART, one of the most recent studies, published by Li et al.[14], was conducted to predict the probability of total fertilization failure in conventional IVF cycles. Using multivariate logistic regression, they identified 13 key predictors, including female age, BMI, infertility duration, and sperm parameters (concentration, motility, and morphology). The final model demonstrated a strong predictive performance, with an AUC of 0.815. While Li et al.'s[14] study represents an important step in predicting ART outcomes, it highlights several limitations associated with traditional statistical methods. Logistic regression models, which rely on linear assumptions, may oversimplify the complex, non-linear relationships inherent in biological processes such as fertilization. van Eekelen et al.[15] examined the effectiveness of IUI-OS compared to expectant management in couples with UI, demonstrating improved pregnancy rates with IUI-OS, particularly in cases with a poor prognosis. However, the study relied on traditional statistical methods, which lack the ability to model complex, non-linear interactions and to incorporate diverse clinical parameters. These limitations emphasize the need for advanced methodologies, such as ML algorithms, which can integrate diverse data types, model complex relationships, and provide more precise and generalizable predictions.

Recognizing the limitations of traditional statistical methods, researchers have embraced AI tools to enhance clinical decision-making, optimize treatment protocols, and improve outcomes throughout the ART process. AI technologies have been increasingly integrated into reproductive medicine since the 2010s, revolutionizing various stages of ART. These tools have been applied to enhance clinical decision-making,

Table 2. Prognosis prediction results of different machine learning algorithms

	Model training results			
	Test		Cross-validation	
Model name	**ROC-AUC**	**Accuracy**	**ROC-AUC**	**Accuracy**
Random forest classifier	0.311	0.435	0.653	0.6
Logistic regression	0.424	0.478	0.479	0.511
LGBM classifier	0.712	0.652	0.648	0.622
Extra trees classifier	0.375	0.391	0.666	0.667
XGB classifier	0.182	0.522	0.647	0.6

LGBM: Light gradient boosting machine, XGB: eXtreme gradient boosting, ROC: Receiver operating characteristic, AUC: Area under the curve

Table 3. Classifier confusion matrix

	LGBM classifier		
Clinical pregnancy	**Yes**	**No**	**%**
Yes	8	3	72.7[a]
No	5	7	58.3[b]

[a]: Sensitivity, [b]: Specifity, LGBM: Light gradient boosting machine

optimize treatment protocols, and improve outcomes at every step of the process, including sperm and embryo selection for transfer[16-18]. As AI applications gained traction, questions arose regarding their superiority over traditional statistical methods in predicting ART outcomes, prompting comparative studies to evaluate the performance of ML algorithms relative to conventional statistical approaches. One such study assessed the efficacy of ML techniques, including support vector machines and artificial neural networks, compared to logistic regression to predict various IVF outcomes[19]. While logistic regression provided baseline predictions based on linear relationships between variables, ML algorithms demonstrated their capacity to model complex, non-linear interactions and to incorporate a broader range of features. The study highlighted that ML methods excel in handling high-dimensional datasets, allowing for the integration of diverse clinical parameters, including patient demographics, hormonal levels, and treatment details[19].

Although previous studies have applied ML approaches to classify UI, identify diagnostic biomarkers, or characterize molecular signatures associated with the disease, few studies specifically focus on predicting IVF CP outcomes in a homogeneous population with UI.

Addressing this gap, we evaluated ML algorithms to predict CP after IVF in UI couples using a multidimensional dataset of clinical, demographic, and laboratory variables. Multiple ML models were evaluated. Although the LGBM model achieved the highest observed test-set ROC-AUC, cross-validation ROC-AUC values were similar among the tree-based models. Therefore, the findings should be considered exploratory and hypothesis-generating rather than evidence of robust predictive performance. Variables with higher feature-importance scores included total and initial gonadotropin doses, total sperm count, progressive motility, and infertility duration, underscoring the value of integrating diverse variables.

Recent studies have further highlighted the growing role of ML in UI. Bachelot et al.[12] developed a couple-based ML model capable of distinguishing fertile and infertile couples with idiopathic infertility, achieving an overall classification accuracy of 73.8% after feature reduction. Their findings emphasized the contribution of anthropometric, metabolic, and oxidative stress-related variables and demonstrated the feasibility of ML approaches in this challenging patient population. Similarly, Depciuch et al.[20] combined Raman spectroscopy with multiple ML algorithms to identify biochemical alterations in follicular fluid from women with UI. They used several classification models, including support vector machines, random forests, and XGBoost, and reported high classification accuracies, demonstrating that ML can detect subtle molecular signatures associated with UI. More recently, Ma et al.[21] applied multiple ML algorithms together with bioinformatics analyses to identify immune-related diagnostic biomarkers in UI. Their study identified six candidate biomarkers and demonstrated strong diagnostic performance across independent datasets, suggesting that ML may contribute not only to outcome prediction but also to elucidating the underlying biological mechanisms of UI.

The study exclusively included true UI couples (normal ovarian reserve, normal sperm parameters, no identifiable cause) to ensure a homogeneous and targeted analysis, though this limited variability in some predictors. The model was designed to predict CP using pre-embryo data to support early clinical decision-making. Therefore, embryo-related variables (quality, number transferred, transfer day) were excluded.

Study Limitations

This study has several limitations. Its retrospective, single-center design and small UI sample limit the generalizability; no power calculation was performed due to its exploratory nature. Embryo-related and laboratory variables (e.g., number and quality of embryos) were not included, which may have reduced accuracy. CP was used as the outcome, although an ongoing pregnancy or a live birth would be more clinically meaningful. AI models are also prone to sampling, historical, label, and algorithmic biases, and testing multiple models on a small dataset increases the risk of overfitting despite cross-validation. External validation with larger, multicenter datasets is needed.

Strengths include well-defined criteria ensuring a homogeneous, true UI population, focusing on couples who failed expectant management or IUI and who were appropriate for IVF. Exclusion of confounders (e.g., diminished ovarian reserve, endometriosis, severe male infertility) improved validity. Unlike prior heterogeneous studies, this work specifically applies ML to predict IVF outcomes in UI, addresses a key gap, and uses a comprehensive dataset that reflects the value of ML in clinical decision-making.

Some models (random forest, extra trees, XGBoost) showed an AUC <0.5 on the test set, suggesting prediction inversion due to a small sample size, overfitting, or poor generalization. Discrepancies between cross-validation and test results indicate optimism bias or data leakage. Additionally, feature selection on the full dataset may have caused data leakage; it should be limited to the training data or embedded within cross-validation. Findings should be interpreted cautiously.

Interpretation

Building on these findings, future research should prioritize external validation of ML models using large, multicenter datasets to improve generalizability and reduce single-center bias. Expanding input variables—such as embryo quality, genetic markers, and lifestyle factors—along with advanced imaging and laboratory data may enhance predictive accuracy. Future studies should also focus on clinically meaningful long-term outcomes, including live birth, maternal health, and neonatal well-being, to better align models with patient-centered care.

In addition, ML applications could extend beyond IVF to support decisions on treatment selection and timing (e.g., IUI

or expectant management). Finally, translating ML models into user-friendly, clinically integrated tools and evaluating them through prospective real-world studies will be essential to ensure effective adoption and impact.

Conclusion

This study represents the first application of ML algorithms specifically aimed at predicting IVF outcomes in couples with UI, addressing a critical gap in the existing literature. By demonstrating the potential of ML to integrate diverse clinical parameters and provide nuanced predictions, this research highlights the transformative role of AI in ART. This study represents an exploratory application of ML methods to predict CP following IVF in couples with UI. Therefore, the findings should be considered hypothesis-generating. Larger multicenter studies with external validation are required before any clinical application can be considered. This important framework not only contributes to the current understanding of UI but also serves as a foundation for future studies, inspiring broader adoption of AI technologies across the field of reproductive medicine.

Ethics

Ethics Committee Approval: Ethical approval was obtained from Ankara Medipol University Non-Interventional Clinical Research Ethics Committee (approval no: 15, date: 22.01.2024),
Informed Consent: All participants provided consent for publication of their data in an anonymized form.

Acknowledgments

We thank Arzu Altun Yavuz at the Department of Statistics, Eskişehir Osmangazi University Faculty of Science, for the statistical analysis.

Footnotes

Authorship Contributions

Concept: Y.K., Y.A., C.K., Design: Y.K., Y.A., C.K., T.T., Ö.Ç., Ş.Ö.K., Data Collection or Processing: T.T., Analysis or Interpretation: Ö.Ç., Literature Search: T.T., Ş.Ö.K., Writing: Y.K., Y.A., C.K., T.T., Ö.Ç., Ş.Ö.K.
Conflict of Interest: No conflict of interest was declared by the authors.
Financial Disclosure: The authors declared that this study received no financial support.

References

1. National Institute for Health and Care Excellence. Fertility: assessment and treatment for people with fertility problems. London: National Institute for Health and Care Excellence; 2013. Available from: https://www.nice.org.uk/guidance/ng257
2. Romualdi D, Ata B, Bhattacharya S, Bosch E, Costello M, Gersak K, et al.; Guideline Group on Unexplained Infertility. Evidence-based guideline: unexplained infertility†. Hum Reprod. 2023;38:1881-90.
3. Ayeleke RO, Asseler JD, Cohlen BJ, Veltman-Verhulst SM. Intra-uterine insemination for unexplained subfertility. Cochrane Database Syst Rev. 2020;3:CD001838.
4. Sunkara SK, Kamath MS, Pandian Z, Gibreel A, Bhattacharya S. In vitro fertilisation for unexplained subfertility. Cochrane Database Syst Rev. 2023;9:CD003357.
5. Man JK, Parker AE, Broughton S, Ikhlaq H, Das M. Should IUI replace IVF as first-line treatment for unexplained infertility? A literature review. BMC Womens Health. 2023;23:557.
6. Raperport C, Desai J, Qureshi D, Rustin E, Balaji A, Chronopoulou E, et al. The definition of unexplained infertility: a systematic review. BJOG. 2024;131:880-97.
7. Nandi A, Raja G, White D, Tarek ET. Intrauterine insemination + controlled ovarian hyperstimulation versus in vitro fertilisation in unexplained infertility: a systematic review and meta-analysis. Arch Gynecol Obstet. 2022;305:805-24.
8. Zippl AL, Wachter A, Rockenschaub P, Toth B, Seeber B. Predicting success of intrauterine insemination using a clinically based scoring system. Arch Gynecol Obstet. 2022;306:1777-86.
9. Simopoulou M, Sfakianoudis K, Antoniou N, Maziotis E, Rapani A, Bakas P, et al. Making IVF more effective through the evolution of prediction models: is prognosis the missing piece of the puzzle? Syst Biol Reprod Med. 2018;64:305-23.
10. Hariton E, Pavlovic Z, Fanton M, Jiang VS. Applications of artificial intelligence in ovarian stimulation: a tool for improving efficiency and outcomes. Fertil Steril. 2023;120:8-16.
11. Medenica S, Zivanovic D, Batkoska L, Marinelli S, Basile G, Perino A, et al. The future is coming: artificial intelligence in the treatment of infertility could improve assisted reproduction outcomes-the value of regulatory frameworks. Diagnostics (Basel). 2022;12:2979.
12. Bachelot G, Lévy R, Bachelot A, Faure C, Czernichow S, Dupont C, et al.; Alifert Collaborative Group. Proof of concept and development of a couple-based machine learning model to stratify infertile patients with idiopathic infertility. Sci Rep. 2021;11:24003.
13. World Health Organization. WHO laboratory manual for the examination and processing of human semen. 6th ed. Geneva: World Health Organization; 2021. Available from: https://www.who.int/publications/i/item/9789240030787
14. Li M, Duan X, Zhang N, Ding F, Wang Y, Liu P, et al. Development and validation of a conventional in vitro total fertilization failure prediction model. J Assist Reprod Genet. 2023;40:1915-23.
15. van Eekelen R, van Geloven N, van Wely M, McLernon DJ, Mol F, Custers IM, et al. Is IUI with ovarian stimulation effective in couples with unexplained subfertility? Hum Reprod. 2019;34:84-91.
16. Siristatidis C, Vogiatzi P, Pouliakis A, Trivella M, Papantoniou N, Bettocchi S. Predicting IVF outcome: a proposed web-based system using artificial intelligence. In Vivo. 2016;30:507-12.
17. Jiang VS, Pavlovic ZJ, Hariton E. The role of artificial intelligence and machine learning in assisted reproductive technologies. Obstet Gynecol Clin North Am. 2023;50:747-62.
18. Fernandez EI, Ferreira AS, Cecílio MHM, Chéles DS, de Souza RCM, Nogueira MFG, et al. Artificial intelligence in the IVF laboratory: overview through the application of different types of algorithms for the classification of reproductive data. J Assist Reprod Genet. 2020;37:2359-76.

19. Barnett-Itzhaki Z, Elbaz M, Butterman R, Amar D, Amitay M, Racowsky C, et al. Machine learning vs. classic statistics for the prediction of IVF outcomes. J Assist Reprod Genet. 2020;37:2405-12.

20. Depciuch J, Paja W, Pancerz K, Uzun Ö, Bulut H, Tarhan N, et al. Analysis of follicular fluid and serum markers of oxidative stress in women with unexplained infertility by Raman and machine learning methods. J Raman Spectrosc. 2023;54:501-11.

21. Ma R, Qin J, Wang S, Guan S, Jia F, Deng Y, et al. Exploration of immune-related diagnostic biomarkers in unexplained infertility by bioinformatics analysis and machine learning. Taiwan J Obstet Gynecol. 2025;64:438-49.