BMC Pregnancy and Childbirth

RESEARCH

Open Access

Predictive modeling of pregnancy outcomes utilizing multiple machine learning techniques for in vitro fertilization-embryo transfer

Ru Bai[1†], Jia-Wei Li[2†], Xia Hong[1], Xiao-Yue Xuan[1], Xiao-He Li[3*] and Ya Tuo[1*]

Abstract

Objective This study aims to investigate the influencing factors of pregnancy outcomes during in vitro fertilization and embryo transfer (IVF-ET) procedures in clinical practice. Several prediction models were constructed to predict pregnancy outcomes and models with higher accuracy were identified for potential implementation in clinical settings.

Methods The clinical data and pregnancy outcomes of 2625 women who underwent fresh cycles of IVF-ET between 2016 and 2022 at the Reproductive Center of the Affiliated Hospital of Inner Mongolia Medical University were enrolled to establish a comprehensive dataset. The observed features were preprocessed and analyzed. A predictive model for pregnancy outcomes of IVF-ET treatment was constructed based on the processed data. The dataset was divided into a training set and a test set in an 8:2 ratio. Predictive models for clinical pregnancy and clinical live births were developed. The ROC curve was plotted, and the AUC was calculated and the prediction model with the highest accuracy rate was selected from multiple models. The key features and main aspects of IVF-ET treatment outcome prediction were further analyzed.

Results The clinical pregnancy outcome was categorized into pregnancy and live birth. The XGBoost model exhibited the highest AUC for predicting pregnancy, achieving a validated AUC of 0.999 (95% CI: 0.999-1.000). For predicting live births, the LightGBM model exhibited the highest AUC of 0.913 (95% CI: 0.895–0.930).

Conclusion The XGBoost model predicted the possibility of pregnancy with an accuracy of up to 0.999. While the LightGBM model predicted the possibility of live birth with an accuracy of up to 0.913.

Keywords Artificial intelligence, In vitro fertilization-embryo transfer, Prediction model, Pregnancy outcome

†Ru Bai and Jia-Wei Li contributed equally to this work.

*Correspondence:
Xiao-He Li
lixiaohelxh@126.com
Ya Tuo
tuoya12@outlook.com

[1]Reproductive Centre, The Affiliated Hospital of Inner Mongolia Medical University, No.1 of North Tongdao Road, Huimin District, Hohhot 010000, Inner Mongolia Autonomous Region, China
[2]Department of Radiology, The Second Affiliated Hospital of Baotou Medical College, Inner Mongolia University of Science and Technology, Baotou 014000, Inner Mongolia Autonomous Region, China
[3]Department of Anatomy, Zhuolechuan Dairy Development Zone, Basic Medical College Inner Mongolia Medical University, Hohhot 010000, Inner Mongolia Autonomous Region, China

Introduction

With the progression of socio-economic development and improvement in living conditions, infertility defined as the inability of a woman to conceive following at least 12 months of unprotected intercourse [1], has emerged as a significant global concern [2]. Notably, recent data indicate a gradual annual increase in infertility rates in China, with figures reaching as high as 30% [3, 4].

In recent years, in vitro fertilization and embryo transfer (IVF-ET) has emerged as a primary technique for treating infertility [5]. However, through IVF-assisted pregnancy, only about one-third of women achieve live births. Most women experience treatment failure and endure physical and psychological challenges. Additionally, IVF-assisted pregnancy is relatively expensive. Consequently, the majority of women with infertility issues discontinue the treatment.

In clinical practice numerous factors influence pregnancy outcomes, leading to significant variations in clinical pregnancy results [6]. Traditional prediction models primarily depend on the woman's age and the impact of various factors on pregnancy outcomes, however, their predictive accuracy is limited. Therefore, an objective and accurate prediction modeling scheme based on individual conditions is needed in clinical practice to enhance the pregnancy and live birth rates and reduce discomfort for women undergoing IVF-assisted pregnancy.

Machine learning is an emerging multidisciplinary and cross-disciplinary field of artificial intelligence (AI), and its application in healthcare has become a focal point in contemporary society [7]. AI is a new technical science to study and develop theories, methods, technologies and application systems for simulation, extension and expansion of human intelligence. Its content mainly includes language recognition, image recognition, natural language processing and expert system, etc. AI can better help doctors through medical image processing, digital pathology, and improve disease prevention, diagnosis and prognosis, through feature analysis and training based on imaging and conventional pathological sections. In recent years, some scholars have been developing machine learning techniques for health management and predictive image pathology diagnosis, and building predictive models [8, 9]. The application of machine learning in assisted reproduction has also been reported in large numbers.

Internationally, researchers have utilized deep learning, convolutional neural networks, artificial neural networks, support vector machines, and other models for data screening, embryo selection, IVF-ET, and outcome prediction [10]. However, there are persisting issues with the current prediction models, such as low prediction accuracy and the absence of standardization. Therefore, in this study, an AI model was developed to predict pregnancy outcomes from clinical data of women undergoing IVF-assisted pregnancy. This model can reduce the physical and psychological stress of women undergoing this procedure, save time and money, improve pregnancy rates, and aid in guiding clinical treatment and patient counseling.

Materials and methods

Study participants

The data for the study were sourced from the clinical data and pregnancy outcomes of women who visited the Reproductive Center of the Affiliated Hospital of Inner Mongolia Medical University and underwent fresh cycle IVF between 2016 and 2022.

Inclusion criteria consisted of individuals meeting the following conditions: ① Aged between 20 and 40 years old; ② women undergoing a fresh cycle of embryo transfer.

Exclusion Criteria consisted of the following conditions: ① Presence of other systemic diseases such as hypertension, diabetes mellitus, heart disease, and other chronic conditions that might significantly affect the progress of IVF and pregnancy outcomes; ② Women who underwent IVF primarily due to male partner infertility factors; ③ Women with chromosomal abnormalities, recurrent miscarriages, and decreased ovarian reserve function; ④ Medical records with significant data gaps.

The clinical data of a cohort of 3880 women with infertility were included, and 1255 women had either missing data or count for follow-up. Finally, 2625 women were enrolled in this study. The cohort was divided into a clinical pregnancy group ($n = 2031$) and a non-clinical pregnancy group ($n = 594$) based on different clinical pregnancy outcomes. Additionally, they were categorized into a live birth group ($n = 1711$) and a non-live birth group ($n = 320$). Each index was thoroughly assessed during the clinical data collection process, with indices exhibiting a missing data rate of ≥ 30% being excluded from the study. (Fig. 1)

Study indicators

Information collection

Medical records were compiled from the enrolled participants. Information on age, infertility factors, years of infertility, body mass index, type of infertility, history of menstruation, disease and surgical history, drug allergies, and smoking and alcoholism history of the study participants were recorded for subsequent statistical analysis.

Detection of the basic female sex hormone levels

Before initiating the IVF-ET transfer cycle, 4 mL of blood was drawn from the median cubital vein on the 2nd to 3rd day of menstruation, and serum was separated for each participant.

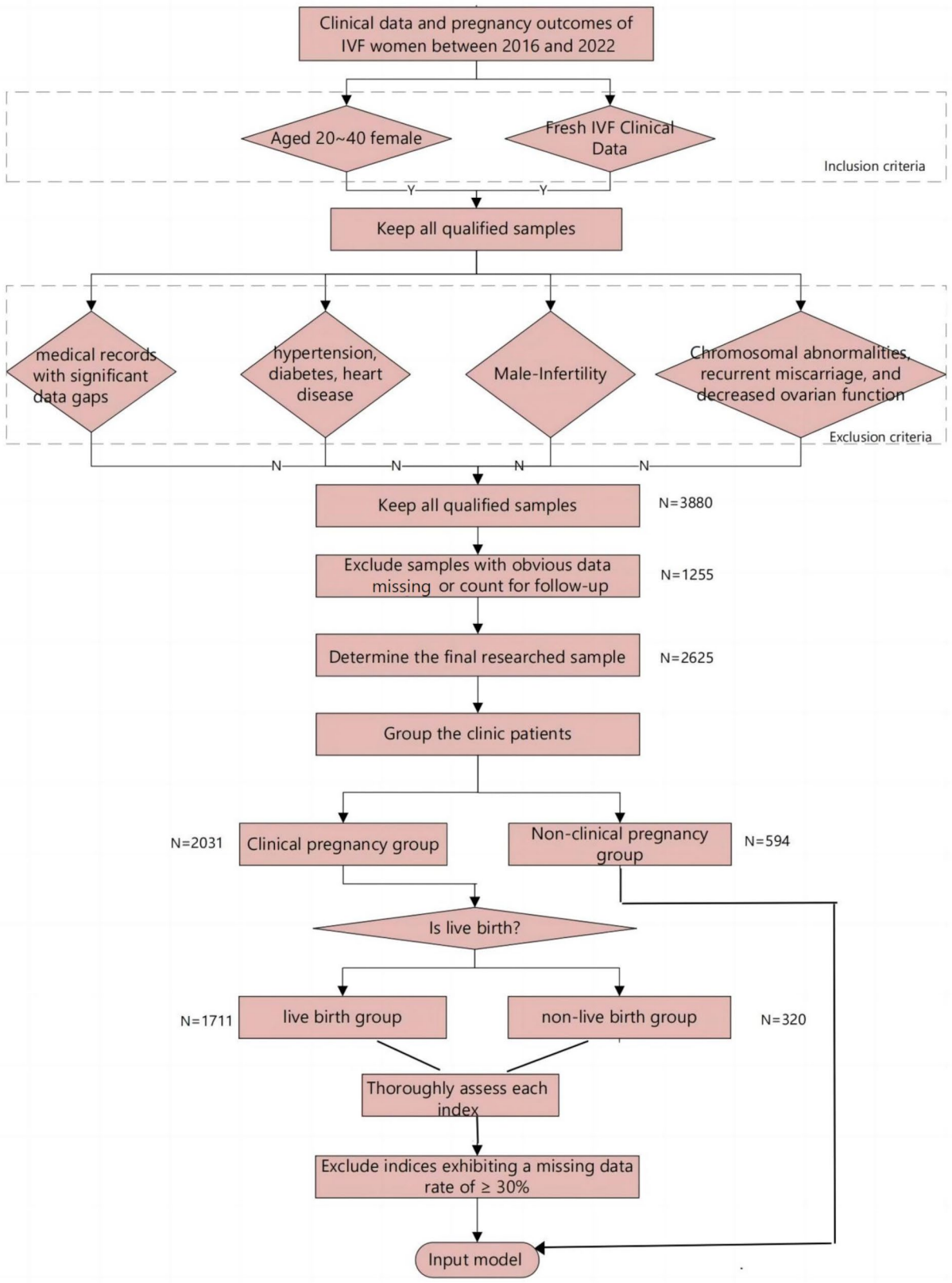

Fig. 1 Flow diagram

Karyotype analysis

For peripheral blood karyotyping, 2 mL of heparin-anticoagulated venous blood was collected for lymphocyte culture and subsequent preparation. This process typically required 2 to 3 weeks to be completed.

Determination of parameters after commencement of the cycle

Upon commencement of the cycle, data were collected for each participant, including initial gonadotropin (Gn) dose, Gn stimulation duration, total Gn dose, the number of oocytes retrieved, and the number of 2PN oocytes. Real-time transvaginal color Doppler ultrasound was utilized to determine the antral follicle count.

Throughout the ovulation induction process, final oocyte maturation was triggered when one follicle reached a diameter of ≥ 20 mm or when two follicles reached a diameter ≥ 18 mm, in accordance with standard practice at our institution. On the day of triggering, serum hormone levels (estradiol, luteinizing hormone, and progesterone) were assessed. Oocyte retrieval was performed approximately 36 to 38 h post-trigger using an oocyte retrieval needle.

On the third-day post-retrieval, ovarian size and the presence of pelvic or abdominal effusion were evaluated via ultrasound. This assessment was performed to monitor for potential ovarian hyperstimulation syndrome (OHSS) and other complications. Additionally, the number of fertilized oocytes, the number of normally fertilized cleaved embryos, the number of high-quality embryos, and the number of implantation cycles were documented.

Embryo grading was conducted on day three post-retrieval, as embryos typically exhibit stable developmental trends by this stage, allowing differentiation between good- and poor-quality embryos. Based on the grading, two high-quality embryos were generally transferred on day three. In certain cases, due to personal or physical factors, embryo transfer was postponed to a subsequent cycle.

Embryo transfer, clinical pregnancy outcome monitoring indicators

The third day embryos were graded, and two embryos were transferred. Embryo grading was conducted because embryos exhibit a stable development trend on day three, allowing differentiation between good- and poor- quality embryos. Most transfers occurred on day three after the embryo grading; however, a small number of patients opted to defer transfer to a subsequent cycle due to personal or physical factors. Embryos were graded as: Grade A (Excellent): The embryonic cell mass is uniform and forms a symmetrical sphere, with consistent size, color and density of individual blastomeres. The development stage matches expected progress. Grade B (Good): The embryonic cell mass exhibits slight irregularities in size, shape, color, or density of individual cells. Grade C (Poor Quality): The embryonic cell mass displays significant irregularities in size, shape, color, or density. Grade D (Degenerated): Embryos that are non-viable due to degeneration or death. Grades A and B were classified as high-quality embryos, Grade C as usable embryos, and Grade D embryos were discarded. High-quality embryos were prioritized for transfer.

Pregnancy monitoring began with the measurement of human chorionic gonadotropin (HCG) levels in blood drawn 14 days post-embryo transfer. An HCG level ≥ 5 IU/L was considered indicative of pregnancy. Clinical pregnancy was confirmed 28 days post-embryo transfer using transvaginal ultrasound to detect the presence of gestational sac and cardiac pulsation.

Participant with confirmed pregnancies were followed up to monitor pregnancy progression. In cases of sustained pregnancy, details such as the type of pregnancy, number of live births, sex of live births, and mode of delivery were recorded. For pregnancy did not continue, the cause and mode of miscarriage were documented, and chromosomal tests were conducted when necessary to identify fetal chromosomal abnormalities.

Research method

Statistical methods

SPSS 26.0 software was used for statistical analysis. Python 3.11 software was used to construct prediction models based on various machine-learning methods.

Machine learning methods

A total of eight common machine learning models were constructed, including support vector machine (SVM) [11], K-nearest neighbor (KNN) [12], random forest (RF) [13], extremely randomized trees (Extra Trees) [14], extreme gradient boosting (XGBoost) [15], multilayer perceptron (MLP) [16], logistic regression (LR) [17], and light gradient boosting machine (LightGBM). The dataset was divided into a training set comprising 80% of the data and a test set containing the remaining 20%.

① SVM is a binary classification model. Its basic model is defined as a linear classifier with maximum intervals on the feature space, which distinguishes it from perceptron machines. SVM is a powerful tool for constructing classifiers. Its objective is to create a decisive boundary between two classes to facilitate label prediction from one or more feature vectors. This decisive boundary, known as the separating hyperplane, is oriented to maximize the distance from the closest data points in each class. These nearest points are called support vectors [18]. As depicted in Fig. 2, the equation $w^*x + b = 0$ represents the separating hyperplane. While a linearly

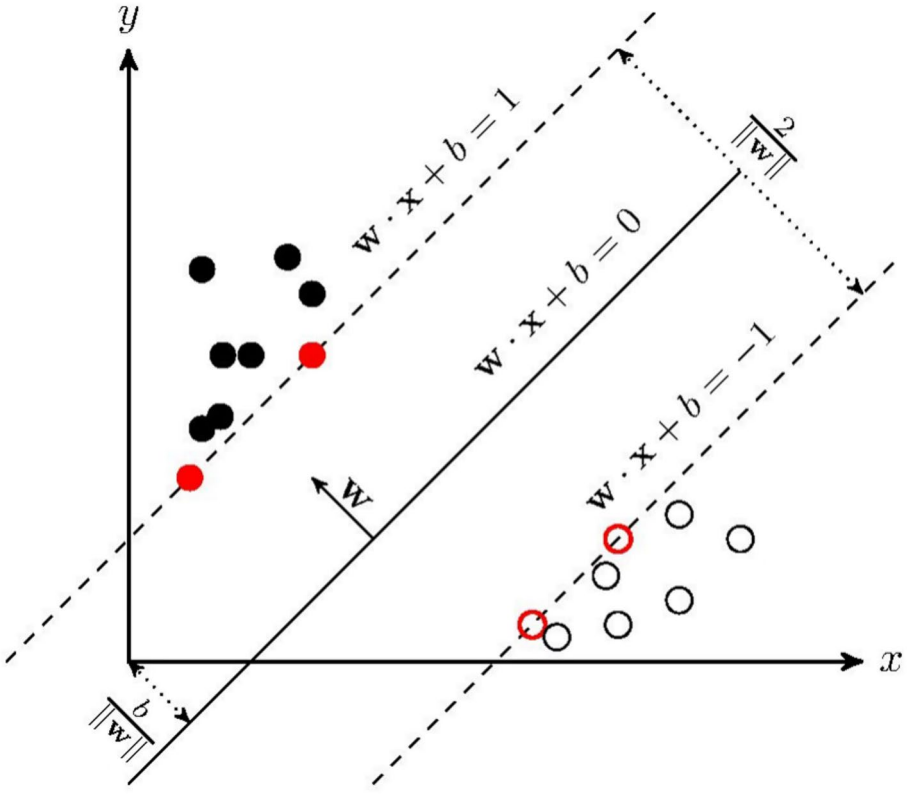

Fig. 2 Operating principle of SVM

separable dataset (i.e., perceptron) may consist of infinitely many such hyperplanes, the separating hyperplane with the largest geometric margin is unique.

② The KNN algorithm is one of the most fundamental and simplest machine learning algorithms. The KNN algorithm is a supervised machine learning algorithm that is primarily used for classification purposes. It has been widely used for disease prediction [19]. It can be used for both classification tasks and regression analysis. The operational principle underlying KNN is intuitive; it classifies data points by assessing the proximity between their feature values [20]. The concept of the KNN algorithm is straightforward: for any n-dimensional input vector representing a point in the feature space, the output is the category label or prediction value associated with that feature vector. It does not have a conventional learning process. It works by partitioning the feature vector space using training data and utilizing the partitioning result as the definitive algorithmic model.

③ RF is an algorithm that integrates multiple trees through the concept of ensemble learning. Its fundamental component is the decision tree, and its principle belongs to a major branch of machine learning known as the ensemble learning method [21]. As a non-parametric technique, RF demonstrates versatility across various types of response variables, accommodating tasks ranging from categorical classification to quantitative outcome prediction and survival analysis [22]. The nomenclature "Random forest" is derived from two keywords: "random" and "forest." The term "Forest" refers to the aggregation of multiple trees. Such an analogy is still very apt. This is also the embodiment of the main idea of RFs—the concept of ensemble learning.

④ ExtRa Trees is an ensemble classifier method. It represents an advancement over the traditional decision tree by incorporating randomness into the construction process. During the learning process, an ensemble of randomized trees is constructed supervised, and then used to forecast the output of previously unseen input vectors [23].

⑤ The XGBoost algorithm is based on the principle of gradient boosting decision trees and utilizes a second-order Taylor expansion methodology to calculate the loss function. This approach is known for its outstanding performance in computational speed and prediction accuracy. XGBoost continuously splits the features to facilitate tree growth and iteratively refines predictions by fitting residuals from preceding models. The training process is followed by the development of K-trees. The features of a specific sample correspond to a distinct leaf

node of each tree and are assigned a score. Subsequently, the predicted value for a specific sample is calculated by aggregating the scores of each tree [24]. XGBoost excels as a robust classifier, leveraging the integration of weak classifiers to yield several advantages including efficient handling of missing values, mitigation of overfitting risks, and expedited runtime performance through parallel and distributed computation strategies.

⑥ MLP is a foundational binary linear classification model and is a classical algorithm in the field of machine learning. The basic principle of the perceptron is to compute a linear combination of input features and then use an activation function to classify the results into binary categories [25].

⑦ LR is a generalized linear regression analysis model and is classified under the category of supervised learning in machine learning. Its derivation process and calculation are similar to the process of regression. However, it is primarily utilized to address binary classification problems and solve multi-classification problems. The model undergoes training utilizing a designated set of n data points, commonly referred to as the training set. Following training, the same set or different sets of data are classified and denoted as test data [26].

⑧ The LightGBM model is a gradient-boosting framework that utilizes decision trees as its fundamental learners. Distinguished by its design for efficient parallel computing, LightGBM offers notable advantages including accelerated training speed, reduced memory usage, support for single-computer multithreading, multicomputer parallel computing, GPU training, and the capability to process large-scale data.

Construction of a prediction model for pregnancy outcomes in IVF-assisted pregnancy

Machine learning prediction model selection

A total of eight common machine learning models were constructed, which included SVM, KNN, RF, Extra Trees, XGBoost, MLP, LR, and LightGBM. The models underwent training using 80% of the available dataset, while the remaining 20% was reserved for testing and evaluation.

Machine learning model construction process

Data selection

The clinical data of women who underwent fresh cycles of in vitro fertilization-embryo transfer at the Reproductive Center of the Affiliated Hospital of Inner Mongolia Medical University between 2016 and 2022 were collected. A cohort comprising 2165 cases of infertile women was included based on the specified inclusion and exclusion criteria. A total of 30 observational indices were collected, including age, body mass index, and other relevant factors including: (1) Basic conditions: infertility factors, years of infertility, body mass index, and infertility types. (2) Detection of basic sex hormone levels: before each participant enters the cycle of in vitro fertilization-embryo transfer, 4 mL of elbow venous blood on the 2 to 3 days of menstruation, Analysis of the basic hormone luteinizing hormone, follicle stimulating hormone, prolactin, progesterone, oestradiol and anti-Mullerian tube hormones was measured by automated immunosorbent assay. (3) Observational indicators monitored during the treatment cycle after the plan had been established. These include: the starting dose of Gonadotropin (Gn), the total days of Gn administration, the total dose of Gn, the total number of follicles, the daily hormone levels of HCG (LH, E2, P), the number of two-pronuclei embryos (2PN), the number of transferable embryos, the number of high-quality embryos, endometrial thickness on the day of blastocyst transfer, the number of embryos transferred, the total number of retrieved oocytes, and other relevant factors. The inclusion of different reproductive centers was to remove bias resulting from data collected by just one reproductive center. Confounding factors such as age and BMI were considered because the clinical work may affect the embryo transfer results.

Data check

A total of 3880 women were included in this study. Among them, 1255 women had missing data or were lost to follow-up, resulting in a final cohort of 2625 women for analysis. Missing data points were not imputed but excluded from the modeling process to ensure data integrity. Improbable values, which were unlikely to occur in actual clinical testing, underwent verification and, if necessary, were corrected.

Data preprocessing and feature selection

① The features were standardized using the Z-score, and the dataset was adjusted to conform to a normal distribution with a mean of 0 and a standard deviation of 1, N ~ (0,1). ② Spearman's correlation coefficient was used to assess the correlation between the features for the extracted histological features. ③ Features exhibiting a correlation coefficient exceeding 0.9 underwent a selection process, whereby one of the two highly correlated features was retaFig. (Fig. 3). The results lead to the following conclusions: the prediction model for pregnancy demonstrated a strong correlation between the counts of left and right follicles and the total follicle count; the count of 2PN and transferable embryos exhibited a strong correlation with the number of inseminated oocytes; and the serum HCG results exhibited a pronounced correlation with the clinical pregnancy outcome. ④ Cross-validation was conducted on the data using Lasso regression to determine the optimal penalization coefficients. ⑤ The dataset was randomly divided into groups: a completely

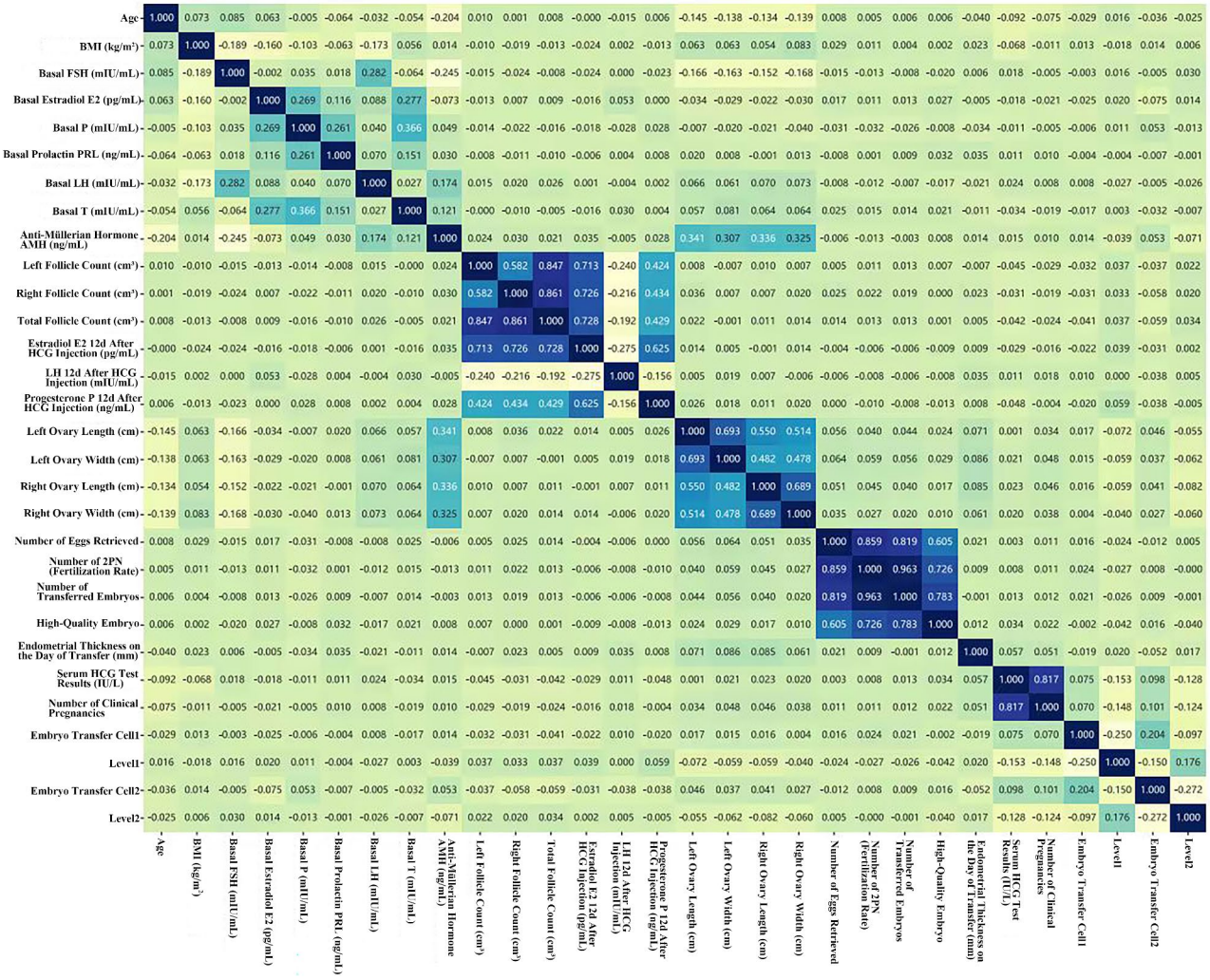

Fig. 3 Spearman correlation coefficient. Level 1: high-quality embryos (Grade A or B); Level 2: usable embryos (Grade C)

randomized approach was used with multiple attempts to ascertain optimal division. N-fold cross-validation was conducted, and the best result was selected. The selected features were those with feature coefficients greater than 0 after cross-validation.

Determination of the training and test set

A total of 3880 women were included in this study, and 1255 women had either missing data or count for follow-up. Therefore, a total of 2625 women were modeled. The test set consisted of 2100 cases, which accounted for 80% of the overall samples, selected randomly. Conversely, the training set consisted of the remaining 525 cases, constituting 20% of the total dataset.

Training set for 5-fold cross-validation

To achieve the optimal training model, the data set designated for training underwent a 5-fold cross-validation procedure, enabling the derivation of the most conducive training model.

Comparison of multiple machine models and their performance

The evaluation involved validation using an independent test set to assess the generalization capabilities, record relevant parameters, and determine the advantages and disadvantages of each model.

① The results of the pregnancy prediction model are displayed in Table 1. Notably, RF, XGBoost, and LightGBM exhibited higher accuracy in both the training and test sets. The recorded accuracy values for the training dataset were 0.997 for RF, 0.996 for XGBoost, and 0.999 for LightGBM. In the test dataset, these models achieved accuracy values of 0.964, 0.986, and 0.990, respectively. A comparative analysis of accuracy is illustrated in Fig. 4. The tabulated results suggest that the XGBoost model

Table 1 Comparison of whether pregnancy prediction models

Model	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Precision	Recall	F1	Task
SVM	0.913310	0.973367	0.9683–0.9785	0.987077	0.911367	0.929114	0.896914	0.929114	0.987077	0.957219	clinical pregnancy-train
SVM	0.882732	0.965351	0.9532–0.9775	0.990148	0.886486	0.917772	0.849624	0.917772	0.990148	0.952587	clinical pregnancy-test
KNN	0.863680	0.933416	0.9256–0.9412	0.827077	0.903924	0.904441	0.826221	0.904441	0.827077	0.864031	clinical pregnancy-train
KNN	0.759021	0.831188	0.8035–0.8589	0.743842	0.786301	0.784416	0.734015	0.784416	0.743842	0.763590	clinical pregnancy-test
RandomForest	0.996777	0.999983	1.0000–1.0000	0.997538	0.998647	0.993884	1.000000	0.993884	0.997538	0.995708	clinical pregnancy-train
RandomForest	0.963918	0.994415	0.9902–0.9986	0.997537	0.943243	0.937500	0.997093	0.937500	0.997537	0.966587	clinical pregnancy-test
ExtraTrees	1.000000	1.000000	nan - nan	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	clinical pregnancy-train
ExtraTrees	0.923969	0.972337	0.9615–0.9832	0.953202	0.899183	0.906323	0.945559	0.906323	0.953202	0.929172	clinical pregnancy-test
XGBoost	0.994844	0.999439	0.9985–1.0000	0.995077	0.997292	0.997526	0.991925	0.997526	0.995077	0.996300	clinical pregnancy-train
XGBoost	0.985825	0.998995	0.9981–0.9999	0.985222	0.991892	0.987654	0.983827	0.987654	0.985222	0.986436	clinical pregnancy-test
LightGBM	0.983887	0.999434	0.9992–0.9997	0.987077	0.990528	0.995595	0.971598	0.995595	0.987077	0.991318	clinical pregnancy-train
LightGBM	0.989691	0.998958	0.9980–0.9999	0.985222	0.994595	0.995025	0.983957	0.995025	0.985222	0.990099	clinical pregnancy-test
MLP	0.915566	0.977672	0.9736–0.9818	0.972308	0.905954	0.927809	0.902649	0.927809	0.972308	0.949537	clinical pregnancy-train
MLP	0.902062	0.970576	0.9603–0.9809	0.967980	0.908108	0.923077	0.880829	0.923077	0.967980	0.944995	clinical pregnancy-test
LR	0.894940	0.970644	0.9655–0.9758	0.980308	0.908660	0.921480	0.868758	0.921480	0.980308	0.949984	clinical pregnancy-train
LR	0.886598	0.972993	0.9633–0.9827	0.972906	0.918919	0.920635	0.854271	0.920635	0.972906	0.946049	clinical pregnancy-test

exhibits enhanced performance in predicting pregnancy outcomes.

② The results of the live birth prediction models are illustrated in Table 2. The models with higher accuracy on the training and test sets were the XGBoost and LightGBM models. The accuracy values on the training set were recorded as 0.884 and 0.857 for the XGBoost and LightGBM models, respectively. In the test dataset, these models achieved accuracy values of 0.712 and 0.744. A comparison of the accuracy is depicted in Fig. 5. The tabulated results suggest that the LightGBM model exhibits enhanced performance in predicting live birth outcomes.

Plotting the area under the ROC curve (AUC) and confusion matrix

AUC: The AUCs for the models predicting pregnancy and live births are displayed in Fig. 6. From the figure, it is evident that the XGBoost model outperforms the other models in predicting pregnancy in both the training and test datasets. The LightGBM model outperformed other models in predicting live birth outcomes in both the training and test datasets.

A confusion matrix is a specific structured tabular representation used to visually assess the performance and efficiency of an algorithm. In this matrix, each row represents the actual category, and each column represents the predicted category. The confusion matrices of the two prediction models presented herein illustrate the outcomes of the different prediction models on the training and validation subsets of the dataset under study. Examination of these matrices reveals a presence of true positives (TPs) and true negatives (TNs) in both the training and test datasets, indicative of the considerable predictive accuracy exhibited by the machine learning models used in this study about pregnancy or live birth prediction (Fig. 7).

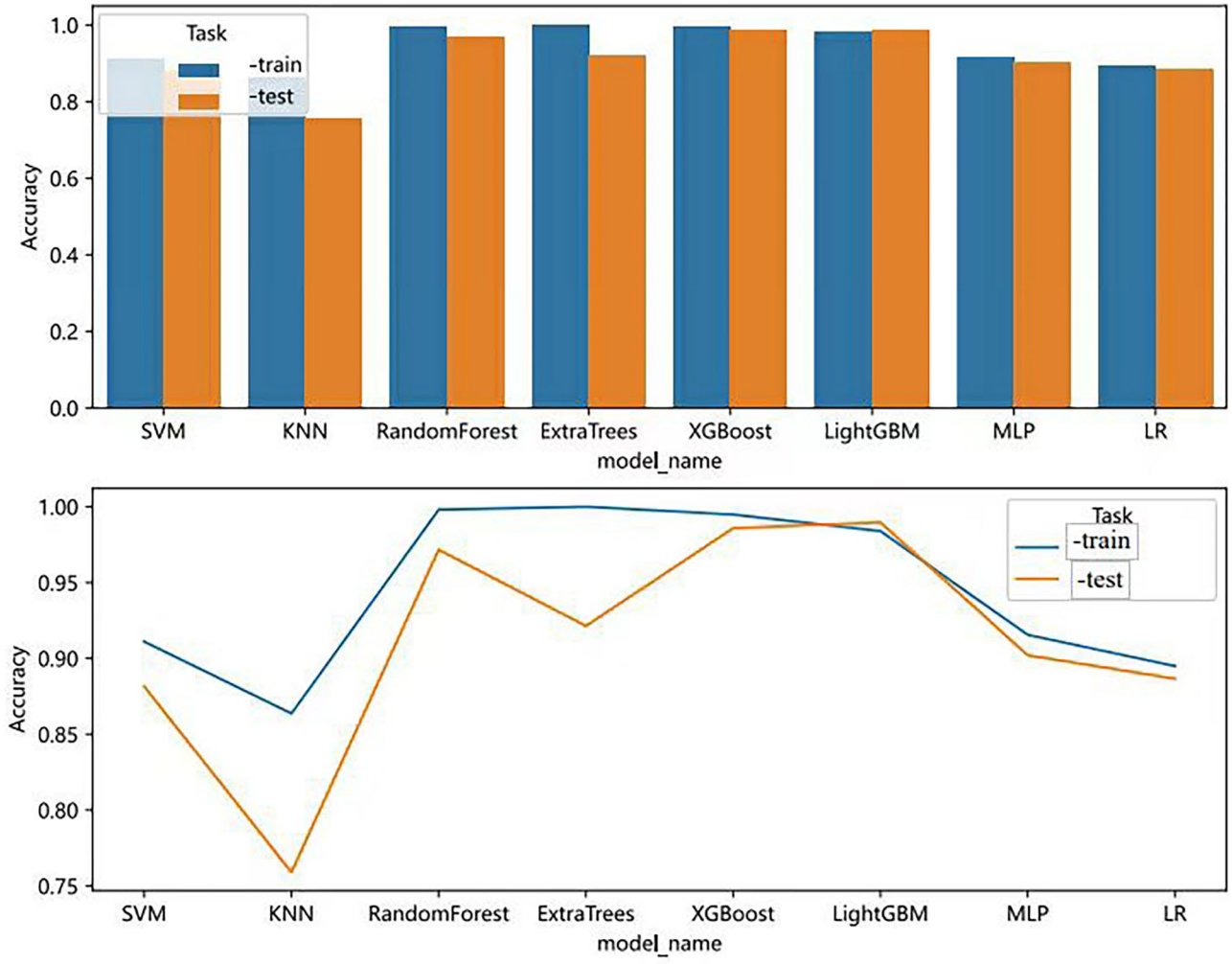

Fig. 4 Comparison of accuracy of pregnancy prediction models

Table 2 Comparison of live birth prediction models

Model	Accuracy	AUC	95% CI	Sensitivity	Specificity	PPV	NPV	Precision	Recall	F1	Task
SVM	0.822769	0.894106	0.8774–0.9108	0.870000	0.784151	0.512766	0.948918	0.512766	0.870000	0.645238	live birth-train
SVM	0.751232	0.661511	0.5958–0.7273	0.720000	0.557576	0.348837	0.859375	0.348837	0.720000	0.469974	live birth-test
KNN	0.666462	0.962911	0.9524–0.9734	0.770000	1.000000	0.355263	0.996198	0.355263	0.770000	0.486202	live birth-train
KNN	0.517241	0.587976	0.5231–0.6529	0.693333	0.570397	0.231111	0.872928	0.231111	0.693333	0.346667	live birth-test
RandomForest	0.991385	0.999335	09986-1.0000	0.993333	0.990943	0.961290	0.998479	0.961290	0.993333	0.977049	live birth-train
RandomForest	0.731527	0.710916	0.6499–0.7719	0.760000	0.557576	0.339623	0,870000	0.339623	0.760000	0.469458	live birth-test
ExtraTrees	1.000000	1.000000	nan -nan	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	live birth-train
ExtraTrees	0.726601	0.655227	0.5945–0.7159	0.733333	0.533742	0.275000	0.837423	0.275000	0.733333	0.400000	live birth-test
XGBoost	0.884308	0.955826	0.9465–0.9651	0.906667	0.875472	0.635922	0.968673	0.635922	0.906667	0.747535	live birth-train
XGBoost	0.711823	0.717744	0.6576–0.7779	0.840000	0.530303	0.294118	0.851974	0.294118	0.840000	0.435685	live birth-test
LightGBM	0.856615	0.912503	0.8951-0.9300	0.856667	0.817358	0.584810	0.943902	0.584810	0.856667	0.695103	live birth-train
LightGBM	0.743842	0.745680	0.6878–0.8036	0.786667	0.612121	0.369369	0.884746	0.369369	0.786667	0.502702	live birth-test
MLP	0.881231	0.949351	0.9386–0.9601	0.930000	0.846038	0.628297	0.968543	0.628297	0.930000	0.749942	live birth-train
MLP	0.726601	0.658651	0.5942–0.7231	0.666667	0.600000	0.320000	0.859477	0.320000	0.666667	0.432432	live birth-test
LR	0.632000	0.702143	0.6731–0.7312	0.830000	0.488671	0.290141	0.897268	0.290141	0.830000	0.429976	live birth-train
LR	0.605911	0.662397	0.6021–0.7227	0.653333	0.621212	0.267760	0.883408	0.267760	0.653333	0.379845	live birth-test

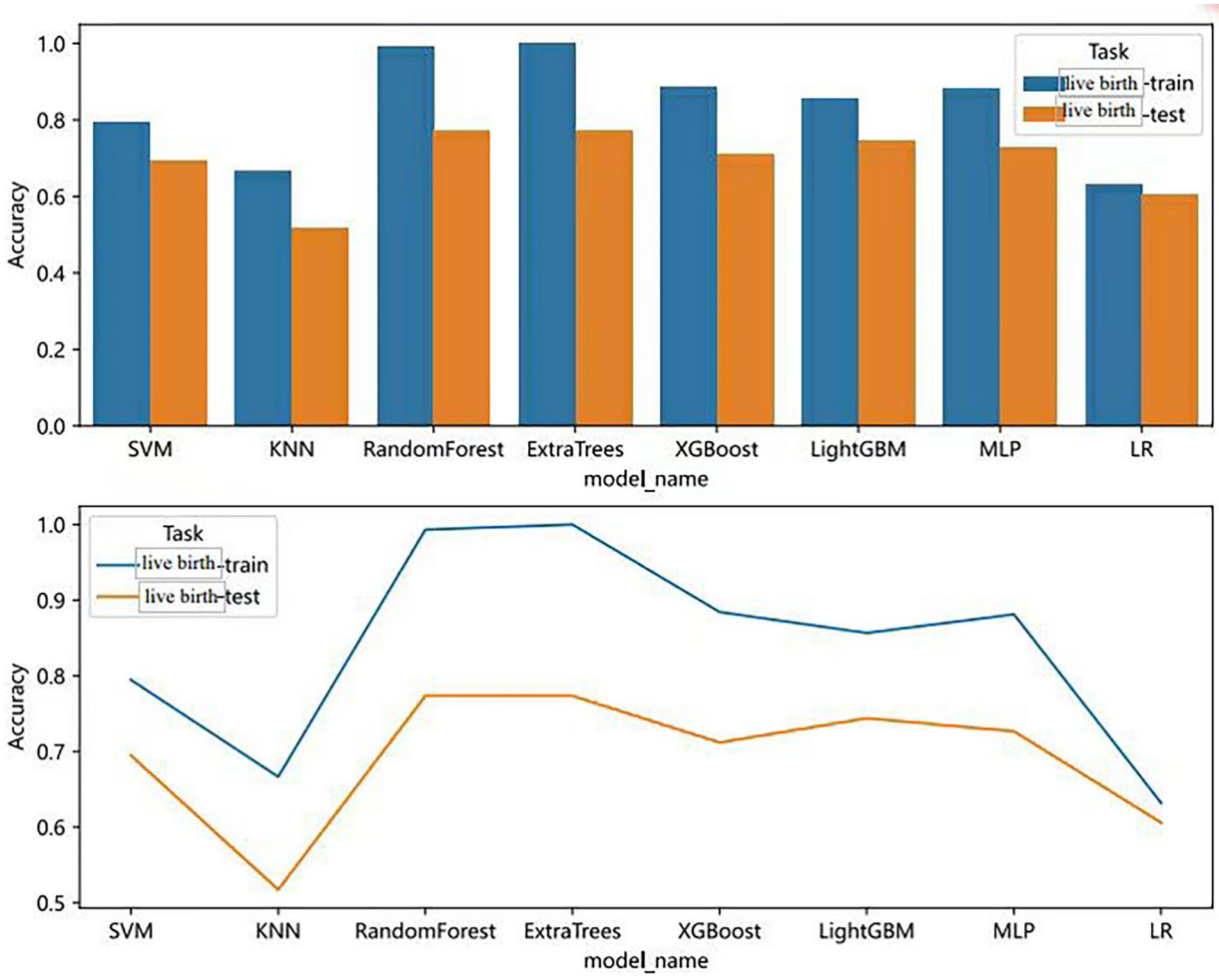

Fig. 5 Comparison of accuracy of live birth prediction models

Discussion

The outcomes of IVF-ET treatment are influenced by various factors, and currently, the extent of their impact on outcomes remains inconclusive [27].

In clinical practice, predictions of IVF-ET treatment success are traditionally based on factors such as the age of the woman undergoing the procedure and the historical success rate of the treatment center, resulting in suboptimal accuracy prediction. Therefore, incorporating AI-based prediction models into IVF-ET processes could potentially improve clinical pregnancy rates and refine the accuracy of these predictions, benefiting clinical practice.

In this study, clinical data from women who underwent IVF-ET at the Reproductive Center of the Affiliated Hospital of Inner Mongolia Medical University were collected. Prediction models were developed using both pre- and post-treatment variables to enhance predictive accuracy for pregnancy and live birth outcomes. While the study demonstrates that it is possible to develop algorithms that predict IVF outcomes once treatment has started, the model does not currently support pre-treatment counseling for couples. This limitation arises because the training and test data included input points from both pre- and during-treatment stages. Despite this, the proposed models hold significant clinical value in reducing unnecessary embryo transfers, alleviating physical and psychological stress, and saving time and costs, which are critical factors for patients undergoing IVF-ET.

Future research should prioritize developing predictive models that focus solely on pre-treatment data, allowing for more practical applications in counseling couples before they begin IVF treatment. Such models would be beneficial in providing couples with a clearer understanding of potential outcomes prior to embarking on the IVF process, which would be a significant advancement in clinical decision-making.

Following data collection and preprocessing, various machine-learning methodologies were utilized to develop prediction models for different pregnancy outcomes. Through comprehensive evaluation encompassing metrics such as accuracy, AUC, 95% confidence

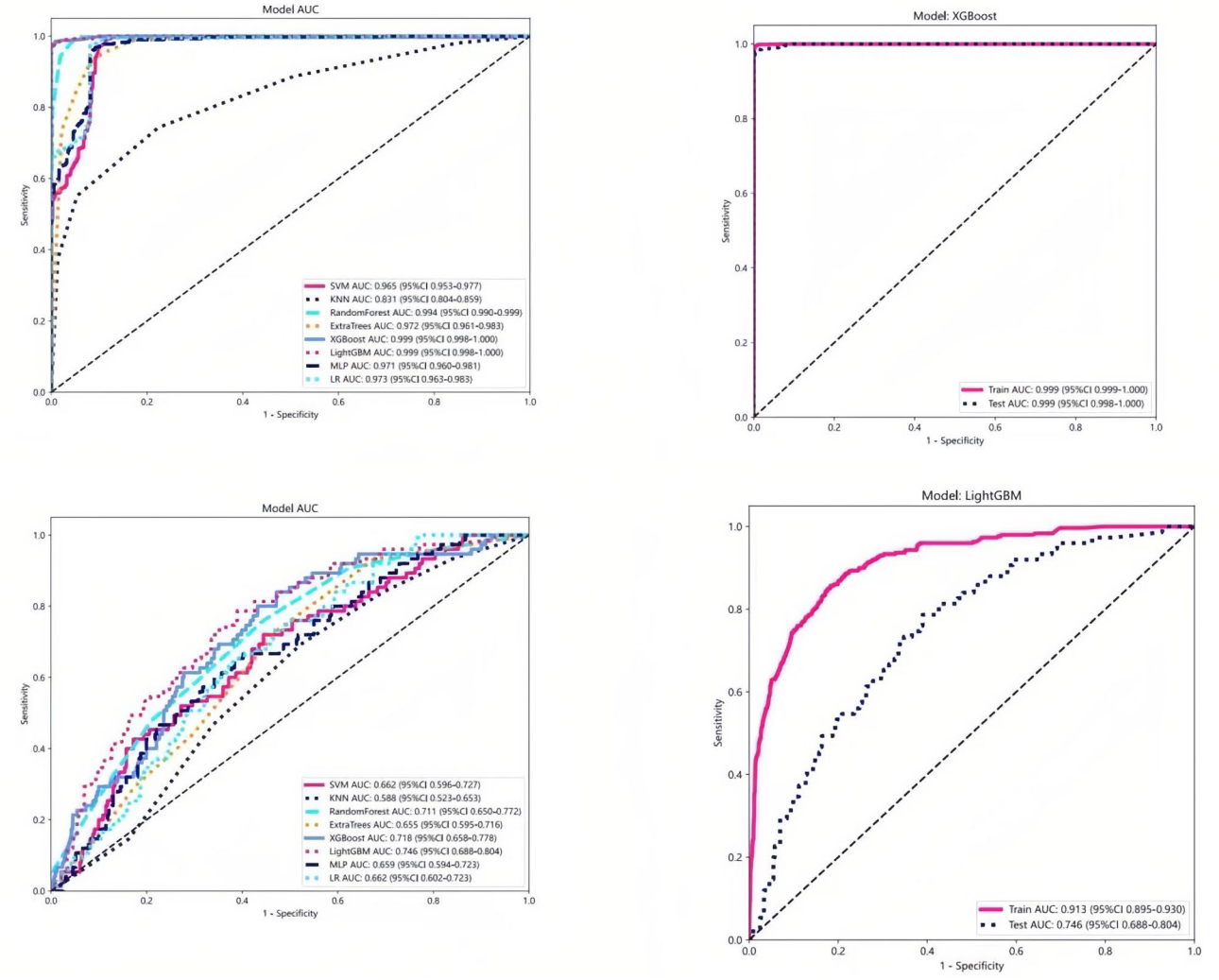

Fig. 6 **A**: Multiple models predict pregnancy AUC. **B**: XGBoost model training set and test set AUC. **C**: Multiple models predict live birth AUC. **D**: LightGBM model training set and test set AUC

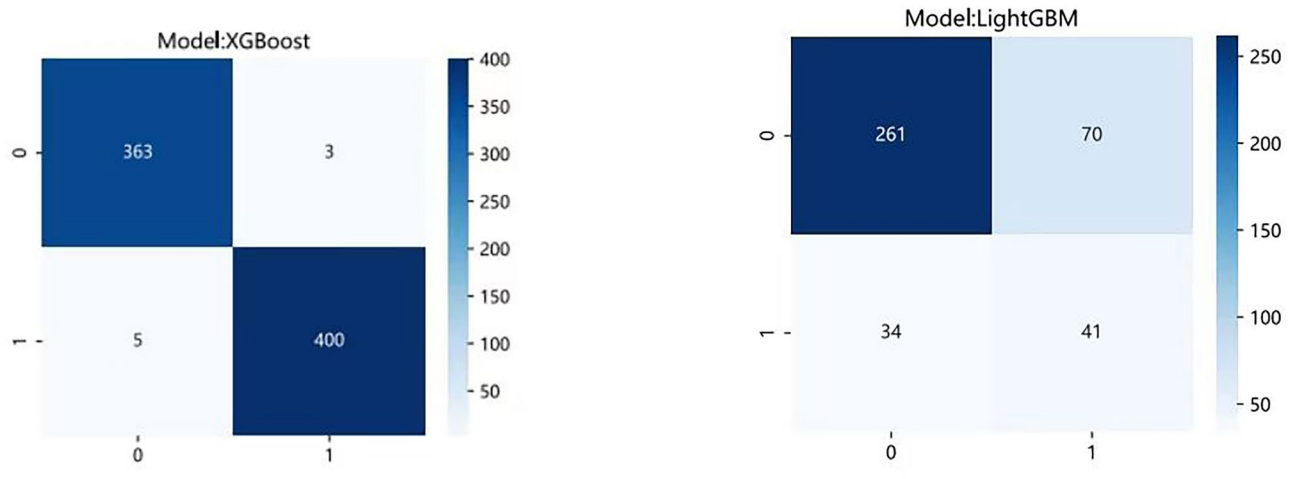

Fig. 7 **A**: XGBoost model confusion matrix. **B**: LightGBM model confusion matrix

interval, sensitivity, specificity, recall, precision, positive prediction rate, negative prediction rate, and F1 index, the model with the best overall predictive efficacy for pregnancy was XGBoost, while the model with the best overall predictive efficacy for live birth was LightGBM. The implementation of these models within clinical settings has the potential to predict pregnancy outcomes and offer additional value.

The integration of AI into assisted reproduction represents a revolutionary advancement, introducing novel opportunities in clinical medicine. Across various fields ranging from patient counseling to embryo transfer, AI-driven solutions have transformed every aspect, particularly in the field of IVF [28]. As cutting-edge research unleashes its comprehensive capabilities, AI holds the potential to revolutionize the IVF stimulation cycle in ways previously inconceivable.

However, persistent shortcomings, particularly of a notable absence of standardization is evident. As an emerging technology, AI lacks uniform standards and norms. Furthermore, there exists variability in data collection standards across different healthcare institutions, resulting in significant discrepancies in results and the potential risk of a data breach [29].

The major limitation of this study relates to the day of embryo transfer. In Inner Mongolia, most women undergoing IVF choose to transfer fresh embryos on the third day, which is different from the standard of care in many other centers that focus on blastocyst transfers and at times freeze all cycles. Therefore, our study conducted model prediction based on the IVF characteristics of local infertile patients, and so inevitably limits the generalizability of our findings. In addition, this is a preliminary study to explore models for predicting pregnancy outcomes in infertile patients, regardless of any factors. Further testing the predictive performance of different models in standard category infertility patients will provide a more favorable evaluation tool for clinical practice.

Conclusion

In conclusion, while AI and machine learning hold significant promise for revolutionizing outcomes in IVF-ET, offering powerful tools for both physicians and women with infertility, these technological advances must be carefully managed to address ethical and standardization challenges. Furthermore, this study highlights the need for future research to develop predictive models focusing solely on pre-treatment data to support counseling couples before initiating IVF treatment. Such advancements would enhance the clinical utility of AI models, providing more comprehensive decision-making support and further improving patient outcomes.

Abbreviations

ExtRa Trees	Extremely Randomized Trees
IVF-ET	In vitro fertilization and embryo transfer
KNN	K-NearestNeighbor
LightGBM	Light Gradient Boosting Machine
LR	Logistic regression
MLP	Multilayer perceptron
RF	Random forest
SVM	Support vector machine

Author contributions

Ru Bai collected the data, drafted, and edited the manuscript. Jia-Wei Li, Xia Hong, and Xiao-Yue Xuan collected, interpreted, and analyzed the data. Xiao-He Li and Ya Tuo conceptualized the study, analyzed the data, obtained the funds, and edited the manuscript. All authors read and approved the final draft.

Funding

This study was funded by the Key Research and Development and Achievement Transformation Plan Projects of Inner Mongolia Autonomous Region (Grant Number: 2023YFSH0011), Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (Grant Number: NJYT22006), Inner Mongolia Medical University General Engineering Project (Grant Number: YKD2023MS019), Natural Science Foundation of Inner Mongolia Autonomous Region (Grant Number: 2021MS08022), Program for Innovative Research Team in Universities of Inner Mongolia Autonomous Region (Grant Number: NMGIRT2227), Key Research Project of Inner Mongolia Medical University (Grant Number: YKD2021ZD001), National Health Commission Hospital Management Research Institute Project (Grant Number: YLXX24AIA005), Inner Mongolia Public Hospital Scientific Research Joint Fund Science and Technology Major Project (Grant Number: 2024GLLH0285), Inner Mongolia Autonomous Region Medical Association Clinical Medical Research and Clinical New Technology Promotion Project (Grant Number: YSXH2024KYF033), and Inner Mongolia Medical University ZhiYuan Talent (Grant Number: ZY20242134). The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study was conducted with approval from the Ethics Committee of The Affiliated Hospital of Inner Mongolia Medical University (Approval Number: S.2022055). This study was conducted in accordance with the declaration of Helsinki. Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 30 August 2024 / Accepted: 6 March 2025
Published online: 19 March 2025

References

1. Infertility Workup for the Women's Health Specialist. ACOG committee opinion, number 781. Obstet Gynecol. 2019;133(6):e377–84. https://doi.org/10.1097/AOG.0000000000003271
2. Sharma RS, Saxena R, Singh R. Infertility & assisted reproduction: a historical & modern scientific perspective. Indian J Med Res. 2018;148(Suppl):S10–4. https://doi.org/10.4103/ijmr.IJMR_636_18
3. Liang S, Chen Y, Wang Q, Chen H, Cui C, Xu X, Zhang Q, Zhang C. Prevalence and associated factors of infertility among 20–49 year old women in Henan

Province, China. Reprod Health. 2021;18(1):254. https://doi.org/10.1186/s12978-021-01298-2
4. Zhang TH, Liang JX, Long DL, Ma M, Chen LG, Lu DX, Jiang XH, Yang XS, Wang G. Tocolysis effects of traditional Chinese medicine and their effective components. World J Tradit Chin Med. 2022;8:188–98. https://doi.org/10.4103/wjtcm.wjtcm_30_21
5. Buckett W, Sierra S. The management of unexplained infertility: an evidence-based guideline from the Canadian fertility and andrology society. Reprod Biomed Online. 2019;39(4):633–40. https://doi.org/10.1016/j.rbmo.2019.05.023
6. Dabbagh Rezaeiyeh R, Mehrara A, Mohammad Ali Pour A, Fallahi J, Forouhari S. Impact of various parameters as predictors of the success rate of in vitro fertilization. Int J Fertil Steril. 2022;16(2):76–84. https://doi.org/10.22074/IJFS.2021.531672.1134
7. Avanzo M, Wei L, Stancanello J, Vallières M, Rao A, Morin O, Mattonen SA, El Naqa I. Machine and deep learning methods for radiomics. Med Phys. 2020;47(5):e185–202. https://doi.org/10.1002/mp.13678
8. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism. 2017;69S:S36–40. https://doi.org/10.1016/j.metabol.2017.01.011
9. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun (Lond). 2020;40(4):154–66. https://doi.org/10.1002/cac2.12012
10. Fernandez EI, Ferreira AS, Cecílio MHM, Chéles DS, de Souza RCM, Nogueira MFG, Rocha JC. 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(10):2359–76. https://doi.org/10.1007/s10815-020-01881-9
11. Chen S, Fang Z, Lu S, Gao C. Efficacy of regularized multitask learning based on SVM models. IEEE Trans Cybern. 2022;PP. https://doi.org/10.1109/TCYB.2022.3196308
12. Barkalov K, Shtanyuk A, Sysoyev A. A fast kNN algorithm using multiple space-filling curves. Entropy (Basel). 2022;24(6):767. https://doi.org/10.3390/e24060767
13. Rigatti SJ. Random forest. J Insur Med. 2017;47(1):31–9. https://doi.org/10.17849/insm-47-01-31-39.1
14. Martiello Mastelini S, Nakano FK, de Vens C, Leon Ferreira de Carvalho ACP. Online extra trees regressor. IEEE Trans Neural Netw Learn Syst. 2023;34(10):6755–67. https://doi.org/10.1109/TNNLS.2022.3212859
15. Moore A, Bell M, XGBoost A, Novel Explainable AI. Technique, in the prediction of myocardial infarction: a UK biobank cohort study. Clin Med Insights Cardiol. 2022;16:11795468221133611. https://doi.org/10.1177/11795468221133611
16. Xie X, Pu YF, Wang J. A fractional gradient descent algorithm robust to the initial weights of multilayer perceptron. Neural Netw. 2023;158:154–70. https://doi.org/10.1016/j.neunet.2022.11.018
17. Wang QQ, Yu SC, Qi X, Hu YH, Zheng WJ, Shi JX, Yao HY. Overview of logistic regression model analysis and application. Zhonghua Yu Fang Yi Xue Za Zhi. 2019;53(9):955–60. https://doi.org/10.3760/cma.j.issn.0253-9624.2019.09.018. Chinese.
18. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of support vector machine (svm) learning in cancer genomics. Cancer Genomics Proteom. 2018 Jan-Feb;15(1):41–51. https://doi.org/10.21873/cgp.20063
19. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inf Decis Mak. 2019;19(1):281. https://doi.org/10.1186/s12911-019-1004-8
20. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep. 2022;12(1):6256. https://doi.org/10.1038/s41598-022-10358-x
21. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Brief Bioinform. 2023;24(2):bbad002. https://doi.org/10.1093/bib/bbad002
22. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat. 2008;2(3):841–60. https://doi.org/10.1214/08-AOAS169
23. Scalzo F, Hamilton R, Asgari S, Kim S, Hu X. Intracranial hypertension prediction using extremely randomized decision trees. Med Eng Phys. 2012;34(8):1058–65. https://doi.org/10.1016/j.medengphy.2011.11.010
24. Wang R, Zhang J, Shan B, He M, Xu J. XGBoost machine learning algorithm for prediction of outcome in aneurysmal subarachnoid hemorrhage. Neuropsychiatr Dis Treat. 2022;18:659–67. https://doi.org/10.2147/NDT.S349956
25. Cai Y, Xie Y, Zhang S, Wang Y, Wang Y, Chen J, Huang Z. Prediction of postoperative recurrence of oral cancer by artificial intelligence model: multilayer perceptron. Head Neck. 2023;45(12):3053–66. https://doi.org/10.1002/hed.27533
26. Schober P, Vetter TR. Logistic regression in medical research. Anesth Analg. 2021;132(2):365–6. https://doi.org/10.1213/ANE.0000000000005247
27. Rienzi L, Cimadomo D, Vaiarelli A, Gennarelli G, Holte J, Livi C, Aura Masip M, Uher P, Fabozzi G, Ubaldi FM. Measuring success in IVF is a complex multidisciplinary task: time for a consensus? Reprod Biomed Online. 2021;43(5):775–8. https://doi.org/10.1016/j.rbmo.2021.08.012
28. Medenica S, Zivanovic D, Batkoska L, Marinelli S, Basile G, Perino A, Cucinella G, Gullo G, Zaami S. 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(12):2979. https://doi.org/10.3390/diagnostics12122979
29. 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(4):747–62. https://doi.org/10.1016/j.ogc.2023.09.003

Publisher's note