ACCSCIENCE PUBLISHING

Eurasian Journal of Medicine and Oncology

ORIGINAL RESEARCH ARTICLE

Bridging clinical knowledge and machine learning: Leveraging large language models to predict *in vitro* fertilization outcomes

Bo Fu[1,2], Di Liu[1,2], and Jing Bai[3]*

[1]Institute of Animal Husbandry, Heilongjiang Academy of Agricultural Sciences, Harbin, Heilongjiang, China

[2]Key Laboratory of Combining Farming and Animal Husbandry, Ministry of Agriculture and Rural Affairs, Harbin, Heilongjiang, China

[3]College of Medical Information Engineering, Heilongjiang University of Chinese Medicine, Harbin, Heilongjiang, China

Abstract

Introduction: *In vitro* fertilization (IVF) and frozen-thawed embryo transfer (FET) are vital components of assisted reproductive technology. However, predicting pregnancy outcomes remains challenging due to various biological and clinical factors. Recent advances in artificial intelligence (AI) and machine learning (ML) have shown the potential in offering innovative solutions for forecasting reproductive success.

Objective: This study explores the use of large language models, specifically ChatGPT-4o, to optimize ML models for predicting pregnancy outcomes in IVF.

Methods: The clinical dataset comprised 1061 IVF patients who underwent FET from 2014 to 2017, including variables such as age, body mass index, infertility duration, endometrial thickness, and serum beta-human chorionic gonadotrophin (β-HCG) levels on the 7th day after FET. ChatGPT-4o was tasked with preprocessing the data, evaluating several ML models, and optimizing performance.

Results: The random forest model emerged as the best-performing model, achieving an accuracy of 85.45% and an area under the receiver operating characteristic curve of 0.8287 after applying the optimal threshold of 0.548, indicating strong predictive capability. Feature importance analysis revealed that serum β-HCG levels on the 7th day after FET were the most influential predictor of pregnancy outcomes. Despite these promising results, the study noted potential overfitting, likely due to the limited training dataset, a constraint largely attributable to the computational limitations of ChatGPT-4o.

Conclusion: ChatGPT-4o shows potential in enhancing ML models in IVF outcome prediction. While AI-driven models can significantly aid clinical decision-making, clinicians should maintain a central role in patient outcome predictions. Future work will focus on improving model generalization with larger datasets and enhanced computational resources.

Keywords: In vitro fertilization; Pregnancy prediction; Large language models; Machine learning; Assisted reproductive technology

***Corresponding author:**
Jing Bai
(baijing@hljucm.edu.cn)

Citation: Fu B, Liu D, Bai J. Bridging clinical knowledge and machine learning: Leveraging large language models to predict *in vitro* fertilization outcomes. *Eurasian J Med Oncol*. 2025;9(4):168-177. doi: 10.36922/EJMO025120058

Received: March 18, 2025

Revised: April 13, 2025

Accepted: May 6, 2025

Published online: June 3, 2025

1. Introduction

Assisted reproductive technology (ART) has been widely applied worldwide to help infertile couples achieve their reproductive goals. However, due to the complex interplay of biological, environmental, and clinical factors, the success rate of ART remains highly uncertain. In recent years, the development of electronic health records (EHR) and the application of advanced data analysis methods such as machine learning (ML) have provided strong support for ART. These advancements hold significant clinical value, particularly in enhancing predictive accuracy for ART outcomes, optimizing clinical decision-making, personalized medicine in reproductive health, and data-driven approaches to improve ART success rates.[1-4]

ML, as a subset of artificial intelligence (AI), has significantly advanced clinical decision-making, particularly in areas such as diagnostics and personalized treatment. Clinical decision support systems powered by ML algorithms, such as decision trees and neural networks, assist in predicting patient outcomes, suggesting treatment options, and monitoring disease progression.[5] However, the development, implementation, and validation of ML models in healthcare require a highly specialized and intricate understanding of both clinical needs and advanced computational methods. While the process of creating these models involves gathering vast amounts of data, selecting the right algorithms, and ensuring high-quality training, it is often challenging for clinicians and researchers due to the technical expertise required for effective model development.[6]

Fortunately, large language models (LLMs), such as ChatGPT, have greatly bridged the gap between human-computer interactions, enabling non-computer professionals to optimize ML models more effectively. LLMs' ability to understand and generate natural language allows users with little or no technical background to interact with complex ML systems through simple conversational prompts. This capability significantly reduces the learning curve, making it easier for non-experts to manage model optimization tasks, such as selecting algorithms and tuning hyperparameters, without requiring deep computational knowledge.[7] Furthermore, the transparency of these models enables users to receive detailed explanations of the decisions made during the optimization process, making it easier to understand the steps involved and facilitating collaboration between human and machine.[8] LLMs, such as ChatGPT, represent a significant leap in the field of ML, offering more advanced capabilities in handling complex natural language processing tasks. By leveraging vast datasets and deep learning architectures, LLMs expand the scope of traditional ML by enhancing text generation, understanding, and reasoning across diverse domains.[9,10] Currently, LLMs are making significant strides in healthcare by enhancing the efficiency of medical processes, particularly in areas like clinical decision support, diagnosis assistance, and medical documentation. Owing to LLMs' ability to process large-scale datasets, identify complex patterns, and provide predictive analytics, LLMs assist with diagnosing diseases, generating medical reports, handling EHR, and providing personalized healthcare advice.[11-16] Interestingly, LLMs, such as ChatGPT, even outperformed human candidates in a virtual objective structured clinical examination in obstetrics and gynecology, achieving higher average scores (77.2%) compared to the human candidates (73.7%), which indicates the potential of LLMs to successfully perform complex clinical reasoning tasks, potentially revolutionizing medical education and assessments.[17] Considering all these factors, we then hypothesize that in the context of *in vitro* fertilization (IVF), LLM-driven methodologies may demonstrate significant potential to enhance understanding of embryogenesis and clinical outcomes.

At present, the validity and reliability of LLMs in advanced data processing and analysis, particularly in predicting clinical pregnancy outcomes in IVF, have not been thoroughly evaluated. This study uses ChatGPT-4o as a representative example to investigate LLMs, as a state-of-the-art chatbot, in terms of their performance in data processing, ML model optimization, and workflow integration, with the ultimate goal of exploring future applications for predicting IVF outcomes.

2. Materials and methods

2.1. Dataset overview

ChatGPT-4o (GPT-4o, https://chatgpt.com/, OpenAI, Inc., United States of America) was tasked to analyze a publicly available dataset from the Reproductive and Genetic Center of the Affiliated Hospital of Shandong University of Traditional Chinese Medicine between 2014 and 2017.[18] A total of 2,582 patients underwent IVF and frozen-thawed embryo transfer (FET). The study population was restricted to patients who were under 38 years old, had a body mass index (BMI) between 18.5 and 25, and had complete follow-up records until delivery. Eligible patients had infertility primarily due to tubal factors, possessed at least four available embryos, and received transferred embryos with a morphological score of 6 or higher according to the Istanbul Consensus criteria. Patients who underwent preimplantation genetic screening or diagnosis were excluded. A maximum of two embryos were transferred per cycle. Additional exclusion criteria included non-pregnancy, abnormal fetal

development identified via color Doppler ultrasound, and spontaneous abortion during the second trimester. Patients who received exogenous beta-human chorionic gonadotropin (β-HCG) for luteal support were also excluded. All patients in this study received standard luteal phase support with Progynova and progesterone after FET. This standardization of treatment was applied to avoid any bias related to concomitant treatments. After screening the clinical data, 1,061 patients' clinical data were included in the analysis.[18] Single and twin ongoing pregnancies were classified as ongoing pregnancy, while ectopic pregnancy, biochemical pregnancy, and single early spontaneous abortion were classified as pregnancy failure, as shown in Table S1. Our ultimate goal is to use ChatGPT-4o to train and optimize ML models, exploring the feasibility of predicting pregnancy outcomes by analyzing clinical indicators such as age (years), infertility (years), BMI (kg/m^2), endometrial thickness on FET day (mm), serum β-HCG levels on the 7th day after FET (mIU/mL), basal follicle-stimulating hormone (U/L), basal luteinizing hormone (U/L), and basal estradiol (pg/mL). The training/testing split was set to 80/20, as shown in Table S2. In this study, stratified sampling was used, and it is particularly important in this dataset because "clinical pregnancy" is likely to be imbalanced (i.e., there may be more positive cases than negative ones). Stratified sampling ensures that both the training and test sets maintain the same class distribution as the original dataset, preventing bias and improving generalization for underrepresented cases.

2.2. Instructing LLMs

The following instructions were given to ChatGPT-4o: (i) Preprocess the dataset by splitting it into training and test sets with stratified sampling, while ensuring balanced proportions of successful and unsuccessful IVF outcomes; (ii) Train and evaluate multiple ML models, compare accuracy, F1-score, precision, and recall, and then select the best-performing model based on performance metrics; (iii) Predict pregnancy outcomes using the trained model, generate a receiver operating characteristic (ROC) curve, then analyze sensitivity, specificity, and Matthew's correlation coefficient (MCC) for further performance evaluation; (iv) Conduct feature importance analysis, apply the Mann–Whitney *U*-test to identify statistically significant differences (p<0.05) between the pregnancy and pregnancy failure groups, and present the top five influential features in a horizontal bar chart. The details of all the prompts in this study are shown in Table S3.

3. Results

ChatGPT-4o generated systematically organized responses to the specified prompt. First, ChatGPT-4o split the data into training and test sets based on the "set" column, then analyzed several ML models and assessed their performance. For classification tasks, ChatGPT-4o evaluated four commonly used ML models: logistic regression, random forest, support vector machine (SVM), and k-nearest neighbors. Subsequently, ChatGPT-4o compared their performance in a horizontal bar chart, as shown in Figure 1. The chart illustrates that the random forest model demonstrated the best overall performance, with the highest F1-score (0.8825), precision (0.8476), recall (0.9272), and accuracy (0.8263). Random forest effectively handles non-linearity, feature interactions, and noise. In addition, its ability to determine feature importance makes this model well-suited for complex clinical datasets. These characteristics may contribute to its superior performance.

Based on the above results, ChatGPT-4o selected the random forest model for training. The model is configured with 300 estimators, a minimum sample split of 10, a minimum sample leaf of 5, and a maximum depth of 20. The random forest model was then trained and saved on a local computer. Next, using the trained random forest model, ChatGPT-4o predicted the pregnancy outcome for the test set, while also generating both the predicted labels and the probabilities, which can later be used for calculating the area under the curve (AUC), as shown in Tables S4 and S5. To analyze the accuracy of the predictions, ChatGPT-4o merged the predictions with the true outcomes. The formula is as follows:

$$\text{Accuracy} = \frac{\text{Number of predictions}}{\text{Total number of correct predictions}} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(I)}$$

where:

TP = True positives (predicted pregnancy = 1, actual pregnancy = 1)

TN = True negatives (predicted no pregnancy = 0, actual no pregnancy = 0)

FP = False positives (predicted pregnancy = 1, actual no pregnancy = 0)

FN = False negatives (predicted no pregnancy = 0, actual pregnancy = 1)

The area under the receiver operating characteristic curve (AUROC) score evaluates the model's ability to distinguish between the two classes (pregnancy and pregnancy failure) across various thresholds. The formula is as follows:

$$AUROC = \int_0^1 TPR(FPR)dFPR \quad \text{(II)}$$

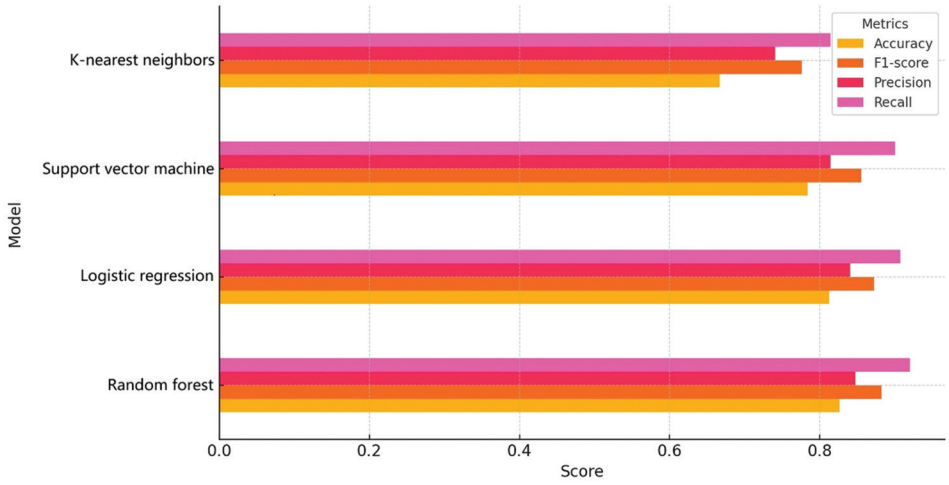

Figure 1. The comparison of machine learning models for predicting clinical pregnancy. This bar chart compares the accuracy of four different classification models: k-nearest neighbors, logistic regression, support vector machine, and random forest. The random forest model achieved the highest accuracy, while the k-nearest neighbors model had the lowest performance among the four models.

where:
TPR (true positive rate) = TP/(TP + FN)
FPR (false positive rate) = FP/(FP + TN)

The AUROC was calculated by plotting the ROC curve and calculating the area beneath it. ChatGPT-4o reported an AUROC of 0.83 and an accuracy of 83.57% on the test set. These results indicate that the model performs well, especially in distinguishing between the pregnancy and pregnancy failure outcomes. Remarkably, ChatGPT-4o even generated the ROC curve, as shown in Figure 2. In addition, ChatGPT-4o performed additional metrics to analyze sensitivity, specificity, and MCC. MCC ranges from −1 (worst performance) to +1 (perfect performance), with 0 indicating random predictions. The results reported a sensitivity (recall) of 92.72%, specificity of 61.29%, and MCC of 0.58. The formulas for calculation are as follows:

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

$$Specificity = \frac{TN}{TN + FP} \quad \text{(IV)}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{(V)}$$

The model correctly identified 92.72% of positive cases, indicating a relatively strong ability in predicting pregnancies. However, a specificity of 61.29% indicated that the model incorrectly classified 38.71% of failed pregnancies as successes, leading to a high false positive rate. This is likely due to class imbalance, and an optimized decision threshold may overcome this problem. Precision-recall curve analysis was then used to optimize the threshold and identify an optimal value of 0.548, as shown in Figure 3.

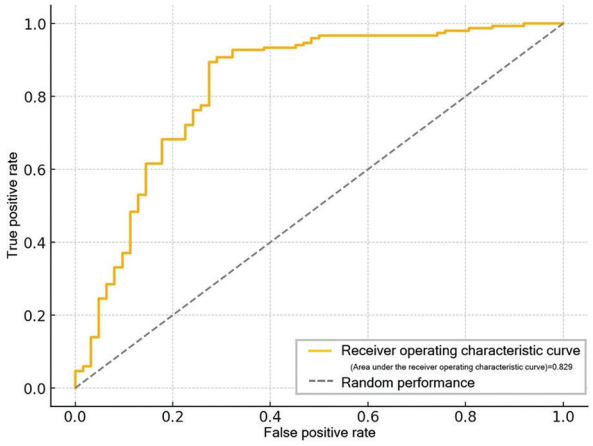

Figure 2. The receiver operating characteristic curve of true positive rate (y-axis) against the false positive rate (x-axis). The yellow curve represents the model's performance, with the AUROC value shown in the legend. The dashed gray diagonal line represents random performance, which corresponds to an AUROC of 0.5. The model's AUROC is significantly higher, indicating stronger predictive power.
Abbreviation: AUROC: Area under the receiver operating characteristic curve.

After applying 0.548 as the optimal decision threshold, the model's accuracy improved from 83.57% to 85.45%, F1-score increased from 88.89% to 90.03%, and precision rose from 85.37% to 87.50%, indicating a better balance between false positives and false negatives. Recall and AUROC remained consistent at 92.72% and 82.87%,

respectively, confirming that the model's ranking ability was unaffected by the threshold adjustment. Additional performance metrics were evaluated after applying the optimized decision threshold. This resulted in improved specificity and MCC while maintaining high sensitivity, leading to a better balance between precision and recall, as shown in Table 1. Finally, to evaluate the contributions of key features on the predictions, ChatGPT-4o applied feature importance (model-specific) on the top five most influential features, and ranked them by their overall impact in a horizontal bar chart, as shown in Figure 4. ChatGPT-4o revealed that among the top five features, serum β-HCG levels on the 7th day after FET are significantly different between the two groups (pregnancy vs. pregnancy failure), with $p<0.05$ (approximately 2.52×10^{-56}), as shown in Table 2.

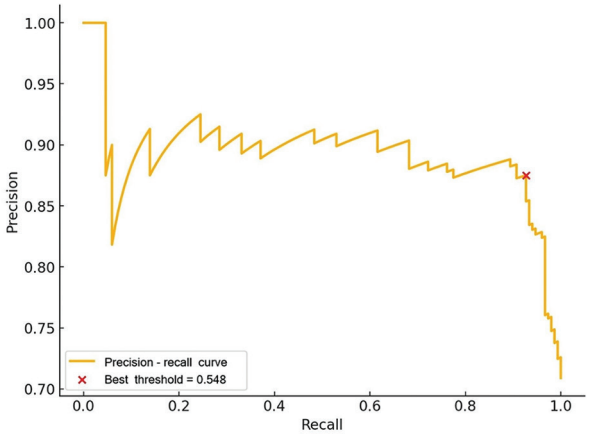

Figure 3. Precision-recall curve with optimized decision threshold illustrating the trade-off between precision and recall for the trained model. The optimal decision threshold is determined to be 0.548, as indicated by the red marker, which achieves the best balance between precision and recall.

4. Discussion

Our results indicate that LLMs like ChatGPT-4o have the potential for data processing and ML model training, thereby providing a new option for predicting clinical pregnancy from IVF treatment.

Although clinicians remain central to medical decision-making, the integration of EHR and ML has advanced rapidly in clinical diagnosis and prediction, and the emergence of MLs provides a powerful tool, enhancing

Table 1. Comparison of model performance before and after applying the optimized decision threshold (0.548)

Metric	Before optimization	After optimization
Sensitivity (recall) (%)	92.72	92.72
Specificity (%)	61.29	67.74
Matthews correlation coefficient	0.5829	0.6352

Table 2. Comparison of features between pregnancy and pregnancy failure

Feature	Pregnancy mean±SE	Pregnancy failure mean±SE	*p*-value
Serum β-HCG levels on the 7th day after FET (mIU/mL)	15.30±0.44	5.53±0.38	2.52×10^{-56}
Endometrial thickness on FET day (mm)	10.95±0.07	11.03±0.11	0.52
bFSH (U/L)	6.19±0.07	6.54±0.14	0.17
BMI (kg/m^2)	21.77±0.08	21.76±0.13	0.58
bLH (U/L)	5.67±0.10	5.51±0.15	0.30

Abbreviations: β-HCG: Beta-human chorionic gonadotrophin; bFSH: Basal follicle-stimulating hormone; bLH: Basal luteinizing hormone; BMI: Body mass index; FET: Frozen-thawed embryo transfer; SE: Standard error.

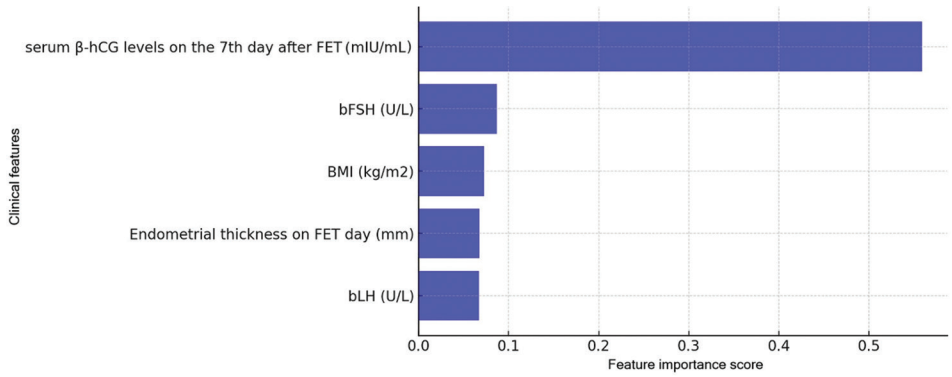

Figure 4. The top five most influential features affecting the success rate of *in vitro* fertilization. The x-axis represents the feature importance score, which quantifies the contribution of each feature to the model's predictive capability. The y-axis lists the clinical features ranked by their significance.
Abbreviations: β-HCG: Beta-human chorionic gonadotrophin; bFSH: Basal follicle-stimulating hormone; bLH: Basal luteinizing hormone; BMI: Body mass index; FET: Frozen-thawed embryo transfer.

diagnostic accuracy and predictive performance.[19-22] Despite the significant potential demonstrated by ML models in clinical diagnosis and prediction, the applications of these models remain challenging for individual clinicians without a background in computer science. The lack of user-friendly interfaces for ML models is a common issue.[23] Non-experts often encounter technical barriers, such as the need for specialized knowledge in data preparation, model training, and result interpretation. These limitations constrain the broader integration of ML in clinical settings.[22,24,25]

Fortunately, the emergence of LLMs represented by ChatGPT-4o has brought a breakthrough to the above-mentioned issues. LLMs capable of processing and generating text, such as ChatGPT-4o, have distinct advantages for non-experts, particularly in terms of accessibility, ease of use, and broad application across various fields. LLMs feature intuitive interfaces that enable non-experts to effectively utilize advanced ML models without requiring specialized technical expertise, making them more accessible to a wider audience.[26] LLMs also offer several advantages for non-experts, including advanced natural language understanding for easy interaction and content generation,[27] the ability to provide contextually accurate feedback with explanations in fields, such as education and healthcare,[28] and enhanced decision support by analyzing complex data and providing actionable suggestions in areas like medicine.[29] Regarding ML model training, LLMs like ChatGPT-4o are designed with intuitive user interfaces that facilitate easy interaction. Non-experts can leverage these models for tasks such as building ML models without requiring advanced programming skills. In addition, LLMs streamline the process of data preparation and preprocessing by automatically handling many of the complexities involved in cleaning and structuring data. This reduces the manual work typically required in traditional ML pipelines, allowing non-experts to avoid common data-related pitfalls. Most importantly, LLMs like ChatGPT-4o improve interpretability by generating text-based explanations for model predictions, making it easier for non-experts to understand, which is particularly beneficial in fields like healthcare and business, where decision-making should be transparent.

In this study, we explore the use of ChatGPT-4o to train and optimize ML models for predicting clinical pregnancy outcomes based on data from IVF and FET patients. ChatGPT-4o was tasked with preprocessing the dataset, performing data splitting, and evaluating multiple ML models, ultimately selecting random forest as the best-performing model. The entire process was seamlessly executed through prompt-based interactions and was completed automatically, without human intervention. The model achieved an AUROC of 0.83 and an accuracy of 85.45%, demonstrating strong predictive capability.

In this context, the specificity was low during the evaluation of the model's classification performance, particularly when analyzing key classification metrics (sensitivity, specificity, and MCC) after making predictions on the test set. This issue became evident when assessing false positives, where a significant portion of failed pregnancies were misclassified as successes. A low specificity often occurs when the dataset has significantly more success cases than failures, causing the model to learn a biased pattern that favors predicting success, which increases false positives. In addition, if the decision threshold is too low (e.g., 0.5), it struggles to confidently classify failures, leading to misclassification and reduced specificity. The precision-recall curve is useful for evaluating the model's performance, particularly in imbalanced classification problems, and helps identify the threshold that minimizes false positives while maintaining high recall. In this context, we use precision-recall curve analysis to optimize the classification threshold, enabling a balanced precision and recall and preventing the model from favoring one class disproportionately. Finding the optimal threshold involves selecting a point that minimizes false positives while maintaining high predictive performance, leading to a better trade-off between sensitivity and specificity. After adjusting the decision threshold to 0.548, the specificity increased from 61.29% to 67.74%, enhancing the model's ability to correctly identify failed pregnancies.

The 'black-box' nature of ML models means they make predictions without providing clear, interpretable explanations for the decisions made.[30] This lack of transparency poses a significant barrier to their acceptance by clinicians and patients, especially when making decisions that could impact patient health. After guiding ChatGPT-4o through a series of prompts to solve this problem, we found that ChatGPT-4o could identify several key features contributing to the predictions. ChatGPT-4o considered "serum β-HCG levels on the 7th day after FET" as the most influential feature. β-HCG is a glycoprotein hormone produced by the trophoblast cells of the placenta following successful embryo implantation. This hormone plays a vital role in maintaining early pregnancy and is essential for embryo implantation and placenta development. Upon implantation, the embryo signals the mother's body to produce β-HCG, which supports the corpus luteum in the ovary to secrete progesterone, a hormone crucial for maintaining pregnancy.[31] Notably, numerous studies have demonstrated that serum β-HCG levels on the 7th day can accurately predict pregnancy outcomes.[32-36] These findings further indicate that by leveraging transparent, data-driven insights, the model can accurately identify critical factors

influencing IVF outcomes. This demonstrates the model's potential for clinical application, improving decision-making processes in reproductive health.

It is important to note that while the fandom forest model trained in this study achieves a good balance between sensitivity and specificity, there remains a potential risk of overfitting. Due to the limited computational resources available in ChatGPT-4o, we have only provided a small sample of the training set to ensure a successful model training task. A small training set may lead to good performance on the training data but hinder the model's ability to generalize to new, unseen data. The model may memorize noise and irrelevant details from the training data rather than learning the actual patterns, leading to overfitting.[37] An AUROC of 0.83 is commendable but not perfect, suggesting possible overfitting to the training data. Although increasing the sample size can help mitigate the issues to some extent, it still relies on OpenAI to provide more computational resources to the users. Taken together, the innovative aspects and the challenges encountered in our study highlight both the potential and limitations of the approach. On one hand, a key strength of this study is the novel applications of LLMs, specifically ChatGPT-4o, to optimize ML models for predicting IVF outcomes. This approach bridges the gap between clinical knowledge and advanced AI technology, making it more accessible to clinicians with limited computational expertise. On the other hand, the relatively small dataset, constrained by the computational resources of ChatGPT-4o, introduces the risk of overfitting, which can impact the model's ability to generalize to new, unseen data. The limited computational resources available during the study hindered the use of larger models and more complex algorithms, which could have further enhanced the model's generalizability. Future work should focus on expanding the dataset, incorporating data from multiple centers, and utilizing more advanced computational resources. These would help to mitigate overfitting and improve the generalizability of the model.

While the use of LLMs like ChatGPT-4o in optimizing ML models for predicting IVF outcomes demonstrates significant promise, incorporating more indicators during clinical data collection will enhance model accuracy and decision-making. For instance, dietary and lifestyle interventions, such as reducing inflammation and adjusting nutrient intake, may help alleviate endometriosis symptoms and could be significant in predicting the outcomes of IVF treatments in women with endometriosis by potentially improving fertility conditions and symptom management.[38] Inositol supplementation, particularly myo-inositol, has shown potential benefits in improving ovarian function and oocyte quality, which may enhance IVF outcomes, especially for women with conditions such as polycystic ovary syndrome or poor oocyte quality.[39] In addition, early reproductive counseling and fertility preservation options can significantly influence IVF outcomes, particularly for those considering parenthood after cancer treatment.[40] Considering the above factors, the future inclusion of these clinical data could potentially enhance the accuracy of ML models based on LLMs for predicting IVF outcomes.

In addition to using ML models to predict pregnancy outcomes in IVF, this approach can also be applied to the prediction of other reproductive diseases. ML models can also be employed to analyze complex genetic data from embryos, enabling precise identification of chromosomal abnormalities, genetic mutations, and other potential issues before the embryo is implanted. For instance, ML models, such as SVM and random forests, have been used to predict the likelihood of successful embryo implantation and identify embryos with optimal genetic profiles. These techniques offer several advantages over traditional methods, including the ability to handle large datasets, reduce human error, and identify patterns in the data that may not be immediately obvious.[41] Additionally, ML models can also assist in identifying ectopic pregnancy by analyzing clinical data such as serum markers, ultrasound findings, and patient history. For instance, a study by Rueangket *et al.*[42] employed neural networks and logistic regression models to predict ectopic pregnancy using 22 clinical features. Their model achieved an impressive AUROC of 0.898, highlighting the potential of ML to outperform traditional diagnostic methods.[42] Moreover, Aljameel *et al.*[43] developed an automated miscarriage prediction system using a gradient boosting classifier, which showed a high accuracy of 93.4% in predicting the risk of miscarriage in early pregnancy. Similarly, Amitai *et al.*[44] used time-lapse imaging data and ML models like XGBoost to predict the risk of first-trimester miscarriage, achieving an AUC of 0.68 to 0.69. These results underscore the utility of ML models in enhancing early miscarriage prediction, which could lead to timely interventions and better pregnancy outcomes.

5. Conclusion

This study explores the applications of LLMs, such as ChatGPT-4o, in optimizing ML models to predict clinical pregnancy outcomes based on IVF data. The study demonstrates that ChatGPT-4o can effectively preprocess data, evaluate multiple models, and optimize performance, achieving an accuracy of 85.45% and an AUROC of 0.83, highlighting its significant potential for enhancing IVF outcome prediction. However, for medical applications, an accuracy of 85.45% is insufficient. Due to the limited computational power of ChatGPT-4o, which restricts the

size of the training set, the trained model still carries the risk of overfitting. We speculate that increasing model size will improve performance, enabling the model to generate more accurate, contextually relevant, and diverse responses across a wide range of topics.[45] Thus, at this stage, it is advisable for clinicians to maintain a leading role in predicting clinical pregnancy outcomes, with AI-trained ML models serving as a supportive tool rather than taking the lead.

Acknowledgments

The authors extend their gratitude to the Reproductive and Genetic Center of the Affiliated Hospital of Shandong University of Traditional Chinese Medicine for providing access to the dataset.

Funding

This work was funded by Heilongjiang Provincial Research Institutes Research Business Fund Project (CZKYF2024-1-B006).

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Bo Fu, Di Liu
Formal analysis: Bo Fu
Methodology: Jing Bai
Writing – original draft: Bo Fu, Jing Bai
Writing – review & editing: Bo Fu, Jing Bai

Ethics approval and consent to participate

The dataset used in this study was obtained from a public database and does not require approval from the ethics committee.

Consent for publication

Not applicable.

Availability of data

The data is available at: https://datadryad.org/dataset/doi:10.5061/dryad.8931zcrnj

References

1. Li B, Chen H, Lin X, Duan H. Multimodal learning system integrating electronic medical records and hysteroscopic images for reproductive outcome prediction and risk stratification of endometrial injury: A multicenter diagnostic study. *Int J Surg*. 2024;110(6):3237-3248.

   doi: 10.1097/JS9.0000000000001241

2. Henderson I, Rimmer MP, Keay SD, *et al*. Predicting the outcomes of assisted reproductive technology treatments: A systematic review and quality assessment of prediction models. *F S Rev*. 2021;2(1):1-10.

   doi: 10.1016/j.xfnr.2020.11.002

3. Huang C, Xiang Z, Zhang Y, *et al*. Using deep learning in a monocentric study to characterize maternal immune environment for predicting pregnancy outcomes in the recurrent reproductive failure patients. *Front Immunol*. 2021;12:642167.

   doi: 10.3389/fimmu.2021.642167

4. Melli B, Morini D, Spaggiari G, *et al*. P-032 sperm parameters can predict the success of assisted reproductive technology. Single-center and retrospective analysis of assisted reproductive technology cycles from 1992 to 2020. *Hum Reprod*. 2022;37:deac107.03032.

   doi: 10.1093/humrep

5. Westbye HJ, Moltu C, McAleavey AA. eXplainable AI for routine outcome monitoring and clinical feedback. *Counsell Psychother Res*. 2025;25(1):e12764.

   doi: 10.1002/capr.12764

6. Chen PH, Liu Y, Peng L. How to develop machine learning models for healthcare. *Nat Mater*. 2019;18(5):410-414.

   doi: 10.1038/s41563-019-0345-0

7. Widyasari R, Lo D, Liao L. *Beyond ChatGPT: Enhancing Software Quality Assurance Tasks with Diverse LLMs and Validation Techniques*. [arXiv Preprint]; 2024.

   doi: 10.48550/arXiv.2409.01001

8. Sadik AR, Ceravola A, Joublin F, Patra J. *Analysis of Chatgpt on Source Code*. [arXiv Preprint]; 2023.

   doi: 10.48550/arXiv.2306.00597

9. Telenti A, Auli M, Hie BL, Maher C, Saria S, Ioannidis JP. Large language models for science and medicine. *Eur J Clin Invest*. 2024;54(6):e14183.

   doi: 10.1111/eci.14183

10. Zheng Y, Koh HY, Ju J, *et al*. *Large Language Models for Scientific Synthesis, Inference and Explanation*. [arXiv Preprint]; 2023.

    doi: 10.48550/arXiv.2310.07984

11. Silver DH, Feder M, Gold-Zamir Y, *et al*. *Data-Driven Prediction of Embryo Implantation Probability Using IVF Time-Lapse Imaging*. [arXiv Preprint]; 2020.

    doi: 10.48550/arXiv.2006.01035

12. Sun L, Li J, Zeng S, *et al*. Artificial intelligence system for outcome evaluations of human *in vitro* fertilization-derived embryos. *Chin Med J (Engl)*. 2024;137(16):1939-1949.

    doi: 10.1097/CM9.0000000000003162

13. Liu R, Bai S, Jiang X, *et al*. Multifactor prediction of embryo transfer outcomes based on a machine learning algorithm. *Front Endocrinol (Lausanne)*. 2021;12:745039.

doi: 10.3389/fendo.2021.745039

14. Nazi ZA, Peng W. Large language models in healthcare and medical domain: A review. *Informatics*. 2024;11(3):57.

doi: 10.3390/informatics11030057

15. Chen S. Potential applications and safety of large language models in healthcare. *Interdiscip Humanit Commun Stud*. 2024;1(6):6.

doi: org/10.61173/f578jp05

16. Rezgui K. Large language models for healthcare: Applications, models, datasets, and challenges. In: *2024 10th International Conference on Control, Decision and Information Technologies*. Maharashtra: CoDIT; 2024. P. 2366-2371.

doi: 10.1109/CoDIT62066.2024.10708253

17. Li SW, Kemp MW, Logan SJ, *et al*. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. *Am J Obstet Gynecol*. 2023;229(2):172.e1-12.

doi: 10.1016/j.ajog.2023.04.020

18. Yuan L, Yu L, Sun Z, *et al*. Association between 7-day serum β-hCG levels after frozen-thawed embryo transfer and pregnancy outcomes: A single-centre retrospective study from China. *BMJ Open*. 2020;10(10):e035332.

doi: 10.1136/bmjopen-2019-035332

19. Kiser AC, Eilbeck K, Ferraro JP, Skarda DE, Samore MH, Bucher B. Standard vocabularies to improve machine learning model transferability with electronic health record data: Retrospective cohort study using health care-associated infection. *JMIR Med Inform*. 2022;10(8):e39057.

doi: 10.2196/39057

20. Ng K, Stewart WF, DeFilippi C, *et al*. Data driven modeling of electronic health record data to detect pre-diagnostic heart failure in primary care. *Circulation*. 2015;132(Suppl 3):A17713

doi: 10.1161/circ.132.suppl_3.17713

21. Hong D, Fort D, Shi L, Price-Haywood EG. Electronic medical record risk modeling of cardiovascular outcomes among patients with type 2 diabetes. *Diabetes Ther*. 2021;12(7):2007-2017.

doi: 10.1007/s13300-021-01096-w

22. Ganesan R, Habraken SC, Van De Vosse FN, Huberts W. Explainable machine learning based prediction of severity of heart failure using primary electronic health records. *Stud Health Technol Inform*. 2024;316:542-546.

doi: 10.3233/SHTI240471

23. Stevens CA, Lyons AR, Dharmayat KI, *et al*. Ensemble machine learning methods in screening electronic health records: A scoping review. *Digit Health*. 2023;9:20552076231173225.

doi: 10.1177/20552076231173225

24. Wang J, Luo J, Ye M, *et al*. *Recent Advances in Predictive Modeling with Electronic Health Records*. [arXiv Preprint]; 2024.

doi: 10.48550/arXiv.2402.01077

25. Iwagami M, Inokuchi R, Kawakami E, *et al*. Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study. *PLOS Digit Health*. 2024;3(8):e0000578.

doi: 10.1371/journal.pdig.0000578

26. Abu-Rayyash H. Revolutionizing translator training through human-ai collaboration: Insights and implications from integrating gpt-4. *Curr Trends Transl Teach Learn E*. 2023;10:259-301.

doi: 10.51287/cttl20239

27. Takayanagi T, Takamura H, Izumi K, Chen CC. *Beyond Turing Test: Can GPT-4 Sway Experts' Decisions*? [arXiv Preprint]; 2024.

doi: 10.48550/arXiv.2409.16710

28. Mondal A, Naskar A. *Artificial Intelligence in Diabetes Care: Evaluating GPT-4's Competency in Reviewing Diabetic Patient Management Plan in Comparison to Expert Review*. [medRxiv Preprint]; 2024.

doi: 10.1101/2024.04.12.24305732

29. Nori H, King N, McKinney SM, Carignan D, Horvitz E. *Capabilities of Gpt-4 on Medical Challenge Problems*. [arXiv Preprint]; 2023.

doi: 10.48550/arXiv.2303.13375

30. Björklund A, Henelius A, Oikarinen E, Kallonen K, Puolamäki K. Explaining any black box model using real data. *Front Comput Sci*. 2023;5:1143904.

doi: 10.3389/fcomp.2023.1143904

31. Schumacher A, Zenclussen AC. Human chorionic gonadotropin-mediated immune responses that facilitate embryo implantation and placentation. *Front Immunol*. 2019;10:2896.

doi: 10.3389/fimmu.2019.02896

32. Sung N, Kwak-Kim J, Koo HS, Yang KM. Serum hCG-β levels of postovulatory day 12 and 14 with the sequential application of hCG-β fold change significantly increased predictability of pregnancy outcome after IVF-ET cycle. *J Assist Reprod Genet*. 2016;33:1185-1194.

doi: 10.1007/s10815-016-0744-y

33. Liu Y, Liu Y, Li X, Jiao X, Zhang R, Zhang J. Predictive value of serum β-hCG for early pregnancy outcomes among women with recurrent spontaneous abortion. *Int J Gynecol Obstet*. 2016;135(1):16-21.

doi: org/10.1016/j.ijgo.2016.03.007

34. Li Y, Xiang YG, Zhang M, Ma LY, Tan L, Zhao DM. Prediction of pregnancy outcome by serum β-hCG and progesterone on the fourteenth day after IVF-ET. *J Int Reprod Health Family Plan*. 2013;32(1):9.

35. Wang L, Jiang Y, Shen H, *et al*. Independent value of serum β-human chorionic gonadotropin in predicting early pregnancy loss risks in IVF/ICSI cycles. *Front Immunol*. 2022;13:992121.

doi: 10.3389/fimmu.2022.992121

36. Ozer G. Initial β-hCG levels and 2-day-later increase rates effectively predict pregnancy outcomes in single blastocyst transfer in frozen-thawed or fresh cycles: A retrospective cohort study. *Medicine (Baltimore)*. 2023;102(42):e35605.

doi: 10.1097/MD.0000000000035605

37. Ying X. An overview of overfitting and its solutions. *J Phys Conf Ser*. 2019;1168:022022.

doi: 10.1088/1742-6596/1168/2/022022

38. Habib N, Buzzaccarini G, Centini G, *et al*. Impact of lifestyle and diet on endometriosis: A fresh look to a busy corner. *Prz Menopauzalny*. 2022;21(2):124-132.

doi: 10.5114/pm.2022.116437

39. Gullo G, Carlomagno G, Unfer V, D'Anna R. Myo-inositol: From induction of ovulation to menopausal disorder management. *Minerva Ginecol*. 2015;67(5):485-486.

40. Zaami S, Melcarne R, Patrone R, *et al*. Oncofertility and reproductive counseling in patients with breast cancer: A retrospective study. *J Clin Med*. 2022;11(5):1311.

doi: 10.3390/jcm11051311

41. Greco E, Litwicka K, Minasi MG, Cursio E, Greco PF, Barillari P. Preimplantation genetic testing: Where we are today. *Int J Mol Sci*. 2020;21(12):4381.

doi: 10.3390/ijms21124381

42. Rueangket P, Rittiluechai K, Prayote A. Predictive analytical model for ectopic pregnancy diagnosis: Statistics vs. Machine learning. *Front Med (Lausanne)*. 2022;9:976829.

doi: 10.3389/fmed.2022.976829

43. Aljameel SS, Aljabri M, Aslam N, *et al*. An automated system for early prediction of miscarriage in the first trimester using machine learning. *Comput Mater Contin*. 2023;75(1):1291-1304.

doi: 10.32604/cmc.2023.035710

44. Amitai T, Kan-Tor Y, Yuval O, *et al*. Embryo classification beyond pregnancy: Early prediction of first trimester miscarriage using machine learning. *J Assist Reprod Genet*. 2023;40(2):309-322.

doi: 10.1007/s10815-022-02619-5

45. Wangsa K, Karim S, Gide E, Elkhodr M. A systematic review and comprehensive analysis of pioneering AI chatbot models from education to healthcare: ChatGPT, Bard, Llama, Ernie and Grok. *Future Int*. 2024;16(7):219.

doi: 10.3390/fi16070219