Prediction of male infertility risk and risk profiles based on questionnaire and machine learning

Peihao Wu[1, 2, 3], MS; Sia Florence Koroma[1, 2], MS; Abdoulie Dukureh[1, 2],MPH; Balansama Marah[1, 2], MPH; Jing Wang[1, 2], MS; Jialin Feng[1, 2], MPH; Yan Yuan[1, 2], MPH; Yixuan Yan[1, 2], MS; Feng Wu[1, 2], MPH; Qi Liu[1, 2], MS; Ying Li[4], MD; Jinqi Ma[4], MD; Feng Pan[5], PhD; Jun Ye[3], MD; Chuncheng Lu[1, 2], PhD; Xinru Wang[1, 2], PhD; Jianquan Chen[6], MD; Qiuqin Tang[7], MS; Rong Ju[8, *], MD; Wei Wu[1, 2, 3, *], PhD

[1] State Key Laboratory of Reproductive Medicine and Offspring Health, Center for Global Health, Nanjing Medical University, Nanjing, China

[2] Key Laboratory of Modern Toxicology of Ministry of Education, School of Public Health, Nanjing Medical University, Nanjing, China

[3] Taizhou Clinical Medical College, Nanjing Medical University, Taizhou, China

[4] The affiliated Wuxi People's Hospital of Nanjing Medical University, Wuxi People's Hospital, Wuxi Medical Center, Nanjing Medical University, Wuxi, China

[5] Department of Urology, Women's Hospital of Nanjing Medical University, Nanjing Maternity and Child Health Care Hospital, Nanjing, China

[6] Central Laboratory, Translational Medicine Research Center, The Affiliated Jiangning Hospital of Nanjing Medical University, Nanjing, China

[7] Department of Obstetrics, Women's Hospital of Nanjing Medical University, Nanjing Maternity and Child Health Care Hospital, Nanjing, China

[8] Department of Obstetrics and Gynecology, The Affiliated Jiangning Hospital of Nanjing Medical University, Nanjing, China

[*] Corresponding author. Wei Wu, State Key Laboratory of Reproductive Medicine and Offspring Health, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Avenue, Nanjing 211166, China. Email: wwu@njmu.edu.cn; Rong Ju, The Affiliated Jiangning Hospital of Nanjing Medical University, Nanjing, China. Email: jurong@njmu.edu.cn.

Word count of the manuscript text: 2882

Key Points

Question Can the male infertility prediction model based on questionnaire data and machine learning effectively predict the risk value of male infertility and calculate corresponding risk characteristics.

Findings BPNN model have good predictive performance and stability with the goal of predicting whether a patient is infertile or not. The average accuracy of BPNN in 10-fold cross validation is 0.845 (95%CI: 0.0.831, 0.859), Kappa value is 0.634 (95%CI: 0.604, 0.665).

Meaning We can use pre-trained machine learning models to help doctors in reproductive centers quickly screen for male infertility and calculate risk profiles to help give sound advice.

Abstract

IMPROTANCE Although clinical guidelines recommend the inclusion of semen analysis as part of the initial evaluation for couples facing infertility, there may be normal biological variability between semen samples collected from the same man at different times, which may result in imprecision.

OBJECTIVE To construct a model of male infertility to assist clinicians in efficiently and accurately assessing patient conditions and risk profiles.

DESIGN, SETTING, AND PARTICIPANTS The data for the study was obtained from a cross-sectional study conducted by Nanjing Medical University. Our study specifically focused on married men of reproductive age, as it would be inappropriate to explore male infertility among unmarried men. In total, we had 2,106 participants who met our inclusion criteria. Out of these participants, 2,012 responded to the questionnaire.

MAIN OUTCOMES AND MEASURES Participants were divided into two groups based on whether they were infertile or not. Male infertility refers to the incapacity of a couple to conceive naturally due to male factors after engaging in regular sexual activity for more than a year without using any contraceptive measures. Independent test set and cross-validation were used to test the efficacy of the model.

RESULTS Of the 2012 participants, 202 individuals (20%) were divided as a test set. The remainder served as the training set, and after balancing by the SMOTE algorithm the training set formed 2556 pieces of data (fertile: infertile = 1:1) passed into the model for training. All features collected from the questionnaire will be subjected to a series of tests and screening. BPNN model have good predictive performance and stability with the goal of predicting whether a patient is infertile or not, and the predicted probability is consistent with the observed risk. The average accuracy of BPNN in 10-fold

cross validation is 0.845 (95%CI: 0.0.831, 0.859), Kappa value is 0.634 (95%CI: 0.604, 0.665), and F1 value is 0.888 (95%CI: 0.877, 0.899).

CONCLUSIONS AND RELEVANCE The BPNN model can help clinicians make efficient and accurate initial judgments about the condition of infertility patients, and the SHAP method can effectively explain the risk characteristics of patients.

Introduction

Infertility is characterized as the incapacity of a couple to achieve pregnancy within a one-year timeframe without the use of contraceptive methods. It is estimated that about 10-20% of couples worldwide are infertile [1]. There is less education, awareness, and research on male reproductive health (both pre- and post-conception) compared to females. Studies that examine the characteristics of male infertility continue to heavily depend on the participation of men who seek medical evaluation for infertility at clinics.

Male infertility can be extremely distressing, leading to serious adverse health and psychological outcomes [2], and is associated with numerous social and economic consequences. Reproductive failure may coincide with other underlying processes, and an increasing body of evidence suggests that infertile men face a greater risk of experiencing issues related to cancer, mental health, cardiovascular disease, diabetes, and endocrine dysfunction when compared to fertile individuals [3,4]. In the existing literature, a substantial volume of population-based studies indicates that underlying male factors and the degree of male factor infertility may heighten the likelihood of mental retardation and autism in offspring [5,6]. Health problems in the offspring of infertile couples using assisted reproduction have attracted continued attention from researchers. Male infertility can arise from various causes, which can be influenced by environmental [7], nutritional [8–12], and lifestyle factors (e.g., smoking, alcohol abuse) [13], social stress [14], genetic history [1,15–19], and physical activity [20,21].

AI and ML technologies can provide a new way to improve diagnosis and prediction, as well as guide the treatment and decision-making process. The annual congresses of the "European Society for Human Reproduction and Embryology" and the "American Society for Reproductive Biology" in 2018 featured significant report that by leveraging AI and ML technologies, it becomes feasible to enhance

the accuracy of assessing sperm morphology and quality, identifying follicle and egg status, predicting embryo developmental stage and quality [22,23], and guiding the optimization of IVF stimulation protocols. Liao et al. develop a dynamic diagnosis grading system which uses ML to assess the condition of women with infertility[24]. It is believed that in the near future, AI researchers are deeply collaborating with reproductive clinicians to use AI models to obtain transparent, comparable, and reproducible results in the clinic to aid in decision-making [25].

Currently, although the initial workup of infertile couples is recommended by clinical guidelines to incorporate semen analysis [26–28], there may be inaccuracies due to normal biological variability between semen samples collected at different times from the same man. At the same time, there exists a considerable overlap in semen parameters between men experiencing infertility and those who are fertile. It is worth mentioning that men with seemingly low-quality semen samples can still achieve pregnancy successfully. The process of semen analysis can often make many men feel uncomfortable, as they may find it embarrassing and costly [29]. Despite comprehensive diagnostic testing, in around 40% of male infertility cases, the underlying cause remains "idiopathic". Although traditional semen analysis is important for clinical decision-making, except in cases where there is an absence of sperm in the semen or the sperm demonstrate significant morphological or functional abnormalities [30], the results of semen analysis cannot be used to predict a man’s fertility. However, it can be used to determine the severity of infertility [31]. Ferlin et al. believe that a couple-oriented approach should be adopted, and that high priority should be given to the assessment, prevention, and treatment of risk factors for infertility [32]. Overall, the results of semen analysis should be used in conjunction with other risk factors for male infertility as a guide for couples’ counseling and clinical decision-making. We are trying to fill this gap using a ML approach, where accurate ML models will help in the prediction of

male infertility risk and calculate individual risk factors.

Methods

A summary diagram of the entire study can be referenced from Figure 1.

Data Source

The data for the study was gathered from a cross-sectional study conducted by Nanjing Medical University. Study participants were recruited from the general population or individuals seeking fertility treatment. Our inclusion criteria targeted married men of reproductive age, as it would be inappropriate to investigate male infertility among unmarried men. The total number of participants was 2,106. Out of these, 2,012 participants responded to the questionnaire.

The questionnaire was designed in 7 sections which can be seen from Figure 1. After collecting all the questionnaires, quality control was carried out on each section of the questionnaire, which included correcting outliers, eliminating variables with over 30% missing values, and filling in any missing values. The method used for filling in missing values is the MICE (Multiple Imputation by Chained Equations) algorithm [33].

Feature Selection

Before Feature selection, variable recoding and normalization were performed (eMethods in the Supplement). Feature selection was performed in multiple steps. Firstly, the selection of larger features based on the variance selection method was performed, which involved calculating the variance of each feature. If a feature had a small variance, it was considered to have a smaller impact on the prediction results. Then, using the logistic regression model, the correlation between the features and the outcomes (eFigure 1 in the Supplement) was calculated. Features with a $P < .05$ were included. In this case, there was no need for multiple-adjust for P values in the correlation calculation as it was

exploratory analysis. This screening method was used for our initial model. Finally, a machine learning wrapper-based feature screening method was also employed, which further screened the features to reduce the complexity of the feature optimization model. The R package *FeatureSelection* helps to perform feature selection which wraps glmnet-lasso, xgboost and ranger.

Statistical methods

After constructing the study dataset, we performed AP clustering [34,35] and obtained a total of 173 clustered cores in the total sample of 2012, which suggests that the dataset itself is highly dispersed and complex, resulting in a diversity of clustering results. Additionally, it indicates that male infertility is indeed a condition with a multifaceted cause. A baseline table of the data can be found in eTable 1 in the Supplement. Due to the imbalance of the total sample data, in which the number of cases in the infertile sample (n = 1419) was larger than the number of cases in the normal sample (n = 593). Therefore, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was used to oversample the training set in the expectation of producing better training results [36,37]. SMOTE can balance the class distribution in the training set by synthesizing new minority class samples to improve the performance of the ML models (eMethods in the Supplement, eFigure 2 in the Supplement).

The technical flowchart of the modeling process is shown in eFigure 3 in the Supplement. Logistic Regression, Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Backpropagation Neural Networks (BPNN) were chosen for modeling. The hyperparameters of each model were fine-tuned using the Particle Swarm Optimization (PSO) algorithm (eFigure 2 in the Supplement) [38]. Following the completion of the modeling process, it was evaluated on a separate test set and validated using a 10-fold cross-validation approach. Several indicators were used for the evaluation of the model, such as accuracy, Kappa, recall, specificity, precision and F1 score (eMethods

in the Supplement).

The SHapley Additive exPlanations (SHAP) method is employed for model interpretation (eMethods in the Supplement). The SHAP method offers an evaluation of the contribution made by each feature or variable towards the prediction outcomes. This facilitates comprehension of the predictive process of the model and the factors that impact the results [39–41].

All computations in this paper are performed on a Linux server, and we use *Mamba* software to build two virtual environments, *ML_env* and *R_env*. The former is a Python environment (*Python* software, version 3.9.18) and the latter is an R environment (*R* software, version 4.2.3). We mainly modeled in the Python environment, which includes reading the data into a "dataframe" form using the *pandas* package, modeling using the *scikit-learn*, *TensorFlow*, and *Keras* packages, calculating the SHAP values using the *shap* package, and using the *caret*, *mice*, *FeatureSelection* and *tidyverse* packages in the R environment for data cleaning, feature screening, model evaluation comparison, and visualization work.

Results

Performance of each model in the test set and robustness of 10-fold cross-validation

The outcomes of the test set demonstrate the generalizability and practical applicability of the final evaluated model, whereas cross-validation is primarily employed to evaluate the model's generalization ability and stability. All the results are presented in Table 1, including the seven metrics used in the method described above. The prediction results of all models after tuning with the PSO algorithm are not bad, which demonstrates the feasibility of using questionnaire data for assessing male infertility. In contrast, the KNN model has the lowest predictive performance, with an average accuracy of less than

0.8 and 0.4 ≤ Kappa < 0.6. This indicates that the model has moderate performance with some accuracy, and the AUC of KNN is significantly lower than the other models. However, despite the SVM model having a Kappa value of only 0.512 on the test set, the average Kappa value, determined through cross-validation, was found to be 0.650 (95%CI: 0.605, 0.695). This indicates that its lower performance in the test set is purely coincidental and a result of the random splitting of the study dataset. For all models except KNN, with 0.6 ≤ Kappa < 0.8, this indicates that the models perform well with high accuracy. It is also surprised to find that both ends of the confidence intervals for the AUC of BPNN, Logistic, and XGBoost were above 0.8, which is a remarkable result.

ROC curves, DCA curves, calibration curves of the models

The ROC curves of all models are depicted in Figure 2. The result of KNN is represented by a broken line, indicating that the model only outputs two values (0/1). This limited output format is not suitable for predicting risk probabilities.

From the DCA curves (eFigure 4 in the Supplement), it is evident that the net benefit of KNN is the lowest and may even be lower than that of the "Intervention for all" group under certain circumstances. The net benefit of SVM is calculated to be poor. From the threshold probability setting of 0.5, BPNN, Logistic, RF, and XGBoost exhibit similar performance.

The calibration curves of all the models are shown in Figure 3. In the calibration curves, when the predicted probability is greater than 0.5, the predicted probability of BPNN agrees slightly better with the actually observed proportion of positives than that of XGBoost, especially near the predicted probability of 0.75, but there is some risk of underestimating the actual probability for both models. Since the high-risk group is our focus, and combining the previous evaluations, BPNN was chosen as the final model to be used for interpretation among the two models.

Model interpretation based on the SHAP method

A random sample of normal and a sample of infertility were selected for interpretation and comparison (Figure 4). After the BPNN modeling the risk probability was calculated to be 0.015 for the normal sample and 0.854 for the infertile sample. In the left radar charts (Figure 4A, Figure 4B), the cumulative risk values for the six categories were calculated. If the point corresponding to the category falls in the outer circle, the category contributed to making a positive prediction——male infertility contributes, and vice versa, it plays a protective effect. The positive/negative contribution of the corresponding tester's characteristics to male infertility is also detailed in Figure 4C and Figure 4D.

Further filtering features based on machine learning wrappers and comparison of their effectiveness

Based on the wrapper mentioned in the methodology, the features are further screened, and the features with union importance greater than 0.5 are selected, and the union importance of all the features is demonstrated in eFigure 5 in the Supplement. The purpose of further screening of the features is to improve the applicability of the model, which is also one aspect of the optimizing the model from a certain point of view. It can be seen in eFigure 6 in the Supplement that after SMOTE, the confidence intervals of all the indicators become narrower, indicating that the model is more stable, which is the desirability of SMOTE oversampling on the training set. After further screening of the features, the predictive performance of the model decreases, especially the Kappa value, the Kappa value in the test set validation and the average Kappa value in the 10-fold cross-validation are both less than 0.6, which is inferior to the previous model, but noteworthy, it still demonstrates the ability to make good predictions.

Discussion

Numerous aspects of fertility treatment can be conducted remotely, encompassing virtual consultations, home testing, and personalized digital fertility assessments, thereby adapting to the specific requirements of each patient [42]. In the near future, with the aid of big data, AI systems will possess the capability to forecast the individual patient's risk and propose suitable treatment alternatives. Male reproductive medicine physicians play an crucial role in a complex decision-making network and must carefully balance the costs and benefits of diagnostic and therapeutic efforts to provide optimal support for infertile couples [43].

The use of preclinical health questionnaires and the collection of data to assess the health of specific populations using machine learning has been proposed in previous study [44]. Avram et al. describe the advantages and disadvantages of this approach, in short, the primary benefit of this approach is its capacity for conducting a direct analysis of the gathered data —— everything is processed by the model; this is two sides of the same coin, the disadvantage of this approach is that in some cases it may not be very accurate, the answers provided in the questionnaire can be subjective, and it is difficult to compare with the precision of medical instruments. It is therefore advisable to limit the application of the model, for example, to patient self-assessment or diagnostic aids.

This study may provide a screening tool for male infertility for families who are uncomfortable with process of semen collection in hospitals. It is further recommended that the results of this model be used by clinical practitioners in conjunction with semen analysis for a better and more comprehensive assessment of male infertility.

For predictive models, there is often a risk of overfitting, i.e., the outcomes observed in the training set exhibit notably superior performance compared to those in the test set, indicating a possibility that the PSO tuning process could potentially encounter a local optimum. However, in this study, a large

PSO tuning range and number of iterations were set to ensure that the tuning process is well-executed. In addition, L1 regularization [45,46] and dropout layers were added to the BPNN model, L1 regularization is employed to restrict the network parameters, thereby enhancing the model's generalization capability and resilience to noise. Additionally, the dropout layer effectively mitigates overfitting and enhances the model's generalization ability.

For model interpretation, two random samples were selected, one of which was a normal male and the other an infertile male. As far as the prediction is concerned, the predicted risk probabilities are consistent with their actual state. Meanwhile, the risk values of all characteristics were calculated, and from the radar graph, it can be seen that normal males played a protective role in all aspects except for the abnormalities in Body parameters; in contrast, dietary habits, lifestyle, and social status, which are all risk factors for the selected infertile male. In this part, the condition of a simple assumption that the risk of infertility in an average male is superimposed is the basis. Also, according to the results of these calculations, evidence-based public health recommendations can be provided by considering the variations in different aspects of the individual's risk profile.

Limitations

Of course, there are certain limitations to this study, starting with the small sample size utilized and the distribution of positive and negative samples is uneven, something that is difficult to avoid; after all, it is generally easier to collect information on patients in hospitals than on healthy individuals. Second, there may be missing key variables, on the one hand, some variables were excluded due to missing greater than 30% during data cleaning, for example, average household income is an important risk factor in most studies, but due to the fact that most of the participants in this study were not willing to disclose their economic status resulting in this variable regrettably being excluded from this study;

on the other hand, because our questionnaire mainly focused on common exposures or potential risks, there is a possibility that other potential risks were left out, especially exposures that individuals themselves may not be aware of, such as insurance status [47].

Conclusions

Our team has developed a tool that can rapidly predict and calculate the risk of male infertility and its characteristics under non-laboratory conditions. Based on the risk profile calculated by the model, rational public health interventions can be delivered. This tool is expected to complement laboratory conditions and improve the diagnosis of male infertility. Additionally, the use of machine learning (on questionnaire data) combined with clinical semen analysis to assess male infertility is recommended. This approach will help in addressing different etiologies of male infertility and making rational interventions.

References

1. Eisenberg ML, Esteves SC, Lamb DJ, et al. Male infertility. *Nat Rev Dis Primer*. 2023;9(1):49. doi:10.1038/s41572-023-00459-w

2. Belladelli F, Muncey W, Eisenberg ML. Reproduction as a window for health in men. *Fertil Steril*. 2023;120(3):429-437. doi:10.1016/j.fertnstert.2023.01.014

3. Keihani S, Hanson B, Hotaling J. Male factor infertility: an opportunity to investigate individual and family health. *BJOG Int J Obstet Gynaecol*. 2019;126(2):149-151. doi:10.1111/1471-0528.15398

4. Punjani N, Lamb DJ. Canary in the Coal Mine? Male Infertility as a Marker of Overall Health. *Annu Rev Genet*. 2020;54(1):465-486. doi:10.1146/annurev-genet-022020-023434

5. Bellver J, Donnez J. Introduction: Infertility etiology and offspring health. *Fertil Steril*. 2019;111(6):1033-1035. doi:10.1016/j.fertnstert.2019.04.043

6. Rumbold AR, Sevoyan A, Oswald TK, Fernandez RC, Davies MJ, Moore VM. Impact of male factor infertility on offspring health and development. *Fertil Steril*. 2019;111(6):1047-1053. doi:10.1016/j.fertnstert.2019.05.006

7. Esteves SC, Humaidan P. Towards infertility care on equal terms: a prime time for male infertility. *Reprod Biomed Online*. 2023;47(1):11-14. doi:10.1016/j.rbmo.2023.04.003

8. Bosdou J, Konstantinidou E, Anagnostis P, Kolibianakis E, Goulis D. Vitamin D and Obesity: Two Interacting Players in the Field of Infertility. *Nutrients*. 2019;11(7):1455. doi:10.3390/nu11071455

9. Khodamoradi K, Parmar M, Khosravizadeh Z, Kuchakulla M, Manoharan M, Arora H. The role of leptin and obesity on male infertility. *Curr Opin Urol*. 2020;30(3):334-339. doi:10.1097/MOU.0000000000000762

10. Mahbouli S, Dupont C, Elfassy Y, Lameignère E, Levy R. Exploring the potential impact of nutritionally actionable genetic polymorphisms on idiopathic male infertility: a review of current evidence. *Asian J Androl*. 2021;23(5):441. doi:10.4103/aja.aja_87_20

11. Sousa ACA, Alves MG, Oliveira PF, Silva BM, Rato L. Male Infertility in the XXI Century: Are Obesogens to Blame? *Int J Mol Sci*. 2022;23(6):3046. doi:10.3390/ijms23063046

12. Service CA, Puri D, Al Azzawi S, Hsieh TC, Patel DP. The impact of obesity and metabolic health on male fertility: a systematic review. *Fertil Steril*. 2023;120(6):1098-1111. doi:10.1016/j.fertnstert.2023.10.017

13. Basic M, Mitic D, Krstic M, Cvetkovic J. Tobacco and alcohol as factors for male infertility—a public health approach. *J Public Health*. 2023;45(2):e241-e249. doi:10.1093/pubmed/fdac042

14. Bräuner EV, Nordkap L, Priskorn L, et al. Psychological stress, stressful life events, male factor infertility, and testicular function: a cross-sectional study. *Fertil Steril*. 2020;113(4):865-875. doi:10.1016/j.fertnstert.2019.12.013

15. Ferlin A, Dipresa S, Delbarba A, et al. Contemporary genetics-based diagnostics of male infertility. *Expert Rev Mol Diagn*. 2019;19(7):623-633. doi:10.1080/14737159.2019.1633917

16. Karimian M, Parvaresh L, Behjati M. Genetic variations as molecular diagnostic factors for idiopathic male infertility: current knowledge and future perspectives. *Expert Rev Mol Diagn*. 2021;21(11):1191-1210. doi:10.1080/14737159.2021.1985469

17. Pini T, Raubenheimer D, Simpson SJ, Crean AJ. Obesity and Male Reproduction; Placing the Western Diet in Context. *Front Endocrinol*. 2021;12:622292. doi:10.3389/fendo.2021.622292

18. Chen T, Belladelli F, Del Giudice F, Eisenberg ML. Male fertility as a marker for health. *Reprod Biomed Online*. 2022;44(1):131-144. doi:10.1016/j.rbmo.2021.09.023

19. Ameratunga D, Gebeh A, Amoako A. Obesity and male infertility. *Best Pract Res Clin Obstet Gynaecol*. 2023;90:102393. doi:10.1016/j.bpobgyn.2023.102393

20. Hwang K, Guo D. Sports-related Male Infertility. *Eur Urol Focus*. 2019;5(6):1143-1145. doi:10.1016/j.euf.2018.04.010

21. Bisconti M, Simon JF, Grassi S, et al. Influence of Risk Factors for Male Infertility on Sperm Protein Composition. *Int J Mol Sci*. 2021;22(23):13164. doi:10.3390/ijms222313164

22. Blank C, Wildeboer RR, DeCroo I, et al. Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective. *Fertil Steril*. 2019;111(2):318-326. doi:10.1016/j.fertnstert.2018.10.030

23. Wei D, Legro RS, Chen ZJ. The application of artificial intelligence in reproductive medicine: baby steps. *Fertil Steril*. 2022;118(1):109-110. doi:10.1016/j.fertnstert.2022.05.002

24. Liao S, Pan W, Dai W qiang, et al. Development of a Dynamic Diagnosis Grading System for Infertility Using Machine Learning. *JAMA Netw Open*. 2020;3(11):e2023654. doi:10.1001/jamanetworkopen.2020.23654

25. Rosenwaks Z. Artificial intelligence in reproductive medicine: a fleeting concept or the wave of the future? *Fertil Steril*. 2020;114(5):905-907. doi:10.1016/j.fertnstert.2020.10.002

26. Oehninger S, Ombelet W. Limits of current male fertility testing. *Fertil Steril*. 2019;111(5):835-841. doi:10.1016/j.fertnstert.2019.03.005

27. Sigman M. Introduction: Male fertility testing: the past, present, and future. *Fertil Steril*. 2019;111(5):833-834. doi:10.1016/j.fertnstert.2019.03.008

28. Pozzi E, Ramasamy R, Salonia A. Initial Andrological Evaluation of the Infertile Male. *Eur Urol Focus*. 2023;9(1):51-54. doi:10.1016/j.euf.2022.09.012

29. Kobori Y. Home testing for male factor infertility: a review of current options. *Fertil Steril*. 2019;111(5):864-870. doi:10.1016/j.fertnstert.2019.01.032

30. Pandruvada S, Royfman R, Shah TA, et al. Lack of trusted diagnostic tools for undetermined male infertility. *J Assist Reprod Genet*. 2021;38(2):265-276. doi:10.1007/s10815-020-02037-5

31. Lamb DJ, Marinaro JA. Can semen parameters predict pregnancy outcomes? *Fertil Steril*. 2023;120(4):709-714. doi:10.1016/j.fertnstert.2023.06.035

32. Ferlin A, Calogero AE, Krausz C, et al. Management of male factor infertility: position statement from the Italian Society of Andrology and Sexual Medicine (SIAMS): Endorsing Organization: Italian Society of Embryology, Reproduction, and Research (SIERR). *J Endocrinol Invest*. 2022;45(5):1085-1113. doi:10.1007/s40618-022-01741-6

33. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067

34. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. *Science*. 2007;315(5814):972-976. doi:10.1126/science.1136800

35. Li N, Latecki LJ. Affinity learning for mixed data clustering. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. AAAI Press; 2017:2173-2179.

36. Chawla N, Bowyer K, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *ArXiv*. 2002;abs/1106.1813.

37. Giulioni C, Maurizi V, Castellani D, et al. The environmental and occupational influence of pesticides on male fertility: A systematic review of human studies. *Andrology*. 2022;10(7):1250-1271. doi:10.1111/andr.13228

38. Kennedy J, Eberhart R. Particle swarm optimization. In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. Vol 4. ; 1995:1942-1948 vol.4. doi:10.1109/ICNN.1995.488968

39. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. Published online 2017.

40. Kuno T, Sahashi Y, Kawahito S, Takahashi M, Iwagami M, Egorova NN. Prediction of in-hospital mortality with machine learning for COVID-19 patients treated with steroid and remdesivir. *J Med Virol*. 2022;94(3):958-964. doi:10.1002/jmv.27393

41. Dimitsaki S, Gavriilidis GI, Dimitriadis VK, Natsiavas P. Benchmarking of Machine Learning classifiers on plasma proteomic for COVID-19 severity prediction through interpretable artificial intelligence. *Artif Intell Med*. 2023;137:102490. doi:10.1016/j.artmed.2023.102490

42. Curchoe CL, Malmsten J, Bormann C, et al. Predictive modeling in reproductive medicine:

Where will the future of artificial intelligence research take us? *Fertil Steril*. 2020;114(5):934-940. doi:10.1016/j.fertnstert.2020.10.040

43. Pallotti F, Barbonetti A, Rastrelli G, Santi D, Corona G, Lombardo F. The impact of male factors and their correct and early diagnosis in the infertile couple's pathway: 2021 perspectives. *J Endocrinol Invest*. 2022;45(10):1807-1822. doi:10.1007/s40618-022-01778-7

44. Avram C, Gligor A, Roman D, Soylu A, Nyulas V, Avram L. Machine learning based assessment of preclinical health questionnaires. *Int J Med Inf*. 2023;180:105248. doi:10.1016/j.ijmedinf.2023.105248

45. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58(1):267-288. doi:10.1111/j.2517-6161.1996.tb02080.x

46. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2324. doi:10.1109/5.726791

47. Persily J, Stair S, Najari BB. Access to infertility services: characterizing potentially infertile men in the United States with the use of the National Survey for Family Growth. *Fertil Steril*. 2020;114(1):83-88. doi:10.1016/j.fertnstert.2020.03.005

Figure legends

Figure 1. Summary chart of the whole study. MICE: Multiple Imputation by Chained Equations. ROC: Receiver Operating Characteristic; DCA: decision curve analysis; Max-Min normalization: $X_{normalized} = (X - X_{min}) / (X_{max} - X_{min})$; SMOTE: Synthetic Minority Over-sampling Technique. Logistic: Logistic regression; SVM: Support Vector Machine, RF: Random Forest, NB: Naive Bayes, BP-NN: Backpropagation Neural Networks.

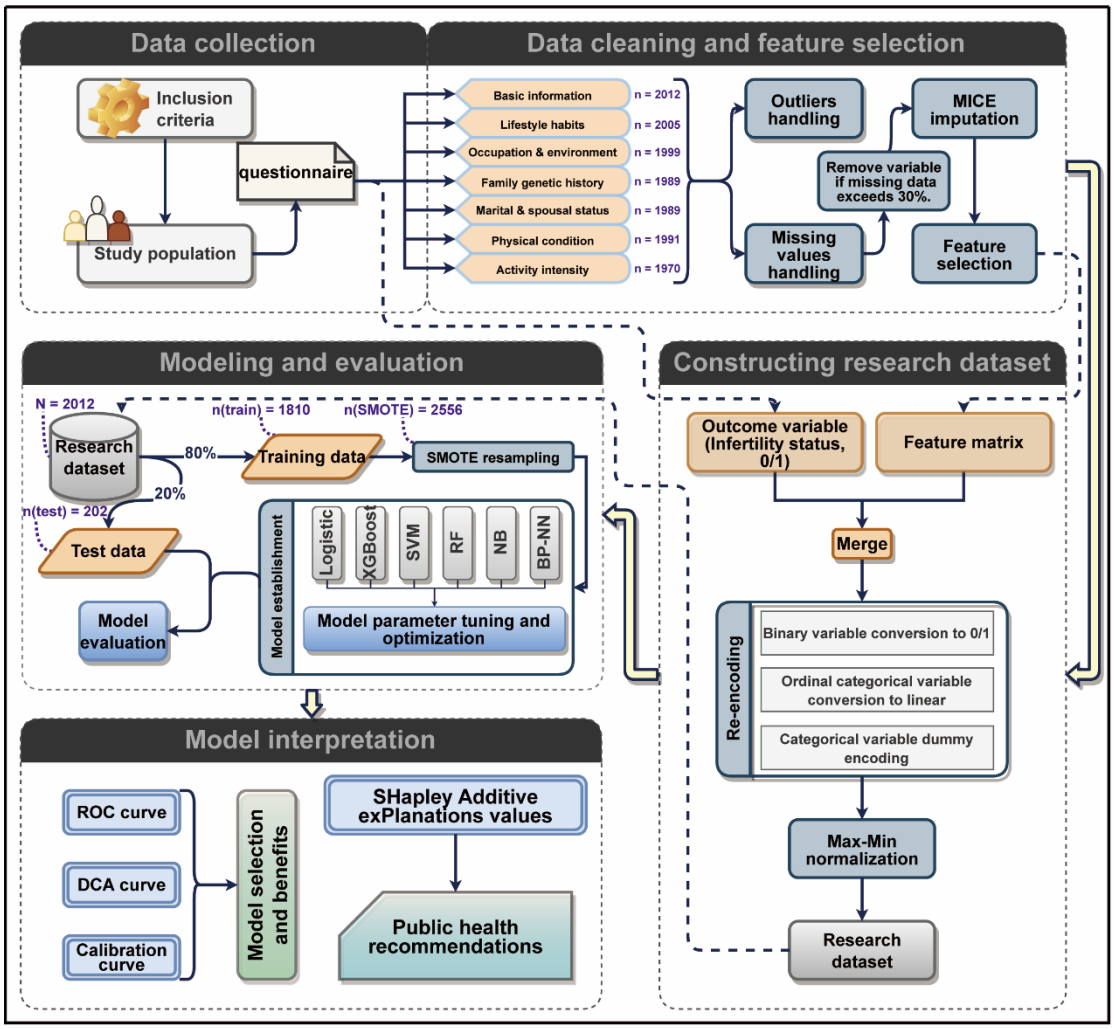


Figure 2. ROC curves established for each model. (A)-(F): The legend of each figure contains the name of each model; The shaded 95% CIs were calculated using the Bootstrap algorithm; AUC: Area Under the Curve. Logistic: Logistic regression; SVM: Support Vector Machine, RF: Random Forest, NB: Naive Bayes, BPNN: Backpropagation Neural Networks.

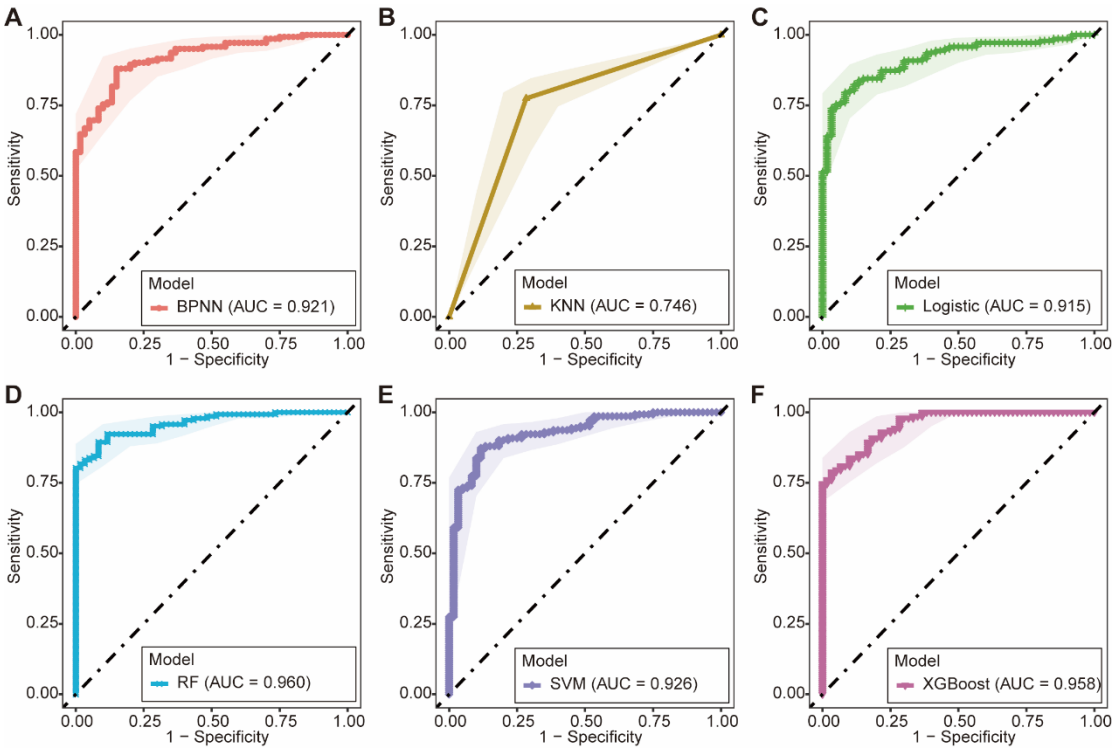


Figure 3. Calibration curves established for each model. (A)-(F): The legend of each figure contains the name of each model; The calibration curve illustrates the correlation between the predicted probabilities from the model and the observed probabilities. The ideal calibration curve is represented by a 45-degree line, indicating that the model's predictions align perfectly with the actual observations. This means that the predicted probability of an event matches precisely with the probability of its actual occurrence. When the model exhibits this level of alignment, it is considered to be fully calibrated. Logistic: Logistic regression; SVM: Support Vector Machine, RF: Random Forest, NB: Naive Bayes, BPNN: Backpropagation Neural Networks.

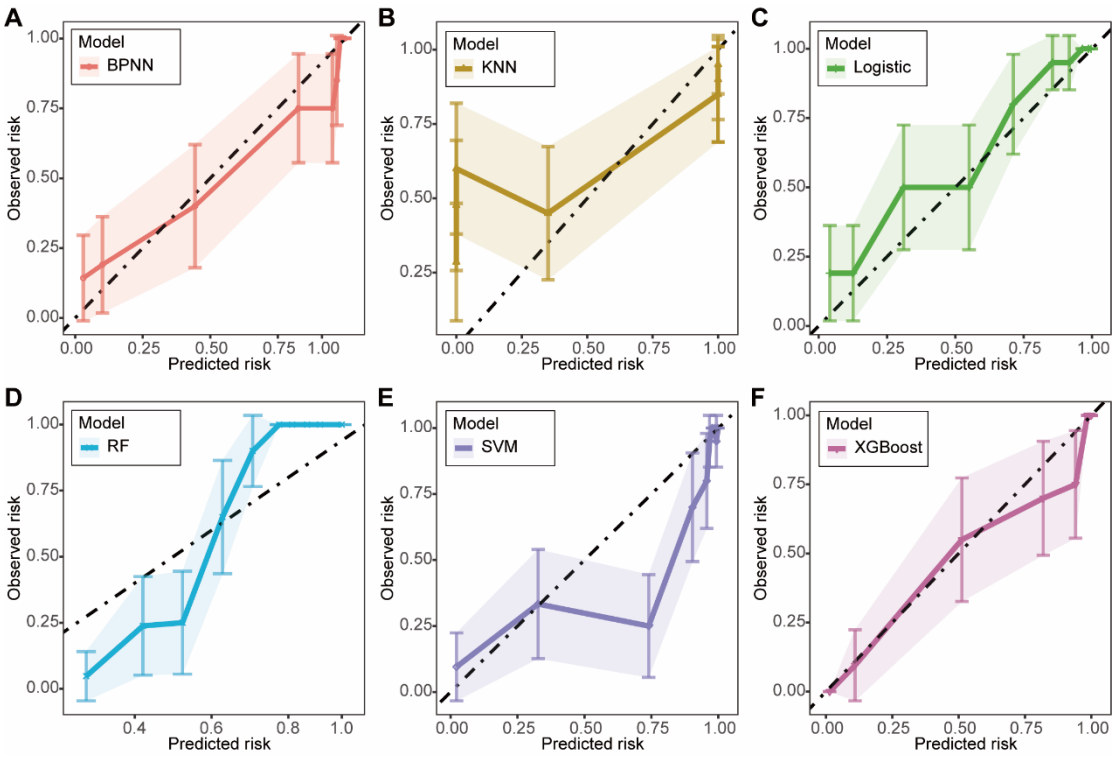
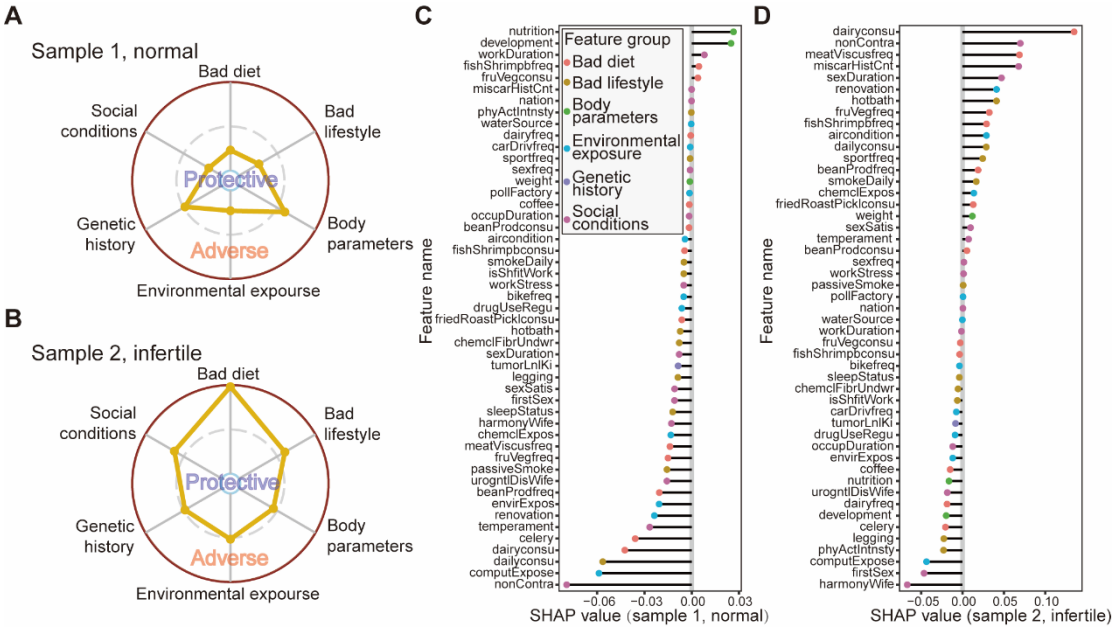


Figure 4. Risk characteristics based on SHAP values on 2 random samples.



Tables

Table 1. Evaluation of each model on the test set and results of cross validation.

Model		Accuracy	Kappa	F1	Recall	Specificity	Precision	AUC
BPNN	*Test*[a]	0.851 (0.795, 0.898)	0.634	0.897	0.915	0.700	0.878	0.921
	CV[b]	0.845 (0.831, 0.859)	0.634 (0.604, 0.665)	0.888 (0.877, 0.899)	0.876 (0.850, 0.902)	0.771 (0.724, 0.817)	0.902 (0.886, 0.919)	0.823 (0.806, 0.841)
KNN	*Test*	0.757 (0.692, 0.815)	0.458	0.818	0.775	0.717	0.866	0.746
	CV	0.746 (0.725, 0.767)	0.454 (0.413, 0.496)	0.803 (0.785, 0.820)	0.734 (0.710, 0.758)	0.774 (0.736, 0.813)	0.886 (0.869, 0.904)	0.754 (0.731, 0.777)
Logistic	*Test*	0.842 (0.784, 0.889)	0.631	0.885	0.866	0.783	0.904	0.915
	CV	0.838 (0.822, 0.855)	0.636 (0.601, 0.671)	0.880 (0.867, 0.892)	0.838 (0.817, 0.859)	0.840 (0.806, 0.874)	0.926 (0.912, 0.941)	0.839 (0.821, 0.857)
RF	*Test*	0.876 (0.823, 0.918)	0.684	0.916	0.958	0.683	0.877	0.960
	CV	0.869 (0.851, 0.887)	0.665 (0.620, 0.710)	0.911 (0.898, 0.923)	0.949 (0.932, 0.965)	0.678 (0.642, 0.714)	0.876 (0.863, 0.889)	0.813 (0.791, 0.835)
SVM	*Test*	0.817 (0.756, 0.868)	0.512	0.879	0.944	0.517	0.822	0.926
	CV	0.863 (0.846, 0.881)	0.650 (0.605, 0.695)	0.907 (0.895, 0.919)	0.946 (0.930, 0.962)	0.666 (0.626, 0.707)	0.872 (0.858, 0.886)	0.806 (0.783, 0.829)
XGBoost	*Test*	0.881 (0.828, 0.922)	0.704	0.918	0.944	0.733	0.893	0.958
	CV	0.886 (0.868, 0.903)	0.721 (0.678, 0.764)	0.920 (0.907, 0.932)	0.929 (0.914, 0.943)	0.783 (0.747, 0.818)	0.911 (0.897, 0.925)	0.856 (0.834, 0.878)

[a] Testing in a test set; [b] 10-fold cross validation.