Current Science
CS 5 (2), 1942-1955 (2025)
Received 22 Jan 2025 | **Revised 20 Feb** 2025 | **Accepted 20 March**. 2025 | **Online Available 25 April** 2025
https://doi.org/ 10.52845/CS/2025-5-2-54

Current Science

Current Science

OPEN ACCESS

ISSN(O)2795-8639

ORIGINAL ARTICLE

Building Personalized Risk Prediction Models for Polycystic Ovary Syndrome Using Machine Learning Techniques: A Retrospective Study

Baoqin Liu[1], Haoming Xia[2], Junning Zhang[3], Yue Hu[4], Qing Wang[1,*]

[1]Department of TCM Gynecology, China-Japan Friendship Hospital, Beijing 100029, China

[2]School of Clinical Medicine, Tsinghua University, Beijing 100084, China

[3]Graduate School, Beijing University of Chinese Medicine, Beijing100029, China

[4]Department of Oncology, Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053, China

[*]Corresponding Author: Qing Wang

Abstract:
Due to the highly diverse pathological features of Polycystic Ovary Syndrome (PCOS), traditional diagnostic methods often struggle to accurately predict the risk of the disease. With the increasing application of artificial intelligence technology in the medical field, machine learning offers a new approach to enhance the accuracy and efficiency of PCOS prediction.In this study, we employed ten machine learning algorithms, including Random Forest (RF), Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM), to analyze clinical retrospective data from multiple centers to identify and evaluate key risk factors associated with PCOS. First, rigorous data preprocessing was conducted, followed by feature selection using the LASSO regression method. Cross-validation techniques were then used to assess the performance of each prediction model on the training dataset. Finally, the models were validated on an independent test set to evaluate their generalization ability.In cross-validation, the RF model excelled in all performance metrics, particularly achieving an average accuracy of 92.56% and an F1 score of 92.63%. Evaluation results on the test set also confirmed the superior performance of the RF model, with LDA showing outstanding performance in specific metrics. Furthermore, the selected 12 key risk factors demonstrated significant clinical relevance for PCOS prediction.This study demonstrates the potential application of machine learning methods in PCOS risk prediction. Our model accurately identifies high-risk PCOS patients, providing robust data support for early diagnosis and personalized treatment. This work not only improves the efficiency and effectiveness of PCOS management but also lays the foundation for future applications of artificial intelligence technology in other complex diseases.

Keyword: Polycystic Ovary Syndrome; artificial intelligence technology; machine learning; feature selection; risk prediction

Introduction

Polycystic Ovary Syndrome (PCOS) is a common clinical endocrine disorder that affects a significant proportion of reproductive-age women, accounting for approximately 20% of women of reproductive age [1]. PCOS presents with diverse clinical manifestations, including hyperandrogenism, ovulatory dysfunction, and ovarian cystic changes seen on ultrasound examination [2]. This condition is not only a leading cause of anovulatory infertility, affecting approximately 72% of patients who are unable to conceive naturally [3], but it is also associated with metabolic disorders such as obesity, insulin resistance, hyperlipidemia, and metabolic syndrome [4]. In the long term, PCOS patients face a significant risk of serious health problems,

including diabetes, cardiovascular disease, and even tumors, posing a severe threat to their life and well-being.

Due to the high heterogeneity and complex pathophysiological features of PCOS, as well as the unclear understanding of its etiology, diagnosing it clinically presents a challenge. PCOS patients exhibit a wide range of clinical presentations, including but not limited to irregular menstruation, hirsutism, obesity, and multiple ovarian cysts seen on ultrasound, posing a diagnostic dilemma for healthcare providers [5,6]. Furthermore, given the chronic nature of PCOS and the absence of a curative treatment, clinical management often requires a combination of medications and lifestyle modifications to optimize patients' health. Additionally, due to the increased risk of complications associated with PCOS, such as type 2 diabetes, cardiovascular disease, and endometrial cancer, early diagnosis and prevention of these complications become crucial [7-9]. In this context, the development of personalized PCOS risk prediction tools holds significant value for early detection of the disease, timely intervention, and improving long-term health outcomes. Personalized prediction tools can offer tailored risk assessments based on individual patients' clinical characteristics and biomarkers, guiding healthcare providers to make targeted interventions, optimize treatment plans, and adjust strategies through continuous monitoring. Furthermore, these tools can assist in the rational allocation of clinical resources by predicting high-risk patient populations and may reduce the cost and complexity of long-term treatment through early intervention. Therefore, personalized PCOS risk prediction can not only enhance the quality of life for patients but also yield benefits for the entire healthcare system.

With the advancement of machine learning technology, new possibilities have emerged for PCOS risk assessment and prediction. In existing research, several scholars have attempted to apply various machine learning approaches to predict risk from multiple perspectives. For instance, Samantha et al. [10] employed a hybrid machine learning model, including Random Forest (RF), Decision Trees (DT), and Logistic Regression (LR), with an optimized and minimal set of parameters for early diagnosis and prevention of PCOS. Baweja A K et al. [11] used five machine learning methods to predict the risk of PCOS and found that using a Multilayer Perceptron (MLP) achieved higher prediction accuracy. Neto et al. [12] compared the application of algorithms such as Support Vector Machines (SVM), MLP, RF, and LR in clinical data mining of PCOS patients and found that Random Forest performed best in terms of prediction accuracy. Denny et al. [13] demonstrated high prediction accuracy with the i-HOPE model constructed using machine learning methods. Zad Z et al. [14] developed a prediction model based on outpatient data, further promoting early diagnosis and risk assessment of PCOS. However, existing prediction methods still lack comprehensive analysis of key factors and clinical interpretability, and there is a need for improved prediction accuracy.

In the context outlined above and based on existing research, this study aims to further explore and enhance PCOS risk prediction models. We conduct in-depth analysis of a large dataset of clinical retrospective data using eight advanced machine learning algorithms, including RF, Linear Discriminant Analysis (LDA), and SVM, to identify the clinical indicators most strongly associated with the risk of PCOS. Particular emphasis is placed on how to select the most predictive factors from a multitude of clinical features and construct an efficient and accurate prediction model based on these factors, as depicted in the overall research workflow in Figure 1.

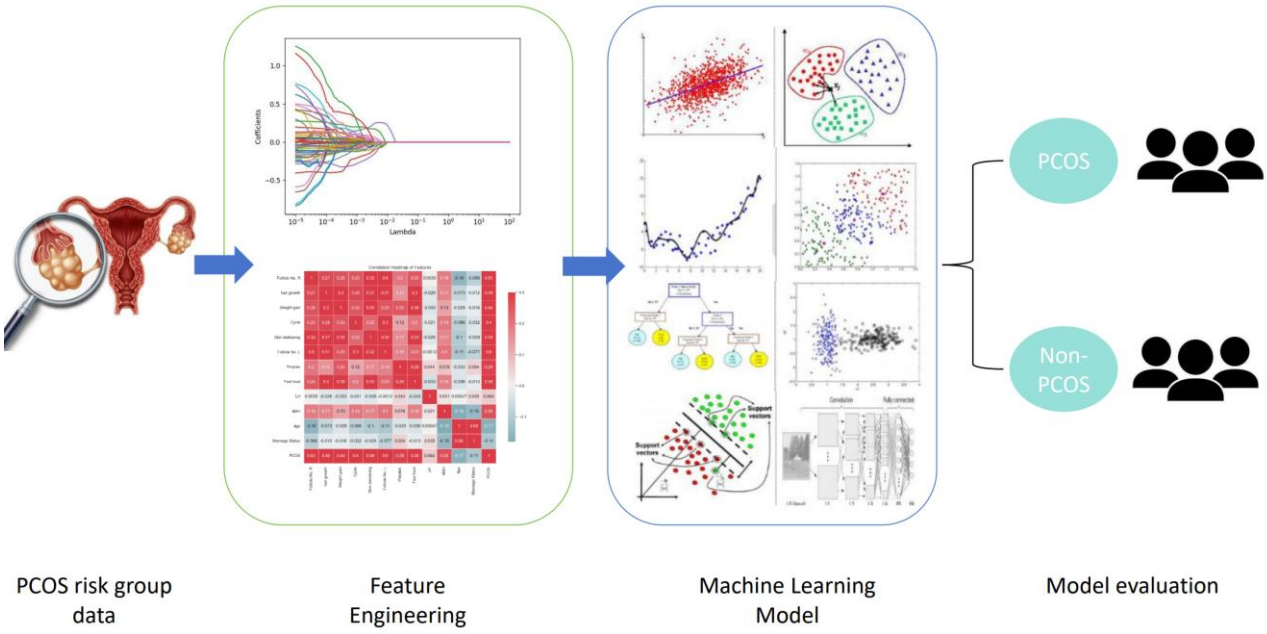

Figure 1: Flowchart of this study.

Materials and Methods

Data Collection

The data for this study were sourced from the internationally renowned data mining repository Kaggle, where a retrospective cohort study dataset on Polycystic Ovary Syndrome (PCOS) patients was uploaded in 2020. This dataset was collected from ten different hospitals in the state of Kerala, India, covering multiple regions to ensure diversity and representativeness of the data. The dataset includes a total of 541 patients, comprising 177 confirmed PCOS patients and 364 non-PCOS patients. This composition of patient groups contributes to a more accurate analysis and prediction of PCOS risk factors.

The dataset encompasses all basic patient information, such as age, weight, height, as well as vital signs like blood pressure and heart rate. Of greater significance, the dataset also incorporates clinical indicators related to Polycystic Ovary Syndrome, such as hormone levels, uterine lining size, follicle count, and more. These comprehensive data provide ample room for feature selection in our machine learning model, enabling the model to more accurately identify and analyze the risk of PCOS. Detailed information about this dataset can be accessed through the following link: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos.

Data Preprocessing

In this study, data preprocessing is a critical step in building the machine learning model to ensure data quality and suitability. Firstly, a significant challenge we faced was dealing with missing values in the dataset. To address these missing values, we employed the K-Nearest Neighbors (KNN) algorithm for imputation. This method estimates missing data points based on the data features of similar patients. Specifically, we imputed missing values by weighted averaging based on the feature values of neighboring samples, preserving the overall distribution characteristics and inherent correlations in the data, thereby providing a more reliable data foundation for subsequent analysis.

Secondly, to address the issue of data imbalance in the dataset, we applied Synthetic Minority Over-sampling Technique (SMOTE) for data augmentation [15-17]. As the number of non-PCOS samples exceeded that of PCOS samples, direct model training could lead to a bias toward the majority class, affecting prediction accuracy. The SMOTE algorithm effectively balances the class distribution by resampling the minority class samples and downsampling the majority class samples. This method improves class balance by generating synthetic samples for the minority

class while preserving the original structural features of the data, providing a more balanced data foundation for model training.

Lastly, we conducted denoising and normalization of the dataset. The denoising step aims to remove noise and outliers from the data, enhancing data quality and usability. Normalization, on the other hand, is performed to eliminate the influence of different scales that may exist among different features, ensuring that all features have equivalent importance during model training. We employed standardization by transforming the feature values of all continuous variables into a distribution with a mean of 0 and a standard deviation of 1. These preprocessing steps create a uniform and clean data environment for machine learning model training, thereby enhancing model accuracy and generalization.

Feature Selection

In this study, feature selection is a crucial step in optimizing the performance of the machine learning model. We chose the Least Absolute Shrinkage and Selection Operator (LASSO) regression method for feature selection [18,19]. LASSO regression introduces a regularization term that effectively shrinks the coefficients of unimportant features to zero, achieving feature selection and dimensionality reduction. To determine the optimal regularization parameter alpha in LASSO regression, we employed 10-fold cross-validation with logistic regression. In this process, we divided the dataset into 10 subsets, training the model on 9 subsets while testing it on the remaining subset, repeating this process to ensure model stability and reliability. We evaluated model performance under different alpha values and selected the alpha value that minimized the cross-validation error as the optimal parameter.

Once the optimal alpha value was determined, we used LASSO regression to analyze and select features highly correlated with the risk of PCOS. Through the coefficient-shrinking property of LASSO regression, we aimed to identify the most predictive features from a pool of candidate features. This not only helps reduce model complexity but also enhances model interpretability and accuracy. Additionally, we generated a series of visualizations to intuitively showcase the feature selection results of LASSO regression, including plots depicting the relationship between feature coefficients and different alpha values, as well as coefficient distribution plots for selected features at the optimal alpha value. These visual results provide crucial visual support for subsequent model training and interpretation.

Construction of Machine Learning-Based Risk Prediction Models

Model Selection and Characteristics

In constructing the risk prediction model for PCOS, we considered 10 mainstream machine learning algorithms for comparison. These algorithms encompass a range of techniques and characteristics: We began with Logistic Regression (LR), a widely applied method in medical settings due to its interpretable outcomes. K-Nearest Neighbors (KNN) was chosen for its non-parametric nature, making it suitable for pattern recognition. Support Vector Machines (SVM) were included for their strong classification capabilities, particularly in high-dimensional data analysis. Decision Tree (DT), known for its transparent decision-making process, is often used in clinical decision support. Random Forest (RF), an ensemble learning technique, was selected to enhance prediction accuracy and stability. Gradient Boosting (GB) was included as it effectively reduces model bias through iterative optimization. Gaussian Naive Bayes (GNB), based on the assumption of feature independence, provides rapid classification decisions. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) were considered for their effectiveness in distinguishing different classes when feature distributions are assumed to be Gaussian. Lastly, Multi-Layer Perceptron (MLP), a versatile feedforward neural network, adapts well to complex non-linear problems.

Model Optimization and Application

In this study, each model underwent careful parameter optimization to ensure optimal performance on our specific dataset. We used grid search (GridSearchCV), which iterates through a defined parameter grid to find the optimal parameter combinations for each model [20,21]. In LR, we tuned the regularization parameter C; for KNN, we optimized the choice of the number

of neighbors (k); for SVM, we adjusted both the penalty parameter C and the kernel type. Parameters such as maximum depth and feature selection criteria for decision trees and random forests were included in the tuning scope. In GB, the learning rate and the number of trees were key parameters for optimization. For GNB, we mainly focused on adjusting prior probabilities. In LDA and QDA, different covariance structures were considered. Finally, for MLP, we fine-tuned network architecture, activation functions, and learning rates. This approach not only provided a detailed assessment and comparison of each model's performance but also ensured optimal performance when handling PCOS data.

Model Evaluation

When evaluating the constructed PCOS risk prediction models, we used multiple metrics, including accuracy (acc), recall, precision (prec), F1 score, receiver operating characteristic (ROC) curves, and their corresponding area under the curve (AUC) values. We ensured the robustness and generalizability of the evaluation through 5-fold cross-validation and independent testing. These metrics collectively consider the model's correctness in predictions, its ability to identify positive cases, reduced misclassification rates, and balanced classification decisions. The AUC value provides a quantitative measure of overall model performance. During the cross-validation process, we used the average values as the final evaluation results, helping us assess the model's consistency and generalization across different data subsets and ensuring that the selected model can provide reliable predictions in real clinical applications.

Results

Experimental Setup

In this study, the pre-processed and feature-selected dataset was divided, with 80% used as the training set for model training and validation, and the remaining 20% reserved as the test set to evaluate the model's generalization ability. During training, 5-fold cross-validation was employed to optimize model parameters and prevent overfitting, ensuring that each model performed optimally on unseen data. The experiments were conducted in an environment equipped with high-performance computing resources, including a 64-bit operating system and a workstation with an NVIDIA RTX 2080 Ti graphics card. Python programming language was used for experimental coding, making use of its rich scientific computing and data analysis libraries such as NumPy and Pandas for data processing. For the construction and training of machine learning models, we utilized the Scikit-learn library, which provides extensive algorithm support and parameter optimization tools like GridSearchCV, crucial for fine-tuning model parameters.

Feature Selection Results

Results of LASSO Algorithm

In our study, when using the LASSO method for feature selection, we first focused on adjusting the model's complexity. The choice of the regularization parameter alpha was based on the performance of a 10-fold cross-validated logistic regression model. By observing the impact of different alpha values on model accuracy (Figure 2a), we identified an optimal alpha value where the cross-validation score reached its highest point. The region to the left of this optimal alpha value indicates higher model complexity, which may lead to overfitting, while to the right, the model becomes too simplified, potentially missing important information. The gray shaded region in the graph illustrates the range of score variations, providing visual evidence of model stability. Furthermore, by examining the trajectory of LASSO regression coefficients (Figure 2b), we could identify which features maintained non-zero coefficients for longer periods during the regularization process, indicating their greater importance in the model. Each line in the graph represents a feature's coefficient, and as the regularization strength (alpha value) increases, most coefficients tend to zero, while certain feature coefficients remain non-zero at higher regularization strengths, suggesting their significant contribution to the model's predictive power.

We evaluated the importance of the selected features (Figure 2c) and conducted three-dimensional visualization using Principal Component Analysis (PCA) (Figure 2d). From the feature importance graph, we could clearly identify which features had the most significant impact on the model, such as the "Follicle No. R" feature with the highest coefficient, indicating its crucial role in distinguishing between PCOS and non-PCOS patients. PCA visualization revealed

the data distribution in the new feature space, demonstrating a clear boundary between PCOS and non-PCOS patients, despite some individual overlap. This confirms the effectiveness of the features selected by LASSO not only in validating our model but also potentially providing clinically significant indicators for PCOS diagnosis.

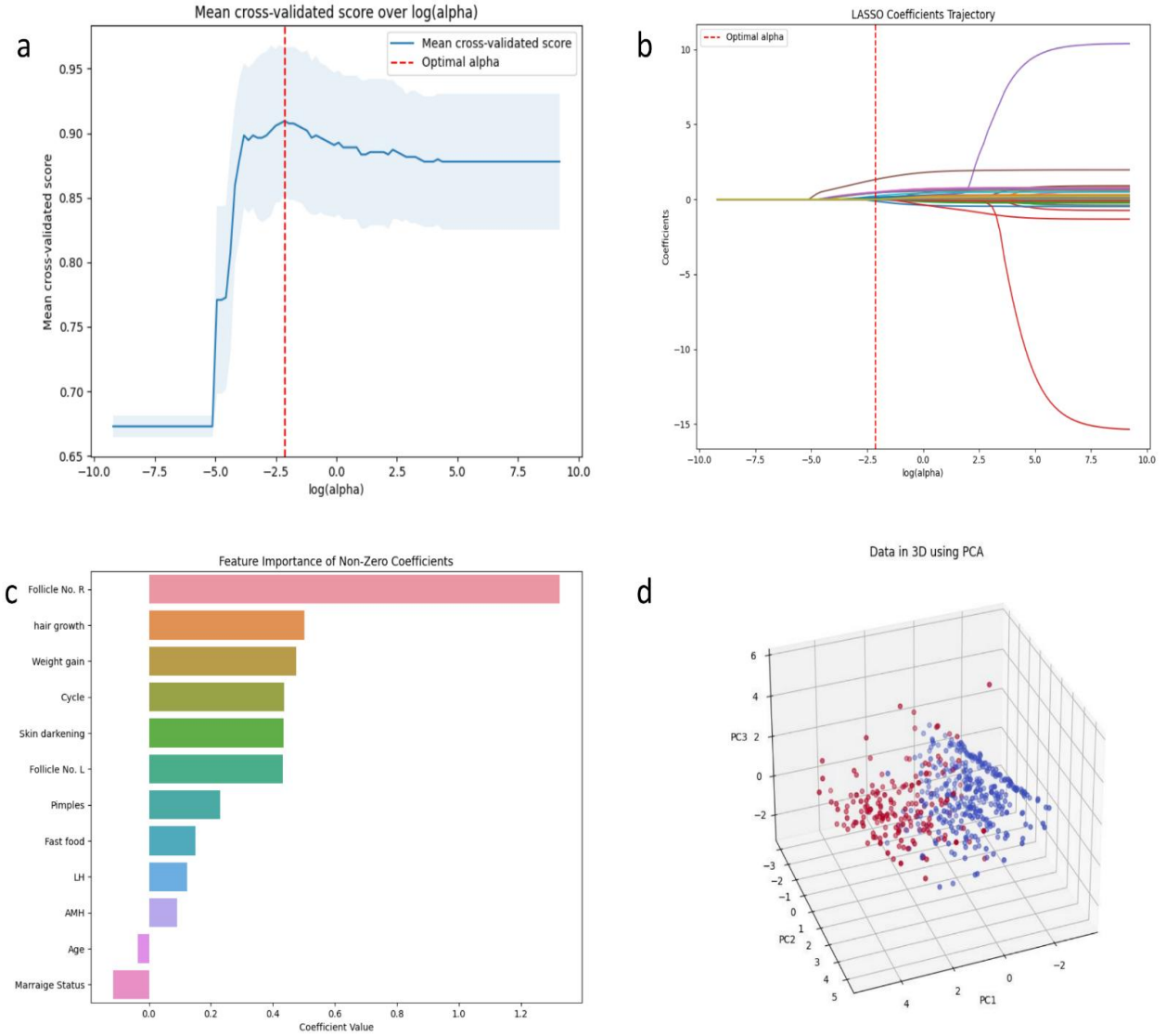

Figure 2: Visualization of Feature Selection in the LASSO Regression Model, where (a) shows the impact of different α values on model accuracy; (b) illustrates the trend of feature coefficients with varying α values; (c) presents the selected features arranged in order of contribution; (d) displays a 3D Principal Component Analysis of the selected features

Distribution Statistics of Selected Features

In this study, we conducted in-depth analysis and clinical interpretation of the selected features. First, the box plots in Figure 3a provided insights into the patterns of the features selected by the LASSO method. For "Follicle No. R" and "Follicle No. L," the box plots revealed the interquartile range of numerical distributions, along with some potential outliers beyond the usual range, indicating extreme variations in follicle counts in certain cases. Box plots for "LH" and "AMH" also displayed relatively wide interquartile ranges, suggesting significant variability in the levels of these biomarkers among the population. The box plot for "Age" showed a relatively consistent distribution, while "Marriage Status" exhibited some higher outliers, possibly indicating a longer marriage history among individual subjects. The correlation

heatmap in Figure 3b showed a strong positive correlation between "Follicle No. R" and "Follicle No. L," which is meaningful in medical terms as they represent follicle counts on the right and left sides. "LH" and "AMH" demonstrated moderate positive correlations with the PCOS label, further supporting their importance in PCOS diagnosis. Low correlations among other features in the heatmap indicated that LASSO had considered feature independence during selection.

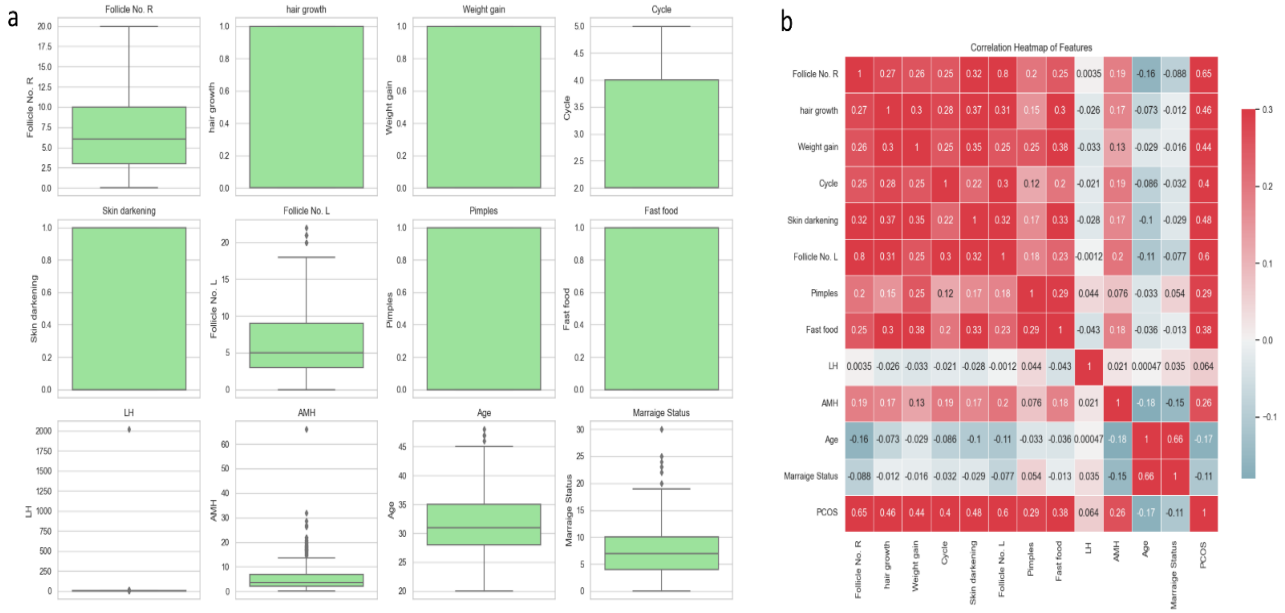

Figure 3: Distribution statistics of selected features, where (a) shows a box plot representation of each feature and (b) presents the correlation matrix between individual features.

Histograms in Figure 4a visually presented the distribution of different feature values. For "Follicle No. R" and "Follicle No. L," the histograms displayed right-skewed distributions, indicating that most samples had fewer follicles, with a minority having more. "LH" and "AMH" distributions also showed right-skewness, with peaks at higher values, suggesting the presence of high-level values. The distribution of "Age" tended toward a normal distribution, while "Marriage Status" exhibited a clear concentration trend, possibly indicating a higher frequency of specific marriage durations. Additionally, we demonstrated feature distributions based on PCOS classification (Figure 4b). "Follicle No. R" and "Follicle No. L" exhibited wider distributions in the PCOS group, implying a higher follicle count among PCOS patients. "LH" and "AMH" showed greater variability in the PCOS group, particularly "AMH," which had a significantly higher distribution in the PCOS group than the non-PCOS group. "Age" had similar distributions in both groups, while "Marriage Status" exhibited different distribution patterns, although whether this difference is directly related to PCOS requires further research for confirmation.

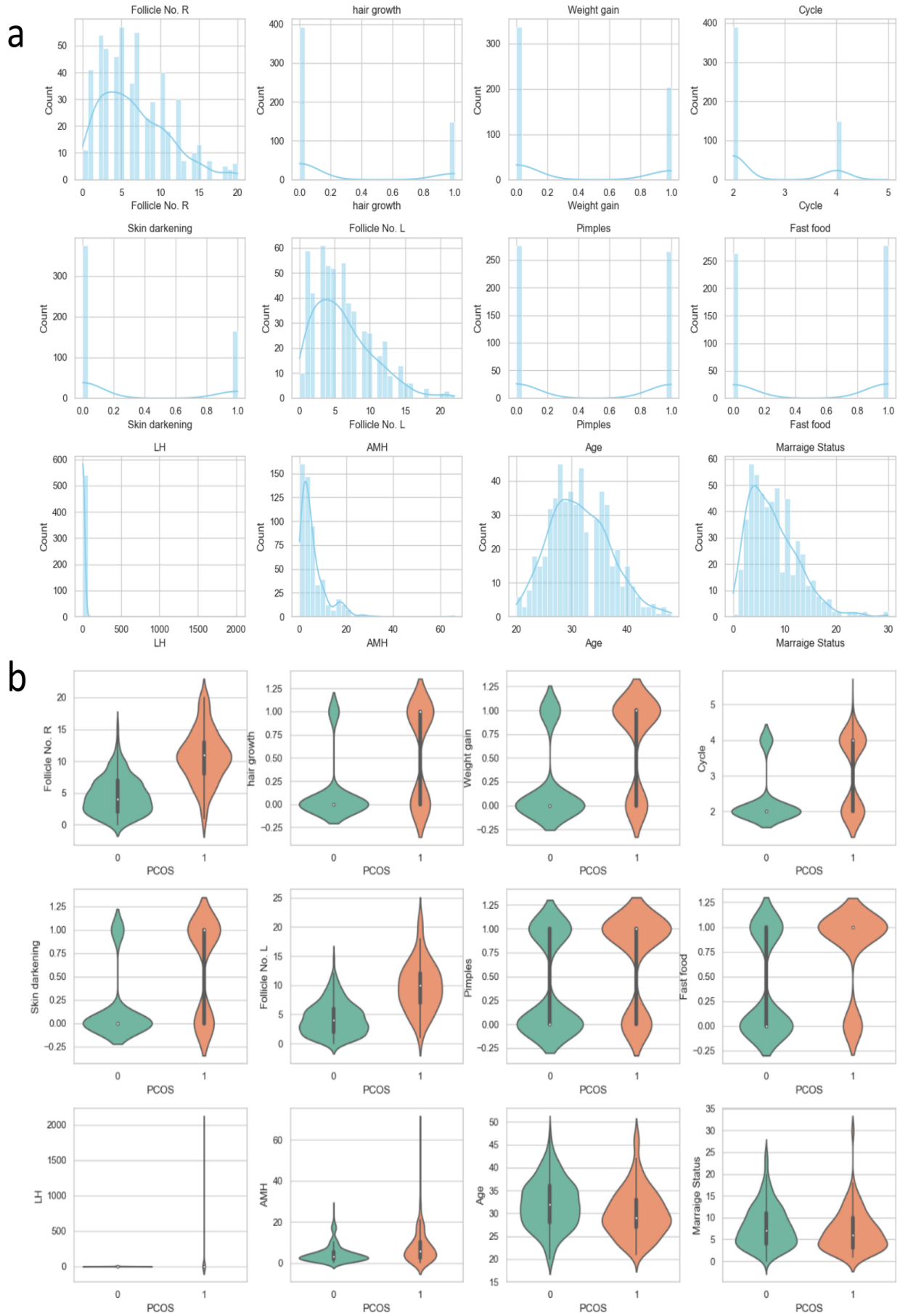

Figure 4: Distribution Statistics of Selected Features, with (a) showing the histogram distribution of each feature and (b) presenting the violin plot for each feature.

Risk Prediction Model Results

Cross-Validation Results

When evaluating the risk prediction models for PCOS, we employed 5-fold cross-validation to ensure robustness and accuracy in the evaluation. The averages of cross-validation results provided an overall statistical performance measure for different models. From Table 1, it is evident that the RF model exhibited outstanding performance across nearly all metrics, particularly in accuracy (CV Accuracy), recall (CV Recall), precision (CV Precision), F1 score (CV F1), and area under the curve (CV AUC), all exceeding the standard of 0.92. This indicates that the random forest model demonstrated high accuracy and reliability in predicting the PCOS dataset. The ensemble learning approach of this model combines predictions from multiple decision trees to enhance prediction accuracy. LDA also showed relatively high accuracy and AUC values, suggesting that even simple linear models can provide good predictive capabilities under certain circumstances. NB and LR similarly demonstrated high precision and F1 scores, possibly due to their probabilistic nature and direct estimation of output probabilities when dealing with such data.

Table 1 Comparison of the performance of various models in cross-validation

Model	CV Accuracy	CV Recall	CV Precision	CV F1	CV AUC
LR	0.8986	0.8919	0.9120	0.8970	0.8982
K-NN	0.8722	0.8754	0.8725	0.8718	0.8721
SVM	0.8456	0.7923	0.902	0.8324	0.8454
DT	0.8454	0.8393	0.8539	0.8417	0.8454
RF	0.9256	0.9224	0.9355	0.9263	0.9256
GB	0.8928	0.8864	0.9027	0.8920	0.8926
GNB	0.9015	0.8860	0.9231	0.8989	0.9011
LDA	0.9076	0.8919	0.9269	0.9046	0.9072
QDA	0.8807	0.8923	0.8782	0.8828	0.8805
MLP	0.8957	0.8868	0.9063	0.8939	0.8956

In contrast, SVM and DT showed slightly weaker performance, especially in recall, which could be attributed to their model complexity not being sufficient to capture all patterns in the data. K-NN and MLP exhibited moderate performance, indicating that these models might benefit from more fine-tuned parameter adjustments to enhance their performance.

Test Set Results

In the test set, RF demonstrated high overall prediction accuracy. Furthermore, it achieved an AUC value of 0.945, indicating its ability to distinguish between PCOS and non-PCOS patients with high accuracy. This result is consistent with the cross-validation findings, further confirming the superiority of the random forest model for this task. Detailed comparative results are shown in Table 2.

Table 2 Comparison of the performance of various models on the test set

Model	Accuracy	Recall	Precision	F1	AUC
LR	0.8672	0.9032	0.7568	0.8235	0.9534
K-NN	0.7823	0.7742	0.6545	0.7094	0.8584
SVM	0.8266	0.8387	0.7091	0.7685	0.9119
DT	0.7897	0.8495	0.6475	0.7349	0.8039
RF	0.8745	0.8710	0.7864	0.8265	0.9450
GB	0.8856	0.8817	0.8039	0.8410	0.9434
GNB	0.8561	0.9462	0.7213	0.8186	0.9530
LDA	0.8745	0.9032	0.7706	0.8317	0.9565
QDA	0.8266	0.9247	0.6825	0.7854	0.9178
MLP	0.8672	0.8602	0.7767	0.8163	0.9061

The average ROC curves depicted in Figure 5 reveal the classification capabilities of various models for true positives and false positives. LDA slightly outperformed random forest with a test AUC score of 0.9565, indicating slightly better discrimination between positive and negative samples, even though their accuracies were similar. This could be attributed to the strong performance of linear discriminant analysis in handling linearly separable data, which can be advantageous in certain cases compared to the nonlinear decision boundaries of random forests. Figure 6 displays the confusion matrices of various models in the test set, providing a detailed distribution of the prediction results.

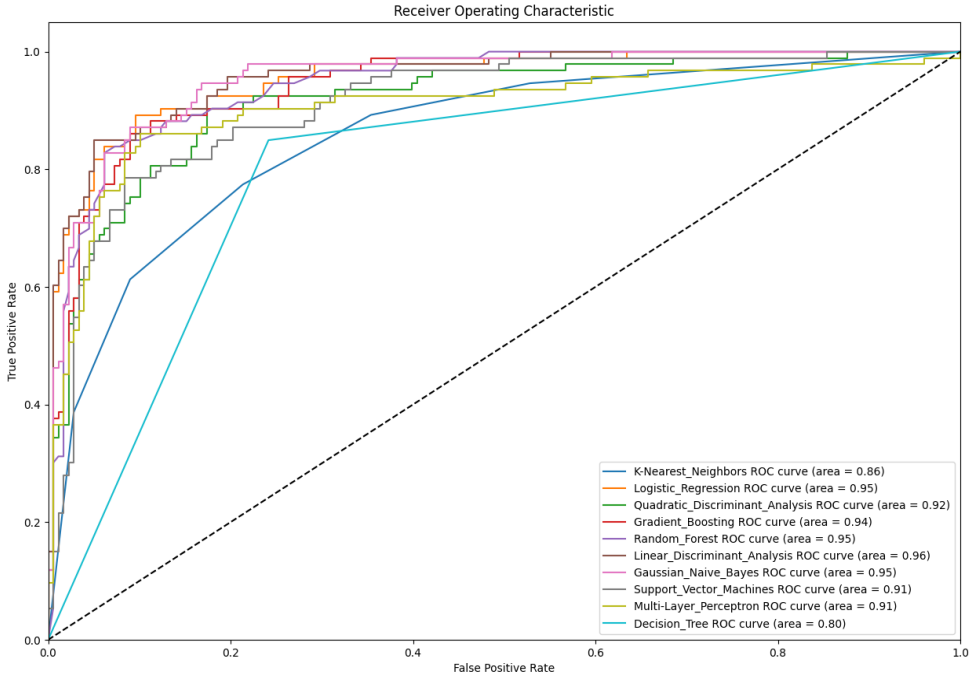

Figure 5: ROC curves and AUC values of each machine learning model.

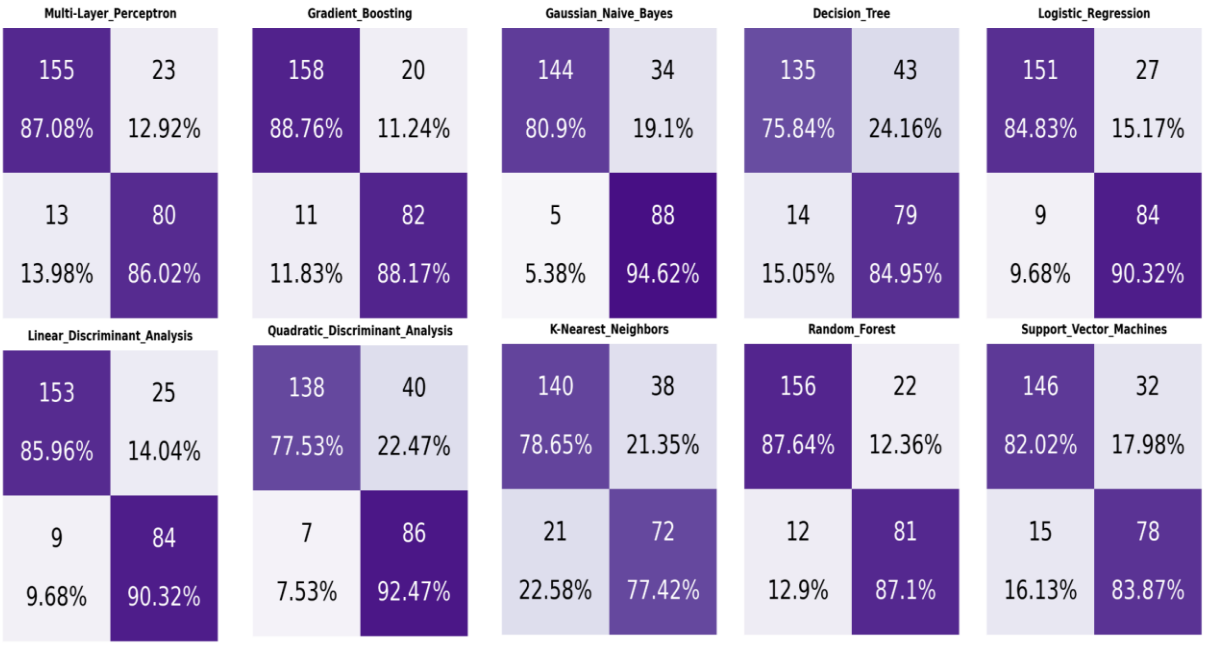

Figure 6: Confusion matrix display of each model.

It is worth noting that SVM and K-NN performed poorly in almost all metrics, especially in precision and F1 score, which may be due to limitations in the performance of these models in high-dimensional spaces and imbalanced datasets. While DT is simple and intuitive, it is prone to overfitting, as reflected in its test AUC value of 0.8039, which is the lowest among all models. Considering all metrics, RF and LDA exhibited the most balanced performance in the test set, displaying stability across various metrics, especially in recall and AUC values. These results offer valuable guidance for future clinical decisions and model selection in the context of PCOS.

Discussion

In this study, we addressed the challenging issue of PCOS diagnosis by developing and evaluating a series of machine learning-based prediction models. PCOS, as a complex endocrine disorder affecting female reproductive health, presents diagnostic challenges due to its involvement of multiple clinical indicators and biomarkers. Initially, we performed feature selection on clinical data using the LASSO regression method to reduce model complexity while retaining the most predictive variables. Subsequently, we employed cross-validation techniques to assess various machine learning models to determine their performance on the training set. During validation, the RF model stood out due to its outstanding performance metrics. Finally, we further validated these models on an independent test set, and the results indicated that both random forest and LDA performed exceptionally well across multiple performance metrics, particularly in their ability to distinguish between PCOS patients and non-PCOS patients. These findings highlight the potential of machine learning to improve PCOS diagnostic methods and predict disease risk with a high degree of accuracy.

Feature engineering and model selection are crucial steps that directly impact the performance of predictive models. The core of feature engineering in this study was LASSO regression, which reduces model complexity through regularization while preserving the most influential features. This approach is particularly suitable for PCOS prediction, as the diagnostic criteria for PCOS involve various clinical indicators, and LASSO effectively selects the most relevant factors from a large pool of potential predictors, reducing unnecessary noise and enhancing model interpretability. In the selection of clinical prediction algorithms, we considered ten different machine learning models, including RF, LDA, SVM, and others. These algorithms were chosen because they have demonstrated high accuracy and reliability in handling classification problems. RF, as an ensemble learning method, is well-suited for medical data like this, as it can handle complex interactions between features and possesses robustness and error tolerance. LDA also exhibited excellent performance in our tests, as its ability lies in finding the optimal linear feature combinations to differentiate between different categories, which is particularly effective when data exhibits clear linear separability. By combining appropriate feature selection and multiple classification algorithms, our study provides a comprehensive framework for accurate PCOS prediction. This approach helps clinicians quickly and accurately identify the most crucial clinical indicators among numerous potential predictive factors for effective PCOS discrimination.

It is noteworthy that the 12 clinical factors selected through LASSO regression have significant clinical relevance for PCOS prediction. These factors include follicle count, hormone levels, reproductive organ characteristics, and lifestyle factors, all of which play important roles in the pathology and physiology of PCOS, aligning with established clinical knowledge. For example, abnormal follicle counts, especially on both sides of the ovaries, are one of the diagnostic criteria for PCOS. In our model, "Follicle No. R" and "Follicle No. L" were identified as features with important predictive value, consistent with clinical observations that PCOS patients typically have more immature follicles. Furthermore, hormone levels such as luteinizing hormone (LH) and anti-Müllerian hormone (AMH) showed a strong correlation with PCOS in our analysis. These hormone indicators are often elevated in PCOS patients and are associated with follicle development and ovulation disturbances. Lifestyle factors, such as "Fast food" consumption frequency, while not direct indicators of PCOS diagnosis, were positively correlated with the incidence of PCOS. Unhealthy dietary habits may

exacerbate insulin resistance, a known risk factor for PCOS development. Other lifestyle-related variables, such as "Weight gain," are closely related to obesity and metabolic syndrome, which are also associated with PCOS. The inclusion of these lifestyle-related variables emphasizes the importance of comprehensive lifestyle changes in PCOS management. Additionally, clinical factors such as changes in skin pigmentation ("Skin darkening") and hair growth ("Hair growth"), which are common clinical manifestations of PCOS, reflect external signs of hormonal imbalance. The inclusion of these features further enhances the clinical applicability of the model.

The machine learning predictive models constructed in this study have significant clinical implications, as they can accurately predict the risk of PCOS in an automated manner, thereby profoundly impacting the improvement of diagnostic processes, the design of personalized treatment plans, and disease management. In clinical practice, early and accurate diagnosis of PCOS is crucial for preventing its related complications, such as diabetes, cardiovascular diseases, and infertility. Machine learning models, through the analysis of patient's historical data and clinical indicators, can help identify those individuals most likely to develop PCOS, even in the early stages when symptoms may not be fully apparent. Furthermore, the predictive capabilities of machine learning models surpass traditional statistical methods because they can handle nonlinear relationships and complex interactions. For example, RF and LDA can not only capture the association of individual features with the risk of PCOS but also explore the mutual influence of multiple features. This deep learning ability means that the models can comprehensively understand the biology and clinical processes of the disease, providing clinicians with more accurate diagnostic support. With the rise of personalized medicine, machine learning-based PCOS prediction models can provide individualized risk assessments for each patient, offering more treatment options and better decision-making guidance for both patients and doctors. Additionally, the application of the model can extend to disease prevention and health promotion by analyzing the impact of lifestyle and environmental factors, providing strategic recommendations for PCOS prevention.

However, this work still has limitations. For example, this study focused solely on PCOS disease risk prediction and did not stratify or classify predictions for clinical symptoms, tests, or long-term impacts on the disease. The data used in this study came from public databases and lacked external test data. Future research should involve multi-center collaboration, external testing of the model based on large samples, and a hierarchical research approach. Furthermore, the model has not yet incorporated traditional Chinese medicine diagnostic indicators. Future efforts can further improve the data by integrating traditional Chinese medicine differentiation with artificial intelligence to build a comprehensive predictive model for PCOS, enriching the disease's diagnosis and treatment methods. Currently, the model is in the exploratory stage and requires further development for clinical translation.

Conclusion

This study explores the application of machine learning in predicting PCOS and confirms the methodological feasibility and accuracy of the PCOS disease prediction model based on machine learning. Through in-depth analysis of a large amount of clinical retrospective data, we carefully screened and analyzed the risk factors for PCOS, and successfully built a machine learning model that can accurately predict the incidence of PCOS, significantly improving the ability to identify patients with high risk factors. Capabilities of risk groups. The application of this model is expected to optimize PCOS prevention strategies and diagnosis and treatment processes, and provide doctors with strong decision-making support. Future research can rely on larger sample sizes and multi-level data to conduct multi-center studies to further verify and improve the robustness and universality of this research model. This will not only bring more personalized prevention and treatment solutions to PCOS patients, but will also lay a solid foundation for the application of artificial intelligence in a wider range of medical fields.

Declaration of Competing Interest

The authors declare no conflict of interest.

Authors' Contributions

L.B.Q. provided the conception and design of the whole project. L.B.Q., X.H.M.,H.Y., were responsible for the statistical analysis. All authors

were involved in the technical support of the proposed models. Z.J.N. assisted with statistical analysis. L.B.Q., X.H.M.and H.Y., were involved in drafting the manuscript, and W.Q. were responsible for reviewing the manuscript. All authors approved the submission of the final manuscript.

Funding

This work was supported by National High Level Hospital Clinical Research Funding (2023-NHLHCRF-YYPPLC-TJ-06).

Supplementary Materials

Not applicable.

Institutional Review Board Statement

This study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Boards of China-Japan Friendship Hospital (Ethical approval number 2022-KY-100).

Reference

1. Khajouei A, Hosseini E, Abdizadeh T, et al. Beneficial effects of minocycline on the ovary of polycystic ovary syndrome mouse model: Molecular docking analysis and evaluation of TNF-α, TNFR2, TLR-4 gene expression. J Reprod Immunol, 2021, 144:103289.
2. Alesi S, Ee C, Moran LJ, Rao V, et al. Nutritional Supplements and Complementary Therapies in Polycystic Ovary Syndrome. Adv Nutr, 2022, 13(4):1243-66.
3. Witchel S.F., E Oberfield S., Peña A.S. Polycystic Ovary Syndrome: Pathophysiology, Presentation, and Treatment With Emphasis on Adolescent Girls. J. Endocr. Soc, 2019, 3:1 545-73.
4. Rostamtabar M, Esmaeilzadeh S, Tourani M, et al. Pathophysiological roles of chronic low-grade inflammation mediators in polycystic ovary syndrome. J Cell Physiol, 2021, 236(2): 824-38.
5. Ajmal N, Khan SZ, Shaikh R. Polycystic ovary syndrome (PCOS) and genetic predisposition: A review article. Eur J Obstet Gynecol Reprod Biol X. 2019 Jun 8;3:100060. doi: 10.1016/j.eurox.2019.100060. PMID:31 403134; PMCID: PMC6687436.
6. Patel S. Polycystic ovary syndrome (PCOS), an inflammatory, systemic, lifestyle endocrinopathy. J Steroid Biochem Mol Biol. 2018 Sep;182:27-36. doi: 10.1016/j.jsbmb. 20 18.04.008. Epub 2018 Apr 17. PMID:29678 491.
7. Dapas M, Dunaif A. Deconstructing a Syndrome: Genomic Insights Into PCOS Causal Mechanisms and Classification. Endocr Rev. 2022 Nov 25;43(6):927-965. doi: 10.1210/endrev/bnac001. PMID: 35026001; PMCID: PMC9695127.
8. Gleicher N, Darmon S, Patrizio P, Barad DH. Reconsidering the Polycystic Ovary Syndrome (PCOS). Biomedicines. 2022 Jun 25;10(7): 15 05. doi: 10.3390/biomedicines10071505. PMID: 35884809; PMCID: PMC9313207.
9. Krishnan A, Muthusami S. Hormonal alterations in PCOS and its influence on bone metabolism. J Endocrinol. 2017 Feb;232(2): R 99-R113. doi: 10.1530/JOE-16-0405. Epub 2016 Nov 28. PMID: 27895088.
10. Swamy S R, KS N P. Hybrid Machine Learning Model for Early Discovery and Prediction of Polycystic Ovary Syndrome [C]//2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE). IEEE, 2022: 1-8.
11. Baweja A K, Kanchana M. Prediction of Polycystic Ovarian Syndrome Using Machine Learning Techniques[C]//Machine Learning, Image Processing, Network Security and Data Sciences: Select Proceedings of
12. Neto C, Silva M, Fernandes M, et al. Prediction models for Polycystic Ovary Syndrome using data mining[C]//International Conference on Advances in Digital Science. Cham: Springer International Publishing, 20 21: 210-221.
13. Denny A, Raj A, Ashok A, et al. i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques[C]//TENCON 2019-2019 IEEE Region 10 Conference (TENCON). IEEE, 2019: 673-678.
14. Zad Z, Jiang V S, Wolf A T, et al. Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records[J]. medRxiv, 2023: 2023.07. 27 .23293255.
15. Nguyen T, Mengersen K, Sous D, Liquet B. SMOTE-CD: SMOTE for compositional data. PLoS One. 2023 Jun 29;18(6):e0287705. doi: 10.1371/journal.pone.0287705. PMID:37384

667; PMCID: PMC10309641.
16. Lee YW, Choi JW, Shin EH. Machine learning model for predicting malaria using clinical information. Comput Biol Med. 2021 Feb;129:104151. doi: 10.1016/j.compbiomed. 2020.104151. Epub 2020 Nov 28. PMID:33 290932.
17. [17] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013 Mar 22;14:106. doi: 10. 1186/1471-2105-14-106. PMID: 23522326; PMCID: PMC3648438.
18. Kang J, Choi YJ, Kim IK, Lee HS, Kim H, Baik SH, Kim NK, Lee KY. LASSO-Based Machine Learning Algorithm for Prediction of Lymph Node Metastasis in T1 Colorectal Cancer. Cancer Res Treat. 2021 Jul;53(3):773-783. doi: 10.4143/crt.2020.974. Epub 2020 Dec 29. PMID: 33421980; PMCID: PMC82 9 1173.
19. Dai P, Chang W, Xin Z, Cheng H, Ouyang W, Luo A. Retrospective Study on the Influencing Factors and Prediction of Hospitalization Expenses for Chronic Renal Failure in China Based on Random Forest and LASSO Regression. Front Public Health. 2021 Jun 15 ;9:678276. doi: 10.3389/fpubh.2021.678 276. PMID: 34211956; PMCID:PMC8239170.
20. Mezzatesta S, Torino C, Meo P, Fiumara G, Vilasi A. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. Comput Methods Programs Biomed. 2019 Aug;177:9-15. doi: 10.1016/j.cmpb.2019.05.005. Epub 2019 May 13. PMID: 31319965.
21. Doğru A, Buyrukoğlu S, Arı M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Med Biol Eng Comput. 2023 Mar;61(3):785-797. doi: 10.1007/s11517-022-02749-z. Epub 2023 Jan 5. PMID: 36602674.