

A Comparative Study of Machine Learning Algorithms for PCOS Diagnosis and Prediction

Nikita Jain, Prof. Savita Ghatte

M.Tech Data Science

Department of Computer Engineering (Data Science)

Abstract

Polycystic Ovary Syndrome (PCOS) is one of the most common endocrine disorders affecting women of reproductive age. Early diagnosis is essential to prevent infertility, metabolic syndrome, cardiovascular diseases, and diabetes. Recent advancements in Machine Learning (ML) have enabled the development of automated PCOS prediction systems using clinical, hormonal, and ultrasound imaging data. This paper presents a comparative study of machine learning algorithms including Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbor (KNN) for PCOS diagnosis and prediction. In addition, ultrasound image segmentation techniques such as Otsu Thresholding and Chan–Vese segmentation are reviewed. The study analyzes the performance, advantages, and limitations of various approaches reported in the literature. The findings indicate that Random Forest and hybrid image-processing-based systems provide superior performance for PCOS diagnosis. Future research directions involving multimodal data integration and explainable artificial intelligence are also discussed.

Keywords— PCOS, Machine Learning, Random Forest, Support Vector Machine, Decision Tree, Ultrasound Imaging, Healthcare Analytics.

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is a hormonal and metabolic disorder affecting approximately 5–15% of women worldwide [4], [5]. The condition is characterized by irregular menstrual cycles, hyperandrogenism, obesity, insulin resistance, and ovarian cyst formation [5], [6].

Traditional diagnostic methods rely on clinical examination, hormonal tests, and ultrasound imaging. However, manual diagnosis is time-consuming and subject to human interpretation errors [1]. Recent developments in Artificial Intelligence (AI) and Machine Learning (ML) have enabled automated systems capable of assisting clinicians in early PCOS detection [1], [2].

The major contributions of this paper are:

1. Review of machine learning techniques used in PCOS diagnosis.
2. Comparative analysis of SVM, DT, RF, NB, LR, and KNN algorithms.
3. Review of ultrasound image segmentation techniques.
4. Identification of research gaps and future opportunities.

II. RELATED WORK

A. Machine Learning-Based PCOS Detection

Prapty and Shitu [1] compared Random Forest, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor algorithms for PCOS prediction using clinical and hormonal data. Their study reported superior performance of Random Forest and Naïve Bayes classifiers.

Several researchers have applied machine learning algorithms to healthcare datasets and demonstrated their effectiveness in disease prediction and classification tasks [17], [18], [20].

B. Ultrasound Image-Based PCOS Detection

Madhumitha et al. [2] proposed an automated follicle recognition framework using ultrasound images and machine learning classifiers. The hybrid model achieved nearly 98% accuracy in identifying PCOS-related abnormalities.

Ultrasound-based approaches offer significant advantages because follicle count and ovarian morphology are important diagnostic indicators [14], [15].

C. Image Segmentation Techniques for PCOS Diagnosis

Nazarudin et al. [3] proposed a hybrid segmentation framework using Otsu Thresholding and Chan–Vese methods for ultrasound image analysis. Their approach improved follicle boundary extraction and segmentation accuracy.

Image segmentation remains a critical step in automated PCOS diagnosis systems because ultrasound images often contain speckle noise and low contrast [3], [16].

III. DATA COLLECTION

A. Dataset Description

The datasets used in PCOS prediction studies are generally collected from hospitals, diagnostic centers, and healthcare databases [1], [13].

B. Clinical and Hormonal Features

Commonly used features include:

- Age
- BMI
- Menstrual Cycle Length
- Weight Gain
- Hair Growth
- Hair Loss
- FSH Level
- LH Level
- AMH Level
- Blood Glucose Level
- Endometrium Thickness

These parameters are widely reported as significant indicators of PCOS [4], [6], [10].

C. Data Preprocessing

The preprocessing phase typically includes:

- Missing value handling
- Data normalization
- Feature scaling
- Class balancing
- Noise reduction

These steps improve model accuracy and reduce bias [17], [19].

IV. MACHINE LEARNING ALGORITHMS

A. Support Vector Machine (SVM)

SVM is a supervised learning algorithm that identifies an optimal hyperplane to separate classes. It performs effectively in high-dimensional datasets and healthcare classification problems [21].

B. Decision Tree (DT)

Decision Trees use hierarchical decision rules to classify data. Their interpretability makes them suitable for medical applications [22].

C. Random Forest (RF)

Random Forest combines multiple decision trees using ensemble learning. It improves prediction accuracy while reducing overfitting [20].

D. Naïve Bayes (NB)

Naïve Bayes applies Bayes' theorem assuming feature independence. It provides fast classification and performs well with healthcare datasets [23].

V. METHODOLOGY

A. Data Collection

Clinical and ultrasound datasets are collected from healthcare sources.

B. Data Preprocessing

Missing values are handled and features are normalized before training.

C. Feature Selection

Relevant clinical and hormonal parameters are selected using statistical methods and feature importance measures.

D. Model Training

Machine learning algorithms are trained using 70:30 and 80:20 train-test splits.

E. Ultrasound Image Segmentation

1) Otsu Thresholding

Automatically determines threshold values for image segmentation [3].

2) Chan–Vese Segmentation

Uses active contour models for object boundary extraction [3].

3) Hybrid Otsu–Chan–Vese Method

Combines thresholding and contour-based segmentation for improved follicle detection.

VI. COMPARATIVE ANALYSIS

Table I. Comparison of Machine Learning Algorithms

Algorithm	Advantages	Limitations
RF	High accuracy	Computational cost
SVM	Effective in high dimensions	Parameter tuning
DT	Interpretable	Overfitting
NB	Fast	Independence assumption
LR	Simple	Limited complexity
KNN	Easy implementation	Slow for large datasets

VII. RESULTS AND DISCUSSION

A. Performance Analysis

Random Forest consistently demonstrates superior performance due to its ensemble structure and ability to handle nonlinear relationships [20].

B. Comparison of Clinical Data-Based Models

Clinical-data-based systems achieve strong performance when hormonal and metabolic parameters are available [1].

C. Comparison of Ultrasound-Based Models

Image-based systems benefit from segmentation and follicle recognition techniques [2], [3].

D. Advantages and Limitations

Hybrid approaches combining clinical and image features provide the most promising results but require larger datasets and increased computational resources.

VIII. RESEARCH GAP

1. Limited Dataset Availability.
2. Lack of Multimodal Systems.
3. Noise in Ultrasound Images.
4. Explainability Issues.
5. Real-Time Clinical Deployment Challenges.
6. Population Generalization Problems.

IX. FUTURE WORK

Future research should focus on the development of multimodal PCOS prediction systems that integrate clinical records, hormonal profiles, genetic information, lifestyle factors, and ultrasound imaging data into a unified framework. The combination of heterogeneous data sources can improve diagnostic accuracy and provide a comprehensive understanding of disease progression. Deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Vision Transformers (ViTs), and hybrid CNN-LSTM models can be explored for automated feature extraction and improved classification performance. Additionally, advanced feature selection and optimization techniques may further enhance model efficiency and generalization capabilities.

Another important direction is the incorporation of Explainable Artificial Intelligence (XAI) techniques to improve transparency and trust in machine learning predictions. Methods such as SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), and attention-based visualization can help clinicians understand the factors influencing diagnostic decisions. Future systems should also focus on real-time deployment through cloud-based healthcare platforms, mobile applications, and decision-support systems for gynecologists. Large-scale multicenter datasets, federated learning frameworks, and privacy-preserving AI models can further improve model robustness, generalizability, and clinical acceptance across diverse populations.

X. CONCLUSION

This comparative study examined various machine learning algorithms and image-processing techniques used for the diagnosis and prediction of Polycystic Ovary Syndrome (PCOS). The reviewed literature demonstrates that machine learning models can effectively assist healthcare professionals in identifying PCOS at an early stage using clinical, hormonal, metabolic, and ultrasound imaging data. Among the analyzed algorithms, Random Forest consistently showed superior performance due to its ensemble learning capability, robustness against overfitting, and ability to handle complex nonlinear relationships among clinical features. Support Vector Machine and Logistic Regression also provided reliable classification results, while Decision Tree, Naïve Bayes, and K-Nearest Neighbor offered simpler and interpretable solutions for specific datasets.

Furthermore, ultrasound image-based approaches have emerged as promising tools for automated PCOS diagnosis. Segmentation techniques such as Otsu Thresholding and Chan–Vese methods improve follicle detection and ovarian morphology analysis, thereby enhancing diagnostic accuracy. Hybrid systems that combine machine learning with image-processing techniques achieve better performance than standalone approaches. Despite these advancements, challenges such as limited dataset availability, class imbalance, image noise, and lack of model interpretability remain significant barriers to widespread clinical adoption. Therefore, further research is required to develop reliable, explainable, and clinically deployable PCOS diagnostic systems.

REFERENCES

- [1] A. S. Prapty and T. T. Shitu, ICCIT, 2020.
- [2] M. Madhumitha et al., IEEE Conference, 2021.

- [3] N. Nazarudin et al., IEEE Conference, 2020.
- [4] Rotterdam ESHRE/ASRM Consensus, Human Reproduction, 2004.
- [5] J. Azziz et al., Nature Reviews Disease Primers, 2016.
- [6] H. Teede et al., International PCOS Guidelines, 2018.
- [7] M. O. Goodarzi et al., Endocrine Reviews, 2011.
- [8] H. F. Escobar-Morreale, Nature Reviews Endocrinology, 2018.
- [9] E. Diamanti-Kandarakis et al., Endocrine Reviews, 2006.
- [10] R. S. Legro et al., Fertility and Sterility, 2013.
- [11] D. A. Dumesic et al., Endocrine Reviews, 2015.
- [12] W. A. March et al., Human Reproduction, 2010.
- [13] D. Dewailly et al., Human Reproduction Update, 2014.
- [14] A. H. Balen et al., Human Fertility, 2016.
- [15] B. Fauser et al., Human Reproduction Update, 2012.
- [16] K. H. Zou et al., Academic Radiology, 2004.
- [17] J. Han and M. Kamber, Data Mining, 2012.
- [18] C. M. Bishop, Pattern Recognition and Machine Learning, 2006.
- [19] T. M. Mitchell, Machine Learning, 1997.
- [20] L. Breiman, Random Forests, 2001.
- [21] C. Cortes and V. Vapnik, SVM, 1995.
- [22] J. R. Quinlan, Decision Trees, 1986.
- [23] R. O. Duda et al., Pattern Classification, 2001.
- [24] S. B. Kotsiantis, Machine Learning Review, 2007.
- [25] A. K. Jain et al., Statistical Pattern Recognition, 2000.
- [26] G. Chandrashekar and F. Sahin, Feature Selection Review, 2014.
- [27] C. Aggarwal, Data Classification Algorithms, 2014.
- [28] M. Kuhn and K. Johnson, Applied Predictive Modeling, 2013.
- [29] A. Géron, Hands-On Machine Learning, 2022.
- [30] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2013.