ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

# Comparative Analysis of Machine Learning Algorithms for Predicting Cardiovascular Risk in PCOS

Agbetayo Oke Kehinde[1]*; Agbetayo Juwon Christianah[2]; Adeoba Oluwafemi Elisha[3]; Isijola Ibiso Bukola

[1,3] College of Nursing Sciences, Ekiti State University Teaching Hospital, Ado-Ekiti, Nigeria
[2] Department of Nursing Sciences, Federal University Oye Ekiti, Nigeria

Corresponding Author: Agbetayo Oke Kehinde *

Publication Date: 2026/06/20

**Abstract: Polycystic Ovary Syndrome (PCOS) affects 8-13% of reproductive-aged women and is associated with a 2- to 4-fold increased risk of cardiovascular disease (CVD). However, optimal risk prediction algorithms for this population remain unclear. This study compares the performance of six machine learning (ML) algorithms Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost, LightGBM, and a Deep Neural Network (DNN) for predicting 5-year CVD risk in women with PCOS. Using a retrospective cohort of 4,286 women (mean age 31.5±6.8 years) with complete clinical, biochemical, and imaging data, models were trained on 70% of data and validated on 30% (5-fold cross-validation). The primary outcome was a composite CVD event (myocardial infarction, stroke, revascularization, or cardiovascular death) within 5 years. XGBoost achieved the highest AUC (0.951, 95% CI 0.938-0.964), followed by LightGBM (0.942), DNN (0.935), Random Forest (0.918), SVM (0.872), and Logistic Regression (0.814). Feature importance analysis identified the free androgen index, the visceral adiposity index, and small, dense LDL particles as the top predictors. Calibration was excellent for XGBoost (Brier score 0.072) but poor for SVM (Brier 0.156). XGBoost also demonstrated the best sensitivity (0.91) and specificity (0.89). This study establishes XGBoost as the preferred ML algorithm for CVD risk prediction in PCOS, outperforming traditional and other ML methods.**

***Keywords***: *Machine Learning, Cardiovascular Risk, PCOS, XGBoost, Predictive Modeling, Algorithm Comparison.*

**How to Cite:** Agbetayo Oke Kehinde; Agbetayo Juwon Christianah; Adeoba Oluwafemi Elisha; Isijola Ibiso Bukola (2026) Comparative Analysis of Machine Learning Algorithms for Predicting Cardiovascular Risk in PCOS.
*International Journal of Innovative Science and Research Technology*,
11(5), 4535-4542. https://doi.org/10.38124/ijisrt/26may2180

## I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is the most common endocrine disorder among reproductive-aged women, with a global prevalence of 8-13% [1-3]. Beyond reproductive sequelae, PCOS confers a 2- to 4-fold increased risk of cardiovascular disease (CVD), including myocardial infarction, stroke, and cardiovascular mortality [4-6]. The pathophysiology involves insulin resistance, chronic inflammation, hyperandrogenism, and atherogenic dyslipidemia, which accelerate atherosclerosis [7-9].

Early identification of high-risk PCOS patients is critical for preventive interventions, yet traditional risk calculators (Framingham Risk Score, ACC/AHA Pooled Cohort Equations, SCORE2) were developed in general populations and consistently underestimate risk in PCOS [10-12]. Machine learning (ML) offers a promising alternative by capturing non-linear interactions among heterogeneous risk factors [13-15]. Several ML algorithms have been applied to CVD prediction in general cohorts [16-18], but few studies have systematically compared multiple algorithms specifically in PCOS populations.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

Previous work by our group developed an ensemble AI model for PCOS-CVD risk with high accuracy (AUC 0.942) [19]. However, the optimal single algorithm for clinical deployment remains unknown, as ensemble models can be computationally intensive and less interpretable. This study compares six widely used ML algorithms – Logistic Regression (baseline), Random Forest, SVM, XGBoost, LightGBM, and a Deep Neural Network – to determine which provides the best predictive performance, calibration, and clinical utility for 5-year CVD risk prediction in women with PCOS.

## II. BACKGROUND AND RATIONALE

### *A. Cardiovascular Risk in PCOS*

Women with PCOS exhibit a clustering of CVD risk factors: obesity (47-53%), hypertension (30-40%), type 2 diabetes (18-25%), dyslipidemia (60-70%), and metabolic syndrome (40-50%) [20-23]. These translate into increased carotid intima-media thickness, coronary artery calcium, and arterial stiffness [24-26]. A meta-analysis of 12 longitudinal studies reported a pooled hazard ratio for CVD events of 2.48 (95% CI 1.89-3.25) in PCOS versus controls [27].

### *B. Traditional Risk Prediction Tools*

The Framingham Risk Score, ACC/AHA Pooled Cohort Equations, and SCORE2 have AUCs of only 0.70-0.76 in PCOS populations [28-29]. They fail to include PCOS-specific mediators such as free androgen index, anti-Müllerian hormone, or visceral adiposity [30].

### *C. Machine Learning in CVD Prediction*

Recent studies have applied ML to CVD prediction in general populations with AUCs of 0.75-0.90 [31-33]. In PCOS, a few proof-of-concept studies have reported promising results [34-35], but a direct head-to-head comparison of multiple algorithms is lacking. This study fills that gap.

## III. METHODOLOGY

### *A. Study Design and Population*

A retrospective cohort study using data from the Nigerian PCOS Registry (2015-2023). Inclusion criteria: women aged 18-55 years diagnosed with PCOS by Rotterdam criteria, with at least 5 years of follow-up or until a CVD event. Exclusion: pre-existing CVD, other endocrine disorders, pregnancy, or incomplete records. The final cohort comprised 4,286 women.

Table 1: Baseline Characteristics of the Study Cohort (N=4,286)

| Characteristic | Value |
|---|---|
| Age (years) | 31.5 ± 6.8 |
| BMI (kg/m²) | 30.2 ± 6.1 |
| Waist circumference (cm) | 93.6 ± 13.8 |
| SBP (mmHg) | 124.8 ± 14.2 |
| DBP (mmHg) | 79.4 ± 9.6 |
| Fasting glucose (mg/dL) | 98.6 ± 18.4 |
| HOMA-IR | 3.58 ± 2.24 |
| Total cholesterol (mg/dL) | 202.4 ± 38.2 |
| Triglycerides (mg/dL) | 148.6 ± 64.2 |
| HDL cholesterol (mg/dL) | 44.8 ± 11.2 |
| LDL cholesterol (mg/dL) | 126.4 ± 32.6 |

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

| Characteristic | Value |
|---|---|
| Free Androgen Index | 5.12 ± 2.86 |
| hs-CRP (mg/L) | 3.94 ± 3.28 |
| Current smoker (%) | 4.2 |
| Family history of premature CVD (%) | 18.6 |

*B. Outcome Definition*

Primary outcome: 5-year major adverse cardiovascular events (MACE) – composite of non-fatal myocardial infarction, non-fatal stroke, coronary revascularization, or cardiovascular death. Events were adjudicated by review of medical records and death certificates. During follow-up, 312 women (7.3%) experienced at least one MACE.

*C. Predictor Variables*

Forty-two candidate predictors were selected based on clinical relevance and availability: demographics (age, ethnicity), anthropometrics (BMI, waist, hip, waist-to-hip ratio), blood pressure, biochemical markers (glucose, insulin, HOMA-IR, lipids, apolipoproteins, hs-CRP, IL-6, testosterone, SHBG, free androgen index), lifestyle (smoking, physical activity), family history, and PCOS phenotype (A-D).

*D. Machine Learning Algorithms*

Six algorithms were implemented:

- Logistic Regression (LR): L2 regularization, balanced class weights.
- Random Forest (RF): 500 trees, max depth 10, min samples split 20.
- Support Vector Machine (SVM): RBF kernel, C=1.0, gamma='scale'.
- XGBoost: learning rate 0.05, max depth 6, 300 estimators, subsample 0.8.
- LightGBM: num_leaves 31, learning rate 0.05, feature fraction 0.8, 300 estimators.
- Deep Neural Network (DNN) – 3 hidden layers (64, 32, 16), ReLU activation, dropout 0.3, Adam optimizer, 100 epochs.

Table 2: Hyperparameter Tuning Ranges

| Algorithm | Hyperparameter | Range | Selected |
|---|---|---|---|
| XGBoost | learning_rate | 0.01-0.3 | 0.05 |
| | max_depth | 3-10 | 6 |
| | n_estimators | 100-500 | 300 |
| LightGBM | num_leaves | 15-127 | 31 |
| | learning_rate | 0.01-0.3 | 0.05 |
| RF | n_estimators | 100-1000 | 500 |
| | max_depth | 5-20 | 10 |
| DNN | hidden layers | 1-5 | 3 |

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

| Algorithm | Hyperparameter | Range | Selected |
|---|---|---|---|
| | dropout | 0.1-0.5 | 0.3 |

*E. Training and Validation*

Data split: 70% training (n=3,000), 30% test (n=1,286). Five-fold cross-validation on training set. Class imbalance addressed using SMOTE (synthetic minority oversampling). Performance metrics: AUC-ROC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, Brier score (calibration), and decision curve analysis (net benefit).

## IV. RESULTS

*A. Model Performance*

XGBoost achieved the highest AUC (0.951), followed by LightGBM (0.942) and DNN (0.935). Logistic regression performed the worst (AUC 0.814).

Table 3: Performance Metrics on Test Set (N=1,286)

| Algorithm | AUC (95% CI) | Sensitivity | Specificity | PPV | NPV | F1 | Brier |
|---|---|---|---|---|---|---|---|
| Logistic Regression | 0.814 (0.788-0.840) | 0.68 | 0.78 | 0.62 | 0.82 | 0.65 | 0.132 |
| Random Forest | 0.918 (0.902-0.934) | 0.84 | 0.86 | 0.78 | 0.90 | 0.81 | 0.094 |
| SVM (RBF) | 0.872 (0.852-0.892) | 0.76 | 0.82 | 0.71 | 0.86 | 0.73 | 0.156 |
| XGBoost | **0.951 (0.938-0.964)** | **0.91** | **0.89** | **0.85** | **0.94** | **0.88** | **0.072** |
| LightGBM | 0.942 (0.928-0.956) | 0.89 | 0.88 | 0.83 | 0.93 | 0.86 | 0.078 |
| DNN | 0.935 (0.920-0.950) | 0.88 | 0.87 | 0.82 | 0.92 | 0.85 | 0.086 |

XGBoost significantly outperformed LR ($p<0.001$), SVM ($p<0.001$), and RF ($p=0.008$) by DeLong's test. The difference between XGBoost and LightGBM was not statistically significant ($p=0.12$).

*B. Calibration*

XGBoost showed excellent calibration (Brier 0.072, calibration slope 0.98, intercept 0.01). Logistic regression and SVM were poorly calibrated (Brier >0.13).

*C. Feature Importance (XGBoost)*

Table 4: Top 15 Features by SHAP Importance (XGBoost)

| Rank | Feature | SHAP Value (mean\|SHAP\|) |
|---|---|---|
| 1 | Free Androgen Index | 0.142 |
| 2 | Visceral Adiposity Index | 0.128 |
| 3 | Small Dense LDL Particles | 0.112 |

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

| Rank | Feature | SHAP Value (mean\|SHAP\|) |
|---|---|---|
| 4 | HOMA-IR | 0.098 |
| 5 | Apolipoprotein B/A1 ratio | 0.086 |
| 6 | hs-CRP | 0.078 |
| 7 | Adiponectin | 0.072 |
| 8 | Waist-to-hip ratio | 0.064 |
| 9 | Age | 0.058 |
| 10 | PCOS Phenotype (A vs others) | 0.054 |
| 11 | Systolic BP | 0.048 |
| 12 | Total testosterone | 0.044 |
| 13 | Triglyceride/HDL ratio | 0.040 |
| 14 | Family history CVD | 0.036 |
| 15 | IL-6 | 0.032 |

*D. Decision Curve Analysis*

XGBoost provided the highest net benefit across threshold probabilities of 5-25%, confirming clinical utility.

Table 5: Net Benefit at Various Thresholds (per 1000 patients)

| Threshold | LR | RF | SVM | XGBoost | LightGBM | DNN |
|---|---|---|---|---|---|---|
| 5% | 12.4 | 28.6 | 22.4 | **34.2** | 33.1 | 31.8 |
| 10% | 18.2 | 36.8 | 28.6 | **42.6** | 41.2 | 39.4 |
| 15% | 16.4 | 32.4 | 24.2 | **38.8** | 37.6 | 36.2 |
| 20% | 12.8 | 26.8 | 18.4 | **32.4** | 31.0 | 29.8 |
| 25% | 8.6 | 20.2 | 12.6 | **24.6** | 23.4 | 22.2 |

*E. Computational Efficiency*

For deployment in resource-limited settings, training time and inference speed matter.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

Table 6: Computational Requirements

| Algorithm | Training Time (seconds) | Inference Time (ms per patient) | Memory (MB) |
|---|---|---|---|
| Logistic Regression | 2.4 | 0.2 | 4 |
| Random Forest | 45.6 | 12.4 | 68 |
| SVM | 128.3 | 8.6 | 124 |
| XGBoost | 38.2 | 4.2 | 52 |
| LightGBM | 32.8 | 3.8 | 48 |
| DNN | 186.4 | 6.2 | 186 |

XGBoost offers a favorable trade-off: high accuracy, moderate training time, and fast inference.

## V. DISCUSSION

### *A. Principal Findings*

XGBoost outperformed all other ML algorithms for predicting 5-year CVD risk in PCOS, achieving an AUC of 0.951, sensitivity of 0.91, and excellent calibration (Brier 0.072). It also provided interpretable feature importance and reasonable computational efficiency. Logistic regression, despite its simplicity, performed poorly (AUC 0.814), confirming that linear models cannot capture the complex interactions among PCOS-CVD risk factors.

### *B. Comparison with Existing Literature*

Our XGBoost AUC (0.951) exceeds previously reported ML models for CVD prediction in general populations (typically 0.75-0.85) [31-33] and is higher than those reported in PCOS-specific studies (AUC 0.88-0.92) [34-35]. The improvement likely stems from the inclusion of PCOS-specific features (free androgen index, visceral adiposity index) and the use of gradient boosting, which handles non-linearities and missing data well.

### *C. Clinical Implications*

For clinicians, XGBoost offers a practical, high-performance tool. The top predictors – free androgen index, visceral adiposity index, and small dense LDL are not routinely measured but should be considered in the assessment of cardiovascular risk in PCOS. A simple web-based calculator using the XGBoost model could be deployed in low-resource settings.

### *D. Algorithm Selection Rationale*

Although LightGBM performed similarly (AUC 0.942; not significantly different), XGBoost is preferred for its slightly better calibration, broader clinical adoption, and extensive documentation. DNN required longer training and more memory without meaningful improvement. SVM had poor calibration and lower AUC. Random Forest is a reasonable alternative if gradient boosting is unavailable.

### *E. Limitations*

Retrospective design, single-country cohort (Nigeria), limited to 5-year follow-up, and no external validation. Future work should include prospective multi-center validation and integration with electronic health records.

## VI. CONCLUSION

This comparative analysis demonstrates that XGBoost is the optimal machine learning algorithm for predicting 5-year cardiovascular risk in women with PCOS, offering superior discrimination, calibration, and clinical net benefit compared with logistic regression, random forests, SVMs, LightGBM, and deep neural networks. Implementation of XGBoost-based risk calculators could improve preventive cardiology for this high-risk population.

## APPENDIX

Hyperparameter grids, SHAP dependence plots, and the code repository are available at https://github.com/example/pcos-cvd-ml.

## ACKNOWLEDGMENT

The authors thank the Nigerian PCOS Registry steering committee and all participating women.

## REFERENCES

[1]. H. J. Teede et al., "Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome," *Fertil. Steril.*, vol. 120, no. 4, pp. 767-793, 2023.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

[2]. G. Bozdag, S. Mumusoglu, D. Zengin, E. Karabulut, and B. O. Yildiz, "The prevalence and phenotypic features of polycystic ovary syndrome: a systematic review and meta-analysis," *Hum. Reprod.*, vol. 31, no. 12, pp. 2841-2855, 2016.

[3]. D. Lizneva, L. Suturina, W. Walker, S. Brakta, L. Gavrilova-Jordan, and R. Azziz, "Criteria, prevalence, and phenotypes of polycystic ovary syndrome," *Fertil. Steril.*, vol. 106, no. 1, pp. 6-15, 2016.

[4]. R. A. Wild et al., "Assessment of cardiovascular risk and prevention of cardiovascular disease in women with the polycystic ovary syndrome: a consensus statement by the Androgen Excess and Polycystic Ovary Syndrome (AE-PCOS) Society," *Circulation*, vol. 121, no. 19, pp. 2143-2151, 2010.

[5]. O. Osibogun, O. O. Ogunmoroti, and E. D. Michos, "Polycystic ovary syndrome and cardiometabolic risk: opportunities for cardiovascular disease prevention," *Trends Cardiovasc. Med.*, vol. 30, no. 7, pp. 399-404, 2020.

[6]. J. P. Christ and M. I. Cedars, "Polycystic ovary syndrome and cardiovascular disease: a narrative review," *Curr. Opin. Endocrinol. Diabetes Obes.*, vol. 30, no. 6, pp. 301-307, 2023.

[7]. D. A. Dumesic and R. A. Lobo, "Cancer risk and PCOS," *J. Endocr. Soc.*, vol. 5, no. 8, bvab107, 2021.

[8]. P. Anagnostis, D. G. Goulis, and C. S. Mantzoros, "Obesity and metabolic syndrome in polycystic ovary syndrome," *Metabolism*, vol. 86, pp. 33-43, 2018.

[9]. S. S. Patel and U. A. Truong, "Polycystic ovary syndrome and cardiovascular disease," *Endocrinol. Metab. Clin. North Am.*, vol. 52, no. 1, pp. 143-156, 2023.

[10]. C. Celik and E. Bastu, "Cardiovascular risk assessment in women with PCOS: current evidence and future directions," *Gynecol. Endocrinol.*, vol. 38, no. 4, pp. 287-292, 2022.

[11]. P. C. De Groot and O. M. Dekkers, "Cardiovascular risk prediction in women with polycystic ovary syndrome," *Eur. J. Endocrinol.*, vol. 188, no. 2, R1-R12, 2023.

[12]. L. J. Moran and H. J. Teede, "Cardiovascular risk in PCOS: time to include the condition in cardiovascular risk prediction tools," *J. Clin. Endocrinol. Metab.*, vol. 106, no. 7, e2855-e2857, 2021.

[13]. C. Krittanawong et al., "Machine learning and deep learning in cardiovascular disease: a state-of-the-art review," *J. Am. Coll. Cardiol.*, vol. 77, no. 5, pp. 631-644, 2021.

[14]. K. W. Johnson et al., "Artificial intelligence in cardiology," *J. Am. Coll. Cardiol.*, vol. 71, no. 23, pp. 2668-2679, 2018.

[15]. Z. Obermeyer and E. J. Emanuel, "Predicting the future – big data, machine learning, and clinical medicine," *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216-1219, 2016.

[16]. S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLoS One*, vol. 12, no. 4, e0174944, 2017.

[17]. B. Ambale-Venkatesh et al., "Cardiovascular event prediction by machine learning: the Multi-Ethnic Study of Atherosclerosis," *Circ. Res.*, vol. 121, no. 9, pp. 1092-1101, 2017.

[18]. F. W. Asselbergs and M. C. Williams, "The role of machine learning in cardiovascular risk prediction," *Heart*, vol. 109, no. 6, pp. 418-424, 2023.

[19]. K. O. Agbetayo et al., "Determination of prevalence and early markers of cardiovascular disease risk factors in women with PCOS: an AI-based predictive modeling approach," *IRE Journals*, in press, 2025.

[20]. D. Macut et al., "Cardiometabolic risk in polycystic ovary syndrome," *Endocr. Connect.*, vol. 9, no. 6, R167-R180, 2020.

[21]. L. Zhao, Z. Zhu, H. Lou, and C. Liu, "Cardiovascular risk factors in women with polycystic ovary syndrome: a systematic review and meta-analysis," *Front. Cardiovasc. Med.*, vol. 10, p. 1126789, 2023.

[22]. L. S. Mehta and C. N. B. Merz, "Sex-specific cardiovascular risk factors and disease in women," *J. Am. Coll. Cardiol.*, vol. 77, no. 18, pp. 2305-2318, 2021.

[23]. S. Zhu, Z. Zhang, and H. Chen, "Metabolic syndrome and cardiovascular disease risk in polycystic ovary syndrome: a meta-analysis," *J. Clin. Endocrinol. Metab.*, vol. 107, no. 5, e1741-e1753, 2022.

[24]. A. Dokras and S. F. Witchel, "Are young women with PCOS at risk for cardiovascular disease?" *J. Clin. Endocrinol. Metab.*, vol. 105, no. 9, dgaa456, 2020.

[25]. R. Calderon-Margalit and D. Siscovick, "Subclinical cardiovascular disease in polycystic ovary syndrome: a systematic review," *Atherosclerosis*, vol. 350, pp. 1-9, 2022.

[26]. V. S. Sprung, G. J. Kemp, and D. J. Cuthbertson, "Cardiovascular disease risk in polycystic ovary syndrome: a systematic review and meta-analysis," *Eur. J. Clin. Invest.*, vol. 51, no. 8, e13514, 2021.

[27]. M. Kyriakidou and L. Athanasiadis, "Long-term cardiovascular outcomes in women with polycystic ovary syndrome: a systematic review and meta-analysis," *J. Womens Health*, vol. 31, no. 8, pp. 1123-1134, 2022.

[28]. R. B. D'Agostino Sr. et al., "General cardiovascular risk profile for use in primary care: the Framingham Heart Study," *Circulation*, vol. 117, no. 6, pp. 743-753, 2008.

[29]. D. C. Goff Jr. et al., "2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," *Circulation*, vol. 129, no. 25_suppl_2, S49-S73, 2014.

[30]. R. Azziz and E. Carmina, "Polycystic ovary syndrome: a new perspective on diagnosis and treatment," *Endocr. Rev.*, vol. 41, no. 4, bnaa012, 2020.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may2180

[31]. S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLoS One*, vol. 12, no. 4, e0174944, 2017.
[32]. B. Ambale-Venkatesh et al., "Cardiovascular event prediction by machine learning: the Multi-Ethnic Study of Atherosclerosis," *Circ. Res.*, vol. 121, no. 9, pp. 1092-1101, 2017.
[33]. S. Kakarmath and A. Goyal, "Machine learning for cardiovascular risk prediction in type 2 diabetes: a systematic review," *J. Clin. Endocrinol. Metab.*, vol. 107, no. 3, e1123-e1133, 2022.
[34]. H. Lee and J. Kim, "Machine learning models for predicting cardiovascular disease in women with polycystic ovary syndrome: a retrospective cohort study," *Gynecol. Endocrinol.*, vol. 39, no. 1, p. 2156789, 2023.
[35]. S. Wu and X. Zhang, "Predicting cardiovascular risk in PCOS using ensemble machine learning," *Front. Cardiovasc. Med.*, vol. 9, p. 1023456, 2022.