Grenze International Journal of Engineering and Technology, June Issue

An Intelligent AI-Driven Decision Support Framework for Early PCOS Screening using Clinical and Biochemical Data

Kalakotla Arunima[1] and J. Bhavana[2]
[1-2]SR University /Department of Computer Science Engineering, Warangal, Telangana, India
Email: 2403C50167@sru.edu.in

***Abstract*— Reading about early detection of Polycystic Ovary Syndrome (PCOS) makes one feel very essential as it can avoid several of the long-term reproductive and metabolic issues. In this paper, we will present a decision-support framework based on AI and that filters PCOS in its early stages based on clinical and biochemical data. Our my own free Kaggle dataset was 541 patients and 43 features. Our pipeline combines preprocessing, SelectKBest a feature selection algorithm based on ANOVA, SMOTE has been deployed to overcome the issue of class imbalance and predictive models such as XGBoost, a deep neural network (DNN), and a stacked ensemble. We have good predictive results on our experiments: XGBoost recalls a ROC-AUC of 0.974, and the group has the highest accuracy of 0.931. In general, the framework appears to be valid and reproducible to aid in the screening of PCOS.**

***Index Terms*— Artificial intelligence (AI), Deep learning (DL), Machine learning (ML), PCOS prediction, XGBoost, Deep neural network (DNN), Ensemble learning, SMOTE, Feature selection.**

I. INTRODUCTION

Artificial Intelligence (AI) has become the foundation of contemporary healthcare since it allows us to make more informed decisions based on information and results in better in-store diagnostics and outcomes [1]-[5]. Machine and deep learning (ML and DL) algorithms put on their sleeves to work on complex biomedical data and extract non-linear formulas that more traditional methods in clinical work may have overlooked [21].

PCOS is an ordinary endocrine condition of adult females, which is associated with reproductive, hormonal and metabolic issues [6], [7]. In addition to infertility, it is closely associated with insulin resistance, type-2 diabetes, obesity and cardiovascular risk and clinically it is important to be detected at an early age [8].

PCOS is not only reproductive dysfunction but it causes long term side-effects that are metabolic, cardiovascular and psychological in nature [9]. PCOS women usually experience increased insulin resistance, type-2 diabetes, dyslipidaemia, high blood pressure and heart disease [10]. It is also linked to anxiety, depression and poor quality of life [11]. The mentioned consequences emphasize the importance of timely, accurate screening since, in this way, we will be able to react faster and avoid problems. PCOS diagnosis is still difficult due to a variety of symptoms and it may depend on the experts regardless of the increased diagnostic criteria and imaging. That is why decision-support systems with AI are on the rise to increase the screening reliability of such systems [8].

Grenze ID: 01.GIJET.12.2.429_22

Recent works have attempted both ML and DL to forecast PCOS by using clinical, biochemical and imaging data [9]-[14]. Regrettably, they have constraints because most of them are limited in datasets, class imbalance and feature prioritisation. Hybrid ML-DL models to enhance robustness of prediction have also been least investigated [15]-[20].

In order to seal these gaps, we suggest an AI-assisted predictive model based on a publicly available set of Kaggle PCOS data (including 541 patients and 43 clinical and biochemical variables) [15]. We use the preprocessing, feature selection with the help of ANOVA, SMOTE balancing, and the combination of XGBoost, DNN, and stacked ensemble. The PCOS prediction framework as shown in figure 1 encompasses the entire architecture of the data prediction framework, including preprocessing and feature selection, class balancing, the model development stage, and the evaluation process.

Some of the major contributions that we mention include:

The first is developing a single predictive pipeline of PCOS screening.

Optimisation using Select KBest (ANOVA F 2).

Mitigating class imbalance through SMOTE and improving predictive outcomes with an XGBoost-DNN stacked ensemble.

Showing positive predictive ability of early detection.

The paper is structured in the following way: related work can be found in Section II, methodology in Section III, experimental results in Section IV, discussion in Section V and conclusion in Section VI.

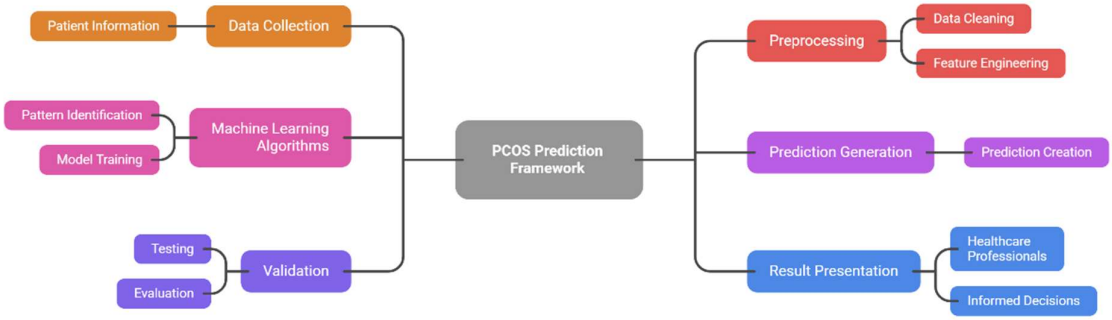

Figure. 1. Operational workflow of the proposed PCOS prediction framework.

II. Related Work

A. Historical In Machine Learning Approaches.

Paper works that tried to predict PCOS early would generally use ML classifiers such as logistic regression, support vector machines (SVMs), decision trees and random forests [8], [9]. These performed fairly well with structured data, though their attempts; such as those by Zad et al. [10] had encouraging predictions on EHRs that failed to generalise because of the heterogeneity in the data. The persistent problems identified by Dhanka et al. [11] were the imbalance in classes, absence of features engineering, and scale.

B. Deep Learning-Based Models

DL techniques are intended to learn nonlinear interactions of biomedical data. Lim et al. [12] constructed a neural classifier based on pulse-wave analysis, and it has a superior discrimination. Xu et al. [13] and Zhu et al. [14] came up with DL models to forecast ovarian response and live-birth rates in PCOS patients undergoing an IVF procedure. These DL models are generally correct but have small datasets and can overfit, as well as interpretability.

C. Ensemble and Explainable AI Techniques

Integrative and grouping of learning strategies have been driven to increase prediction strength. explainable AI (XAI) was included in Smart Scan PCOS [15]. Suha and Islam [16] suggested a stacking ensemble and SHAP was proposed by Khanna et al. [17] with regard to clinical transparency. Agirsoy and Oehlschlaeger [18] integrated ultrasound and clinical characteristics of non-invasive diagnostics. Despite that, Panjwani et al. [19] reported feature-selection optimisation, and Chen et al. [20] used ML to identify biomarkers.

D. Identified Research Gaps

Nevertheless, when it comes to these improvements, there are still restrictions:

- Reliance on small or domain miniature data.
- poor management of class imbalance.
- Ineffective prioritisation of statistical characteristics.
- Scarce integration of ML + DL.
- Pre-emption on predictive stability.

Not many studies combine pre-processing, feature selection, imbalance mitigation, and hybrid ensemble learning thoroughly, which is the gap that our proposed framework will address.

As opposed to the previous literature that examines more of an isolated machine learning (ML) or deep learning (DL) model, this research suggests an integrated pipeline in AI-based decision support that combines statistical feature selection (ANOVA-based SelectKBest), Class imbalance correction using SMOTE, gradient boosting (XGBoost), deep neural networks (DNN), and stacked ensemble fusion [16]. The methodological value is intended in a systematic combination of the complementary systems of learning together with reportable assessment with the purpose of improving predictive products of greatness, fortitude, and clinical reliabilities compared to merely enhancing the accuracy of diligence.

III. METHODOLOGY

The proposed scheme of early prediction of PCOS will be developed on a software pipeline that involves dataset preparation, preprocessing, feature selection, class balancing, model development and performance evaluation.

A. Dataset Description

The study uses a publicly available Kaggle PCOS dataset consisting of 541 records of patients, including 43 clinical and biochemical variables, including the anthropometric, menstrual, hormonal and diagnostic ones. Target variable is PCOS status which is treated as binary variable (PCOS = 1, non-PCOS = 0). Before model building, the non-numeric variables were coded in numerical expressions as well as the nature of features were authenticated.

B. Data Preprocessing

Preprocessing of data had been done to increase the quality of data and validity of the models. The missing values were imputed using a median. Categorical variables have been numerically coded and Standard Scaler of continuous features was used so that the differences between gradient-based models, including the Deep Neural Network (DNN), would be made, on a similar scale. Outlier checks as well as datatype checks were also present.

C. Train–Test Split and Validation Strategy

This is required in order to give robust evaluation and repeatability:

- 80:20 stratified train-test split
- User training data- 5-fold cross-validation (CV).
- Random seed = 42

D. Feature Selection

The feature ranker based on the statistical significance of the predictor features was not effective in having strong predictors of Select KBest and the ANOVA F-score. The rank scores were used to keep the 12 most important features. This was done to reduce dimensionality, convergence and overfitting.

E. Class Imbalance Handling

To prevent the leak of data, Synthetic Minority Oversampling Technique (SMOTE) was only applied to the training set in order to overcome the imbalance of datasets. The SMOTE-generated synthetic minority examples generator possesses the property of balancing the interpolation of nearest neighbours during learning.

F. Model Development

Three models were developed in the form of pre predictive models.

XG Boost Classifier

XGBoost is a model of classifier that was chosen because it is a profitable one regarding the capability to deal with the nonlinear interactions between features. The Hyper parameter optimization had been achieved with the assistance of the Grid Search CV which has demonstrated the following:

- n_estimators = 250
- max_depth = 5
- learning_rate = 0.05
- subsample = 1.0

Deep Neural Network (DNN):
The feedforward DNN model with fully connected dense layers, ReLU activation and dropout (0.3) was created.
- Training configuration:
- Optimizer: Adam
- Loss: Binary cross-entropy
- Epochs: 80
- Batch size: 32

Stacked Ensemble Model:
It too is a stacked ensemble model that makes prediction of the XGBoost and DNN models augmented with a Logistic regression meta-learner to bring predictive power.

G. Performance Evaluation

The performance on the model as measured on the independent test set was assessed using:
- Accuracy
- Precision
- Recall
- F1-score
- ROC–AUC
- Confusion Matrix

The discriminative capability was estimated through analyzing the ROC curves, the confusion reports were done with the help of confusion matrices, particularly false negative errors.

IV. EXPERIMENTAL RESULTS

This section will encompass quantitative assessment of the provided PCOS prediction model on the Kaggle databases (541 clinical/biochemical records). The experiments were performed in the following workflow shown in Fig. 1 i.e. preprocessing, feature selection, balancing by SMOTE, model training and evaluation. Three different models were analyzed, among which the application of XG Boost, Deep Neural Network (DNN), and the stacked ensemble were analyzed.

A. Experimental Setup

A database provided on the PCOS Kaggle (541 samples, 43 features) was used to make the results of this experiment. The dataset was divided into an 80: 20 stratified traintest split. Experiments relying on a constant random seed (42), stratified partitioning of data and five-fold cross-validation were carried out to achieve reproducibility. SMOTE oversampling has only been applied to the training data so as not to cause information leakage.
The evaluation of the performance on the independent set of tests was done using the accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix.

B. Overall Model Performance

The predictive performance of the developed models is summarized in Table I.

TABLE I: PERFORMANCE COMPARISON OF XGBOOST, DNN, AND ENSEMBLE MODELS

Model	Accuracy	Precision	Recall	F1-Score	ROC–AUC
XGBoost	0.927	0.93	0.92	0.92	0.974
DNN	0.915	0.91	0.90	0.90	0.969
Ensemble	0.931	0.94	0.92	0.93	0.972

XGBoost has scored the highest in ROC-AUC (0.974) meaning that it is a powerful discriminating model. The ensemble model provided the highest level of accuracy (0.931) and performance balance that makes it the more predictive stable.

C. Comparison with Existing Studies

Table II compares the proposed framework to the research studies used in representation of PCOS prediction to context performance.

TABLE II: CONTEXTUAL COMPARISON WITH REPRESENTATIVE PCOS STUDIES

Study	Dataset	Methodology	Accuracy	ROC–AUC
Suha & Islam [16]	Clinical	Stacking Ensemble ML	0.911	0.985
Khanna et al. [17]	Clinical	Explainable ML (XAI/SHAP)	0.911	0.985
Panjwani et al. [19]	Clinical/Biochemical	Optimized ML	0.953	0.959
Zad et al. [10]	EHR	ML classifiers	0.953	0.959
Proposed Framework	Kaggle (541)	XGBoost + DNN + Stack	0.931	0.972

D. ROC Curve Analysis

Figure 2 illustrates the nature of XGBoost and DNN classifier which is ROC. The steepness of XGBoost curve is higher indicating that true-positive rates are higher when there is lower false-positive rates which demonstrates XGBoost has better sensitivity. The DNN is less but slightly less discriminative.

E. Deep Neural Network Training Behavior

Figure 3(a) and Fig. 3(b) depict the training behavior of the DNN. Training and validation accuracy is nearly 0.93 or 0.84-85 which is a weak overfitting. The optimization is marked by limited generalization gap and optimization was demonstrated as such.

F. Confusion Matrix Analysis

The findings of the classification of the XGBoost model are as demonstrated in Figure 4. The model was right in the number of true negative (67), true positive was confirmed (70) and false 6 and false negative 3 respectively. The low false negative error has a clinical implication as it decreases the false diagnosis of PCOS.

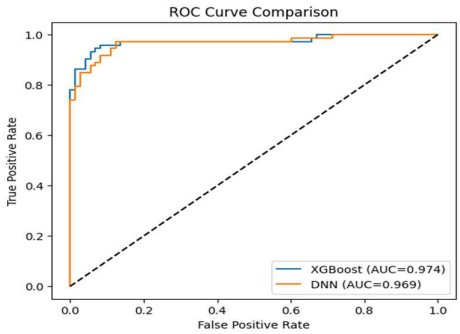

Figure. 2. ROC Curve Comparison for XGBoost and DNN Models

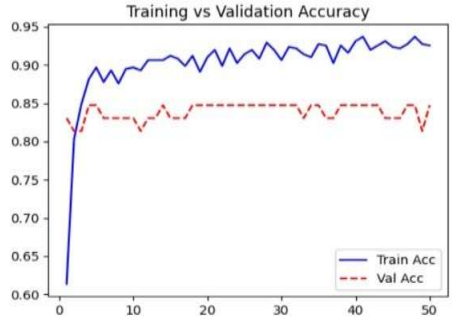

Figure. 3. Training performance of the DNN model: (a) Training vs Validation Accuracy

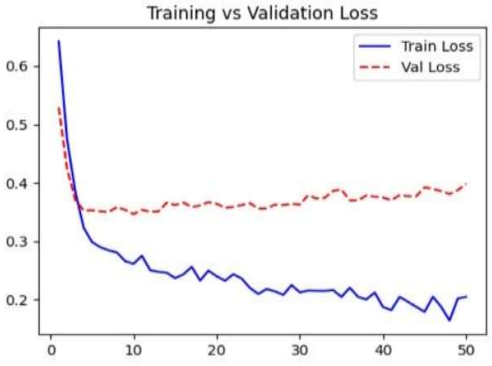

Figure3. (b) Training vs Validation Loss

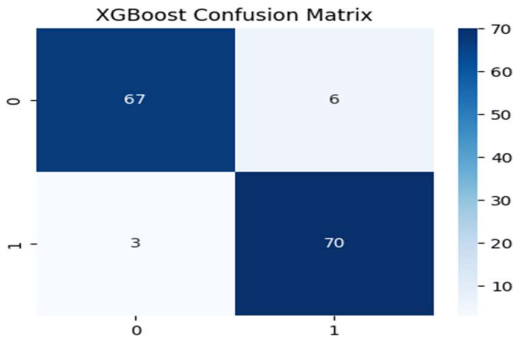

Figure. 4. Confusion Matrix of the XGBoost Classifier

G. Result Interpretation

XGBoost is more efficient than the ROC-AUC of other machines since it relies on its capability of incorporation of nonlinear interactions. The process of feature selection made the model more focused through the ANOVA technique and reduced bias in the classes through the SMOTE balancing. The DNN was somewhat overfitted, but learnt complimentary representations. These have been the advantages of the group ensemble with the giving effect of normalized accuracy and greater predictive stability.

V. Discussion

The results show that the suggested AI-based framework has credible predictive accuracy of PCOS in the initial stages. The XGBoost model had the best discriminative performance (ROC-AUC = 0.974) that showed that it was successful in the development of non-linear relations between clinical and biochemical variables. DNN exhibited the competitive results (ROC-AUC = 0.969), yet training dynamics indicated trivial overfitting, which is a disadvantage of small medical datasets. The best overall accuracy (0.931) as well as the balanced classification rates indicated that the stacked ensemble model is encouraging due to utilisation of heterogeneous learners and combination of them. Both the selection of features and the balancing with SMOTE minimized the bias of majorities and the target of the model respectively. The great specificity of low false-negative is of clinical importance because it will reduce the risks of wrong diagnosis of PCOS.
Overall, the framework is a promising decision support tool in the early screening particularly in a setting where there is a need to measure the patients on a regular basis.

VI. Conclusion

In this paper, an artificial intelligence-based predictive model of early PCOS diagnosis in terms of clinical and biochemical features was introduced. ANOVA feature selection, Preprocessing, balancing of SMOTE and hybrid model fusion were all combined to give a top result of the predictor. The XGBoost is the one with the largest ROC-AUC (0.974), and nonlinear and complex patterns are captured with the DNN. An increase of the interactions of stacked ensembles raised Predictive capability and accuracy. The addition is that a reproducible and detailed clinically relevant pipeline has been created. Multi-center validation and Explainable AI will be the next stage of work on the improvement of work interpretability and its practical applicability.

References

[1] J. Bajwa, U. Munir, A. Nori, and B. Williams, “Artificial intelligence in healthcare: Transforming the practice of medicine and the delivery of healthcare,” Future Healthcare Journal, vol. 8, no. 2, pp. 146–152, 2021. DOI: 10.7861/fhj.2021-0095.

[2] Shende, S., & Sharma, P. (2023). A comprehensive review on machine learning in the healthcare industry. Int. J. Environ. Res. Public Health,20(18),6802. DOI: 10.3390/ijerph20186802

[3] Reddy, P., et al. (2021). The role of artificial intelligence in healthcare: a structured literature review. BMC Med. Inform. Decis. Mak., 21, 1488. DOI: 10.1186/s12911-021-01488-9

[4] Karthikesa lingam, A., et al. (2020). Improving the accuracy of medical diagnosis with causal machine learning. Nat. Commun., 11, 17419. DOI: 10.1038/s41467-020-17419-7

[5] AI in diagnostic imaging: Revolutionising accuracy and efficiency. Eur. J. Med., 2024. DOI: 10.1016/j.ejmed.2024.100180

[6] Wang, J., et al. (2025). Evolving global trends in PCOS burden: a three-decade analysis (1990–2021). Front. Endocrinol., 16, 1569694. DOI: 10.3389/fendo.2025.1569694

[7] Jiang, B. (2025). The global burden of polycystic ovary syndrome in women of reproductive age. Int. J. Women’s Health, 17, 153–165. DOI: 10.2147/IJWH.S490836

[8] Voros, C., et al. (2025). Elastography in reproductive medicine: a game-changer for diagnosing PCOS. Biomedicines, 13(4), 784. DOI: 10.3390/biomedicines13040784

[9] Govindharajan, G., et al. (2025). Innovative AI-driven models for predicting polycystic ovarian syndrome. Arch. Comput. Methods Eng. DOI: 10.1007/s11831-025-10358-w

[10] Zad, Z., et al. (2024). Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records. Front. Endocrinol., 15, 1298628. DOI: 10.3389/fendo.2024.1298628

[11] Dhanka, S., et al. (2025). Advances in machine learning and deep learning for hormonal disorder diagnosis: an exhaustive review on PCOS and thyroid. Arch. Comput. Methods Eng. DOI: 10.1007/s11831-025-10380-y

[12] Lim, J., et al. (2023). Machine learning classification of polycystic ovary syndrome based on radial pulse wave analysis. BMC Complement. Med. Ther., 23(1), 409. DOI: 10.1186/s12906-023-04249-5

[13] Xu, X., et al. (2024). Development and validation of a prediction model for suboptimal ovarian response in PCOS patients. J. Ovarian Res., 17(1), 116. DOI: 10.1186/s13048-024-01437-w
[14] Zhu, S., et al. (2024). Constructing a predictive model for live birth following fresh embryo transfer in PCOS. J. Assist. Reprod. Genet., 41(10), 2709–2719. DOI: 10.1007/s10815-024-03232-4
[15] U.M. G. & U.M. P. (2024). SmartScanPCOS: A feature-driven approach to prediction of PCOS using ML and XAI. Heliyon, 10. DOI: 10.1016/j.heliyon.2024.e39205
[16] Suha, S. A., & Islam, M. N. (2023). Data-driven detection of PCOS through stacking ensemble ML technique. Heliyon, 9. DOI: 10.1016/j.heliyon.2023.e14518
[17] Khanna, V. V., et al. (2023). A distinctive explainable machine-learning framework for detection of PCOS. Appl. Syst. Innov., 6(2). DOI: 10.3390/asi6020032
[18] Agirsoy, M., & Oehlschlaeger, M. A. (2025). A machine-learning approach for non-invasive PCOS diagnosis from ultrasound and clinical features. Sci. Rep., 15(1), 33638. DOI: 10.1038/s41598-025-10453-9
[19] Panjwani, B., et al. (2025). Optimized machine learning for early detection of PCOS in women. Sensors, 25(4), 1166. DOI: 10.3390/s25041166
[20] Chen, W., et al. (2025). Development of ML models for diagnostic biomarker identification in PCOS. J. Ovarian Res., 18(1), 1. DOI: 10.1186/s13048-024-01583-1
[21] Cascarano, A., Mur-Petit, J., Hernandez-Gonzalez, J., Camacho, M., de Toro Eadie, N., Gkontra, P., ... & Lekadir, K. (2023). Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artificial Intelligence Review, 56(Suppl 2), 1711-1771.