Diagnosis of PCOS in Adolescent Girls Using Traditional and Ensemble Machine Learning Methods

Priyanka Pariyawala
priyanka_pariyawala@yahoo.com

Veer Narmad South Gujarat University
Pushpal Y Desai
Veer Narmad South Gujarat University

Research Article

Keywords: Machine Learning, Polycystic ovarian syndrome (PCOS, feature selection, disease prediction

Posted Date: August 8th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7009734/v1

Additional Declarations: No competing interests reported.

Diagnosis of PCOS in Adolescent Girls Using Traditional and Ensemble Machine Learning Methods

Priyanka Pariyawala (0009-0007-7836-3943) [1], Dr Pushpal Y Desai [2]

[1] Research Scholar, Department of Computer Science,Veer Narmad South Gujarat University, Surat, Gujarat, India.

[2] Professor, Department of ICT, Veer Narmad South Gujarat University, Surat, Gujarat, India.

[1] priyanka_pariyawala@yahoo.com, [2] pydesai@vnsgu.ac.in

ABSTRACT:

Polycystic ovary syndrome (PCOS) affects women of childbearing age. PCOS is a condition where women have irregular menses, obesity, hyperandrogenism, type 2 diabetes mellitus, cardiovascular diseases, etc. Awareness of PCOS has been one of the most important concerns for female fertility as PCOS patient has different risks associated with health, among the above-listed risks, the major risk is infertility. In this study, we developed a Model for predicting PCOS among adolescent girls. The questionnaire was prepared by considering different symptoms related to PCOS and applying machine-learning techniques to detect and raise awareness of PCOS. A total of 21 features were available then feature selection techniques 14 best features among them were selected, we used, SMOTE(Synthetic Minority Oversampling Techniques) with under sampling to solve the problem of class imbalance, and then traditional and Ensemble ML models were created, we compared output from both conventional and ensemble models, and select the best model among them. We split the dataset 80:20, and results showed that the ensemble technique did better than a traditional method, where Extra Trees had 91.70% and BRF achieved an accuracy of 90.30%.

Keywords: Machine Learning, Polycystic ovarian syndrome (PCOS, feature selection, disease prediction.

I. INTRODUCTION

Polycystic ovarian syndrome is one of the common endocrinopathy disorders during women's reproductive phase, which leads to infertility [1]. Ovaries play an important part in the reproductive system; Women have 2 ovaries that grow eggs and secrete estrogen and progesterone hormones. The ovaries make the egg that is released each month as part of a healthy menstrual cycle. If women have PCOS, the egg may not develop as it should, or it may not be released during ovulation as it should be [2]. PCOS was first reported by Stein and Leventhal in 1935.[42]

In the Medical field, bulk of data is generated regularly, which consists of all the information related to patients, So the paper is part of research where the author follows questions when doctors examine patients through physical examination and questions. In this Model prediction system, all the parameters that are important for predicting the disease are included, so it is possible to predict the disease efficiently and more accurately. Our research provides model descriptions to confirm efficiency, effectiveness, and trust in the developed model. The model works on traditional ensemble model techniques. Feature selection techniques are compared with different ML models, and proposed models to forecast PCOS in adolescent girls.

This paper's primary contributions are as follows.

1. Developed a simple, automated, accurate, and reliable machine-learning model for predicting PCOS in adolescent Girls
2. Applying feature selection (FS) to reduce data dimensionality and select the optimal feature, and applying SMOTE for class imbalance,
3. We build a traditional model and an ensemble model to predict PCOS and select the best model out of them based on accuracy, precision (P), recall (R), F1 score (F1), and area under the receiver operating characteristic (ROC) (AUC) curve.

This paper is divided into five sections: Section 2 describes existing research and review, Section 3 and Section 4 methodology and results compared with traditional and ensemble techniques, and Section 5 Conclusion.

2 Background Study

PCOS leads to many health problems, including anovulation, type 2 diabetes, cardiovascular problems, hypertension, insulin resistance, and so on[10-12].The past studies related to predicting PCOS are analyzed in this section. Our major reviewed papers from 2018 to 2024, 180 papers were reviewed, and 30 papers for this study were finalized. PCOS condition is most challenging for young women as it directly affects infertility at a later stage [3]. Polycystic ovary syndrome or PCOS is an illness that impacts women of childbearing age and is characterized by several features[23,24]. Anyone suffering from the condition needs to go through an appropriate diagnosis as the first step to receiving the right therapy for PCOS.

[4] described symptoms, causes, and treatment for PCOS. She has to describe various data mining techniques such as fuzzy logic, Machine learning techniques, statistics, and Rough set techniques. The Author also stated that ANN will be used for future work. Emir et al [5] developed a predictive model from six-week observation data to predict the response to pregabalin for the treatment of neuropathic pain. The training dataset had information about 9,187 patients and the testing dataset contained 6,114 patient information. They used 8 predictors to analyze resu5k, 54lts. The prediction suggested that adhering to a pregabalin medication regimen is essential for an ideal end-of-treatment result.

[13]compared different models, ANN, CNN, SVM, DT, and KNN also applied feature selection methods to diagnose PCOS. They get the highest accuracy from RF. [14] used Gini importance to select features. They applied different ML models: KNN, DT, SVM, LR, and NB, to detect PCOS. Based on the accuracy, DT recorded the highest rate. For feature selection, Agrawal et al. Author [15] has applied Chi-Square, ANOVA, and Mutual Information to identify insignificant features from the data. Author used selected features to detect PCOS by applying SVM, LR, DT, NB, XGBRF, RF, and CatBoost. The CatBoost classifier performed with the best accuracy.To determine whether a woman had PCOS, they used SVM, LR, NB, and KNN

in [21]. The top 30 characteristics were chosen using chi-square feature selection techniques. The maximum rate of accuracy has been attained by RF.

[22] carried out a study to improve to understand which clinical factors play the biggest role in determining a woman's risk of developing Polycystic Ovary Syndrome (PCOS). They used a public dataset available from Kaggle, which included medical details from 541 women and covered 44 different clinical features, along with information on whether or not each individual was diagnosed with PCOS. The data came from ten hospitals across Kerala, giving the study a diverse and well-rounded sample where a hybrid model combining a Multilayer Perceptron (MLP) and a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, which achieved an impressive 93% accuracy in predicting PCOS.

[24] suggested that a machine learning model based on late fusion that achieved 96% accuracy in identifying β-Thalassemia carriers on a dataset comprising 5066 cases.

3. Proposed Methodology and Implementation:

The proposed paper predicts PCOS. The Prediction is based on important parameters affected by PCOS. We build different traditional and ensemble ML models that combine with feature selection techniques to get good results.

Dataset: The author has collected questionnaire data from adolescent girls from different colleges in Surat City. The study's sample size is 1086. Among them, 748 are Negative PCOS, and 335 are positive PCOS.21 attributes were present in the dataset. A questionnaire was prepared with the help of a team of gynecologists. Once the data was collected, 50% was classified and validated by a gynecologist; the rest of the dataset was classified by the semi-supervised machine learning for annotation, and the doctor again validated the entire dataset. An Exploratory Data Analysis EDA was performed with Python programming, where anomalies, outliers, and missing values were removed. After EDA, data columns that have categorical features were converted to binary. Python supports great libraries, visualization, and plotting & framework for machine learning, so we chose Python. Table 1 shows the answer summary of different girls.

Table 1: Respondent's Answers Summary

Features		Percentage
Age (Years old)	17-19	72%
	20-22	25%
	23-25	3%
Do you have periods regularly?	YES	85%
	No	15%
Periods Flow	2 Days	3%
	3 Days	19%
	4 Days	28%
	More than 4 Days-6 Days	50%
Do you have pain in your periods	Yes	80%
	No	20%
	Yes	9%

Do you take medicine during periods?	No	91%
Do you use tempon?	Yes	6%
	No	94%
Are you suffering from weight gain?	Yes	29%
	No	71%
Are you suffering from Skin Darkning issues?	Yes	45%
	No	55%
Are you suffering from Hair Loss issues?	Yes	69%
	No	31%
Do you exercise regularly?	Yes	14%
	No	86%
Do you have pimples?	Yes	64%
	No	36%
Do you have any Outside food cravings?	Yes	70%
	No	30%
Do you have Sleep Disorder?	Yes	30%
	No	70%
Do you have any family history of Diabetes?	Yes	26%
	No	73%

```
Procedure to convert the object to an integer
def g(s):
    if(s=="YES"):
        return 1
    if(s=="NO"):
        return 0
    if s=="Yes":
        return 1
    if s=="No":
        return 0
    #convert object to integer
newdf['First_Period'] = newdf['First_Period'].astype('category')s
newdf["Periods_Regular"]=newdf.Periods_Regular.apply(g)
newdf["PainFul_Periods"]=newdf.PainFul_Periods.apply(g)
newdf["Medicine_Periods"]=newdf.Medicine_Periods.apply(g)
newdf["Use_Tempon"]=newdf.Use_Tempon.apply(g)
newdf["Weight_Gain"]=newdf.Weight_Gain.apply(g)
newdf["Skin_Darkning"]=newdf.Skin_Darkning.apply(g)
newdf["Exercise"]=newdf.Exercise.apply(g)
newdf["Pimples"]=newdf.Pimples.apply(g)
newdf["Food_Cravings"]=newdf.Food_Cravings.apply(g)
newdf["Know_PCOS"]=newdf.Know_PCOS.apply(g)
newdf["Sleep_Disorder"]=newdf.Sleep_Disorder.apply(g)
newdf["Hair_Loss"]=newdf.Hair_Loss.apply(g)
```

Figure 1: Code for Conversion in Python(Source: Self)

We observed that dataset is highly imbalanced, so we have combined the Synthetic Minority Oversampling Technique (SMOTE) + under-sampling =hybrid sampling method to resolve the class imbalance problem.

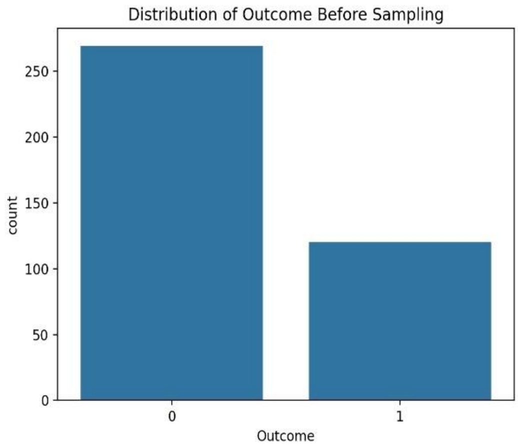


Figure 2: Before applying Smote (Source: Self)

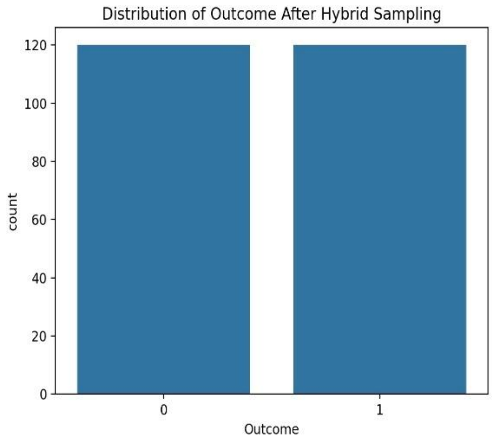


Figure 3: After applying Smote (Source: Self)

Table 2: Importance Scores of **PCOS** Features

Priority	Attributes	Importance
1	Height	0.093115
2	BMI	0.141593
3	Weight	0.172187
4	Do you have periods regularly?	0.055553
5	Age	0.056754
6	Do you have pain in your periods?	0.021521
7	Do you take medicine during periods?	0.041638
8	Do you use tampon?	0.023451
9	Are you suffering from weight gain?	0.035065
10	Are you suffering from Skin darkening issues?	0.047469
11	Are you suffering from Hair Loss issues?	0.028060
12	Do you exercise regularly?	0.027356
13	Do you have pimples?	0.033998
14	Do you have any Outside food cravings?	0.033794
15	Do you know about PCOS?	0.033449
16	Do you have Sleep Disorder?	0.046100
17	Do you have any family history of Diabetes?	0.059403
18	Have you ever visited a gynecologist?	0.032520

In **Table 2** Importance of features is collected with feature importance = (rf_importance + et_importance + xgb_importance) / 3, then data frame has been created for ranking.

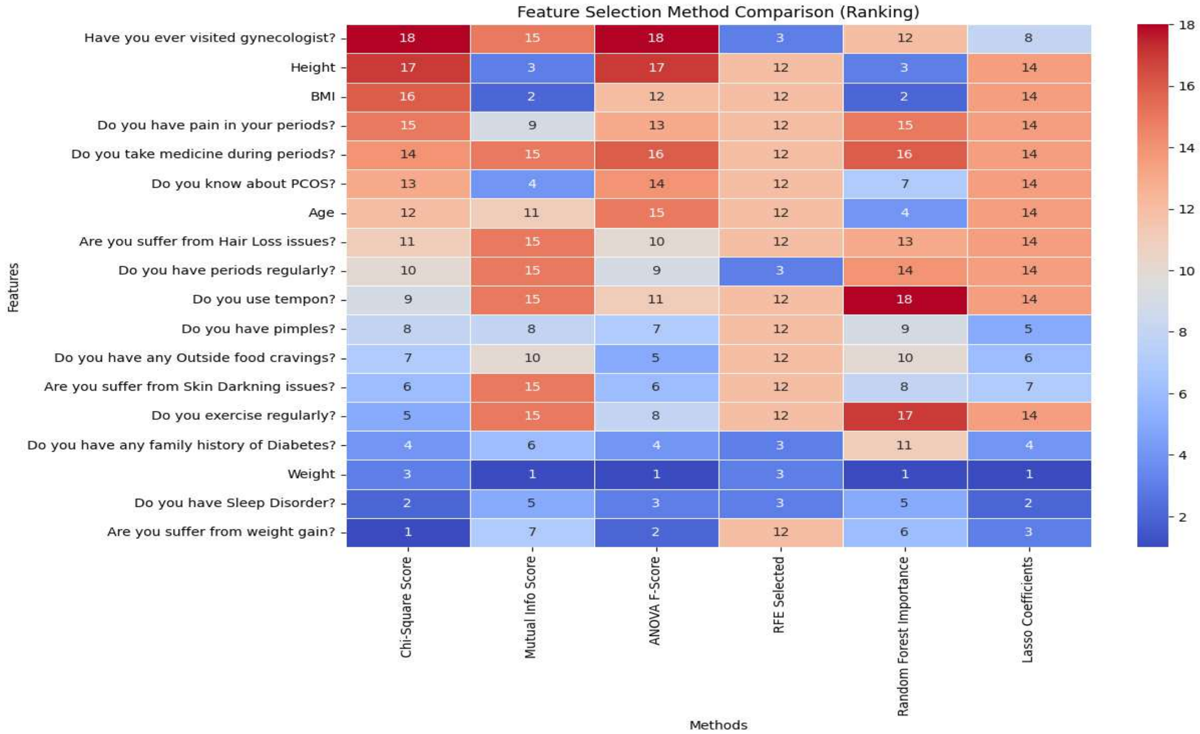


Figure 4. Shows the Visualization of feature importance rank Chi-Square, ANOVA, Mutual Info, RFE, Random Forest, Lasso. (Source: Self)

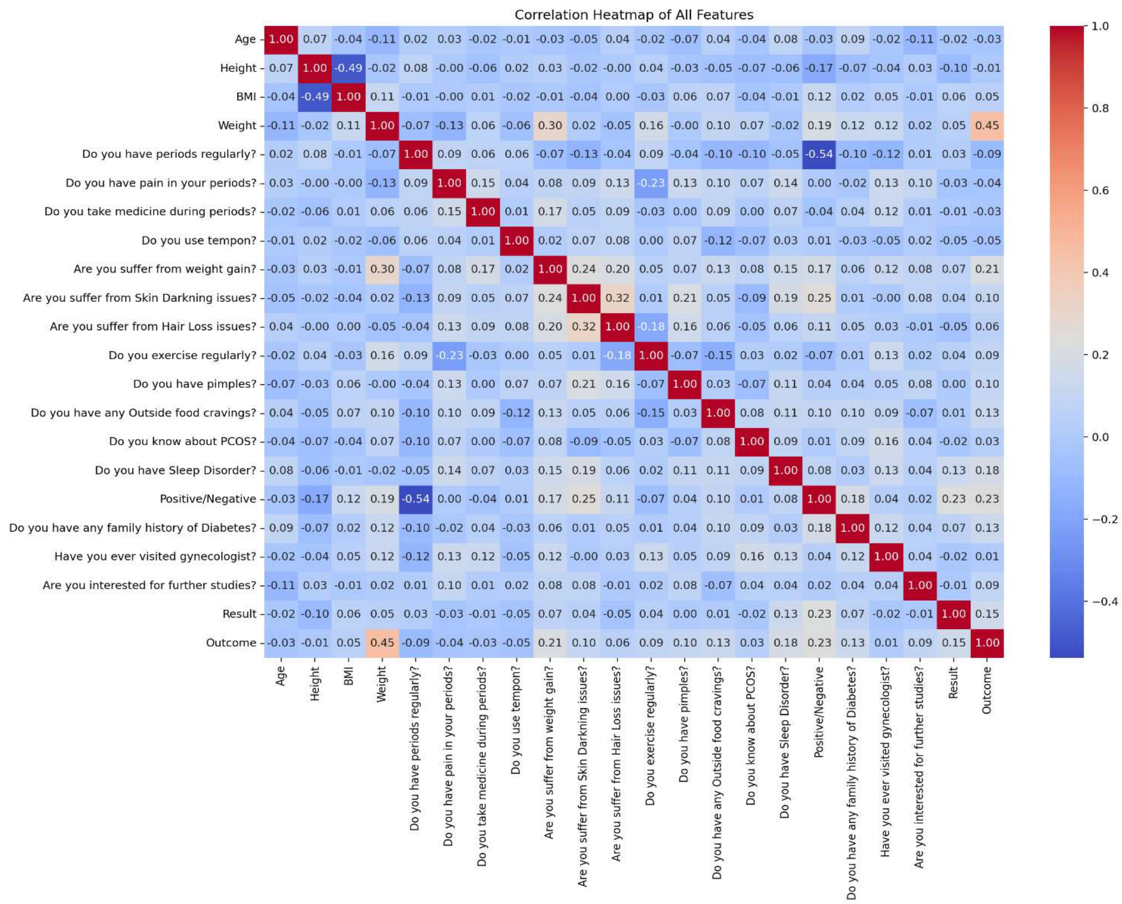


Figure 5. Feature correlation Heatmap(Source: Self)

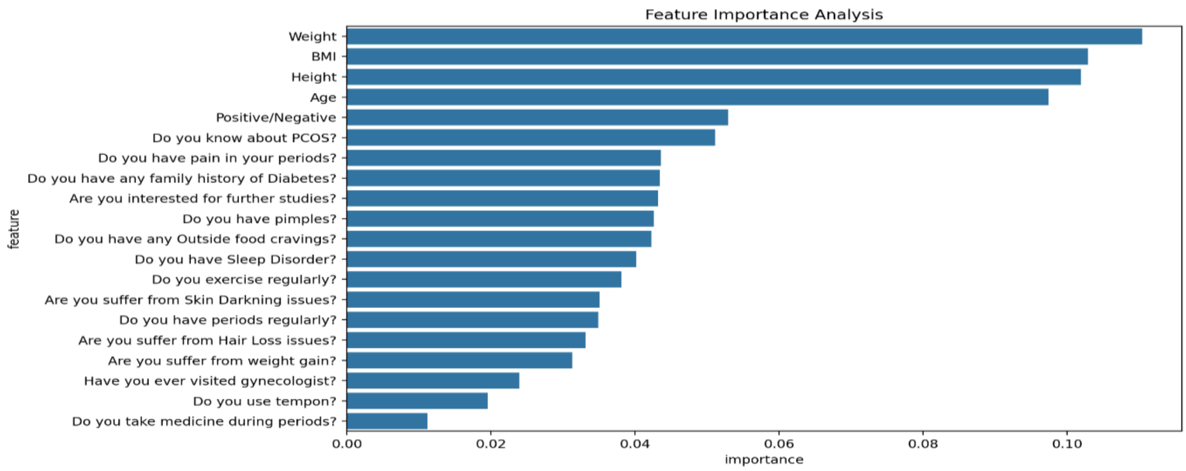
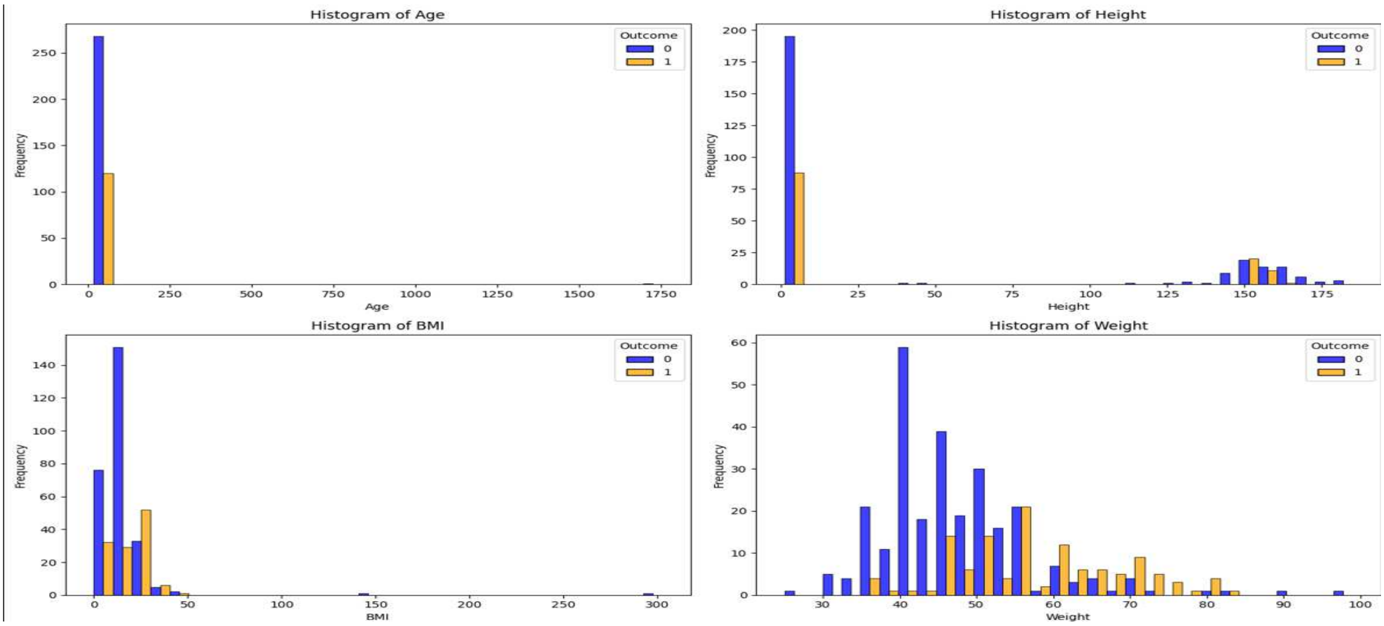
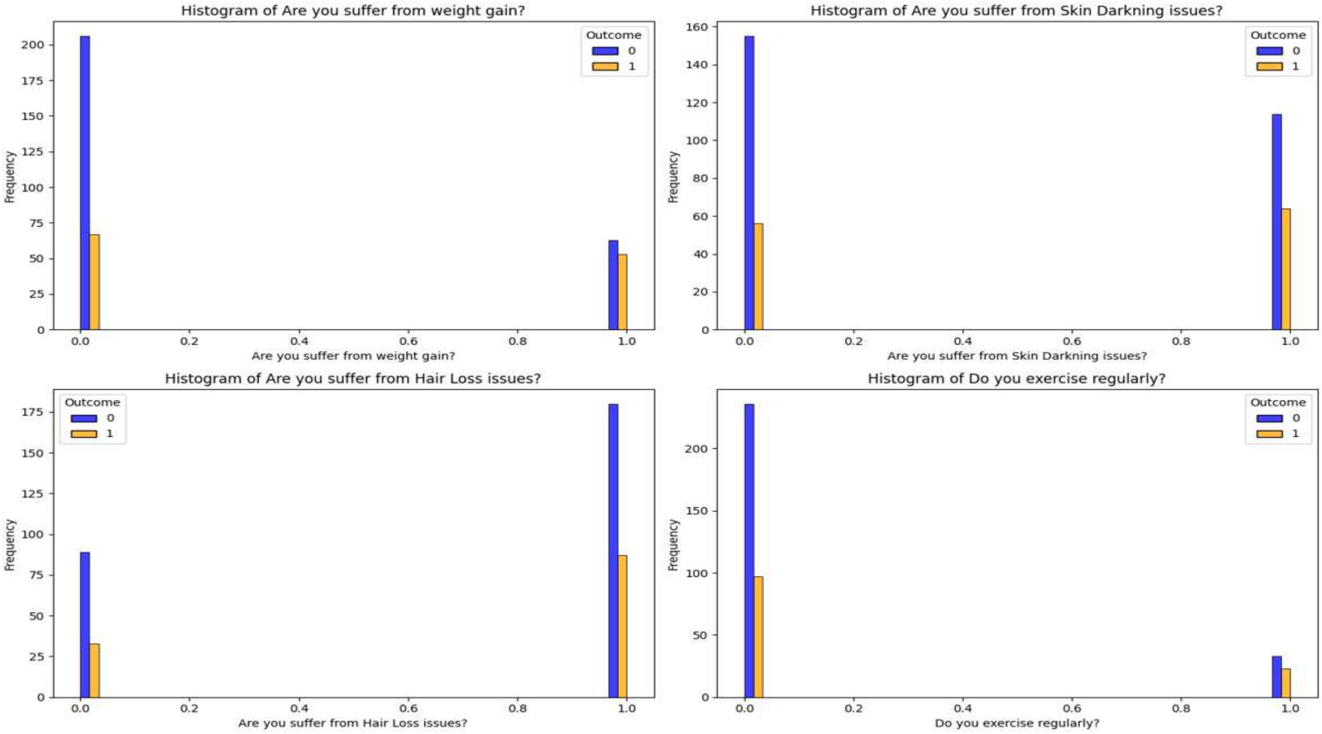
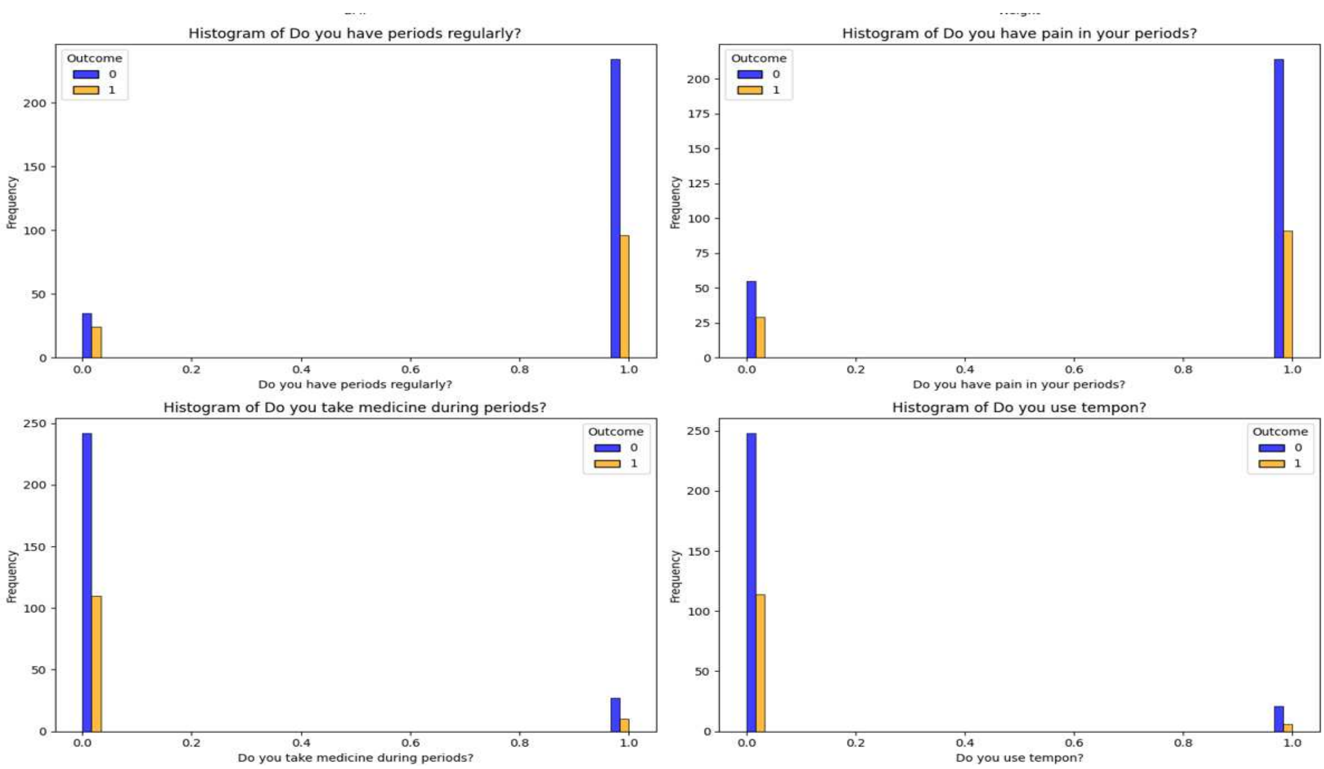


Figure 6. Feature importance (Source: Self)

In Figure 5 Heatmaps use color to depict the strength of correlations, with red representing a positive correlation and blue indicating a negative correlation. The color's intensity reflects the correlation's strength: colors closer to red or blue denote stronger correlations [25], while white signifies no correlation. Weight (0.45), BMI (0.30), Do you have periods regularly? All have a significant good relationship. Weak or No Correlation is Are you interested in further studies? Have you visited a gynecologist? Has zero correlation with outcome.




Histogram of Are you suffer from weight gain?
Are you suffer from weight gain?
Histogram of Are you suffer from Skin Darkning issues?
Are you suffer from Skin Darkning issues?
Histogram of Are you suffer from Hair Loss issues?
Are you suffer from Hair Loss issues?
Histogram of Do you exercise regularly?
Do you exercise regularly?
Frequency
Outcome
0
1


Histogram of Do you have periods regularly?
Do you have periods regularly?
Histogram of Do you have pain in your periods?
Do you have pain in your periods?
Histogram of Do you take medicine during periods?
Do you take medicine during periods?
Histogram of Do you use tempon?
Do you use tempon?
Frequency
Outcome
0
1

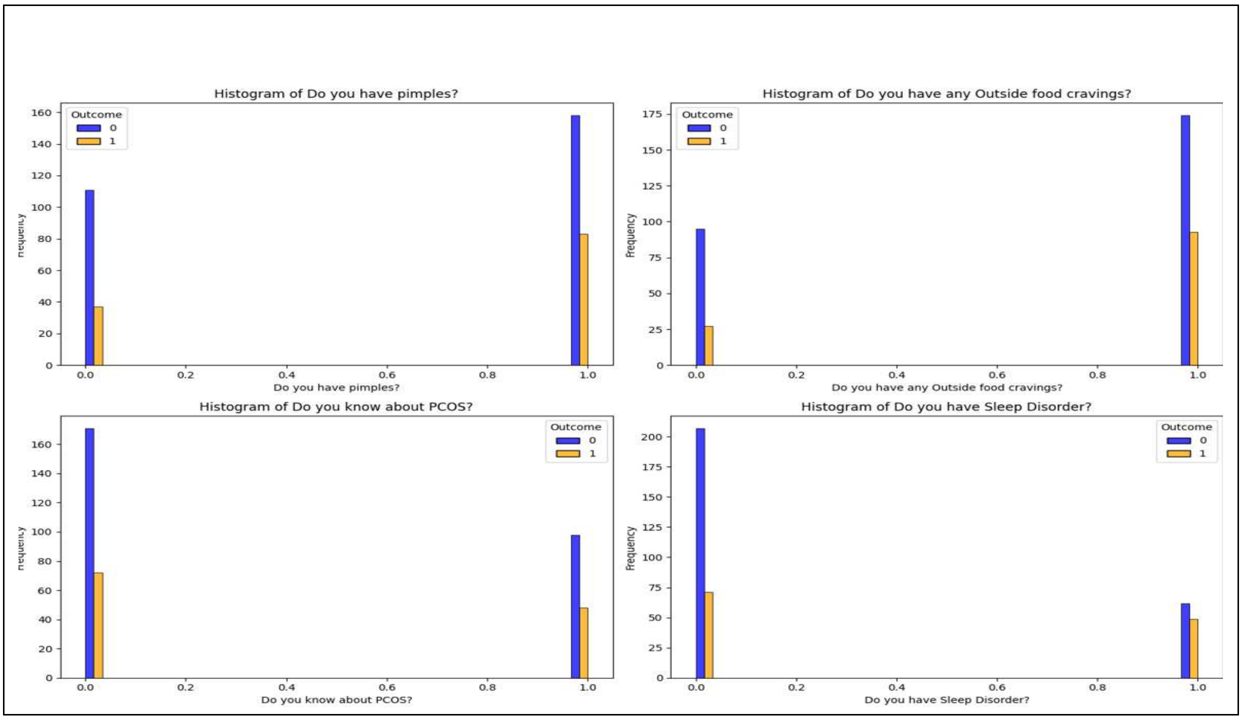


Figure 7 Data Distribution of all features (Source: Self)

In this paper, we have compared the traditional ML model with the Ensemble ML Model.

Algorithm 1: Working Procedure of PCOS Prediction

Input: Collected Dataset from Different colleges

Output: Outcome Predicted as PCOS Positive/Negative

1. Begin
2. data ← Load dataset
3. If the dataset value is empty or nan apply data imputation
4. Replace it with mean imputation
5. Data pre-processing
6. If data.dtypes is equal to object /string
7. Perform encoding of data
8. x ← data.drop[target=Outcome]
9. y ← data.target
10. xtrain,ytrain-80% of dataset
11. xtrain,ytrain-20% of dataset
12. Ensemble Learning, Traditional Learning methods and Algorithms Models=[All]
13. Model ← applies different classifiers
14. Predict ← find out the best accuracy
15. Compute performance matrix
16. End

Model	Accuracy in %	Precision	Recall	F1 Score	AUC-ROC	Observations
Logistic Regression	76.9%	75.0%	37.5%	50.0%	84.1%	Best AUC-ROC, but low recall So misses many PCOS cases.
Random Forest	79.5%	72.2%	54.2%	61.9%	83.6%	Balanced performance, robust choice.
SVM	79.5%	83.3%	41.7%	55.6%	79.9%	High precision but low recall So it misses many positives.
KNN	74.4%	61.1%	45.8%	52.4%	71.9%	Weakest model with lowest AUC-ROC.
Decision Tree	74.4%	58.3%	58.3%	58.3%	77.3%	Simple but out performed by Random Forest.
Naïve Bayes	74.4%	70.0%	29.2%	41.2%	76.5%	Poor recall; fails to detect many PCOS cases.
Tuned Random Forest	78.2%	70.6%	50.0%	58.5%	82.3%	Slight improvement over

Table 3. Traditional ML Model Performance(Source: Self)

In Table 3 shows comparison of Traditional Model, here We can see that Logistic Regression 84.1% has Highest AUC-ROC, which distinguishing best between PCOS and non-PCOS girls. Random Forest has high accuracy and good balance between precision and recalls has lower recall but also avoided falls positive. Naïve Bayes has lowest recall 29.2% so it fails to predict PCOS girls, whereas Weakest model is KNN Lowest AUC-ROC: 71.9% shows poor classification ability.

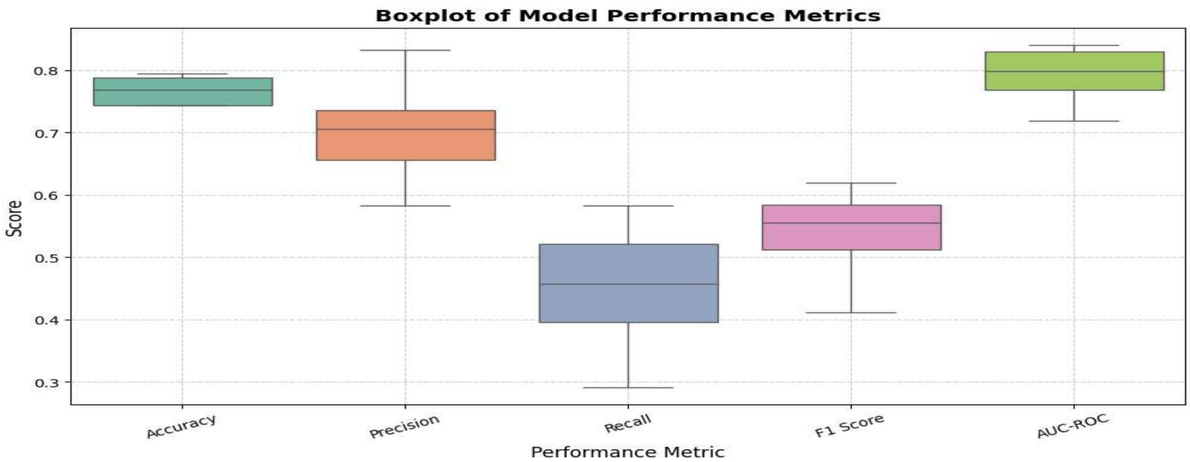


Figure 8 Boxplot Traditional Model Performance (Source: Self)

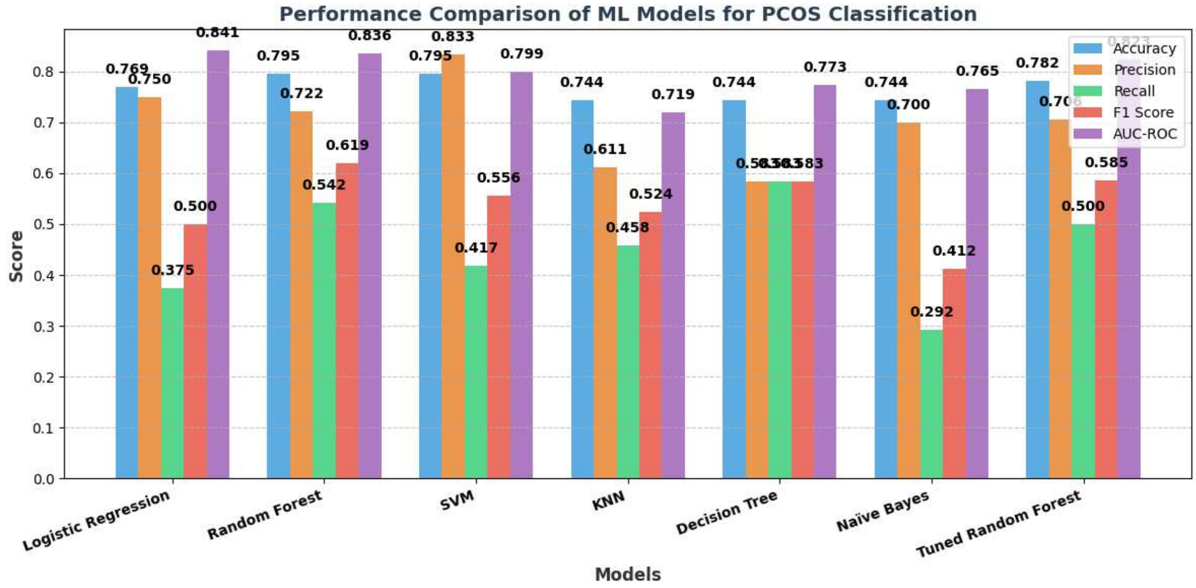


Figure 9. The graph to show Performance(Source: Self)

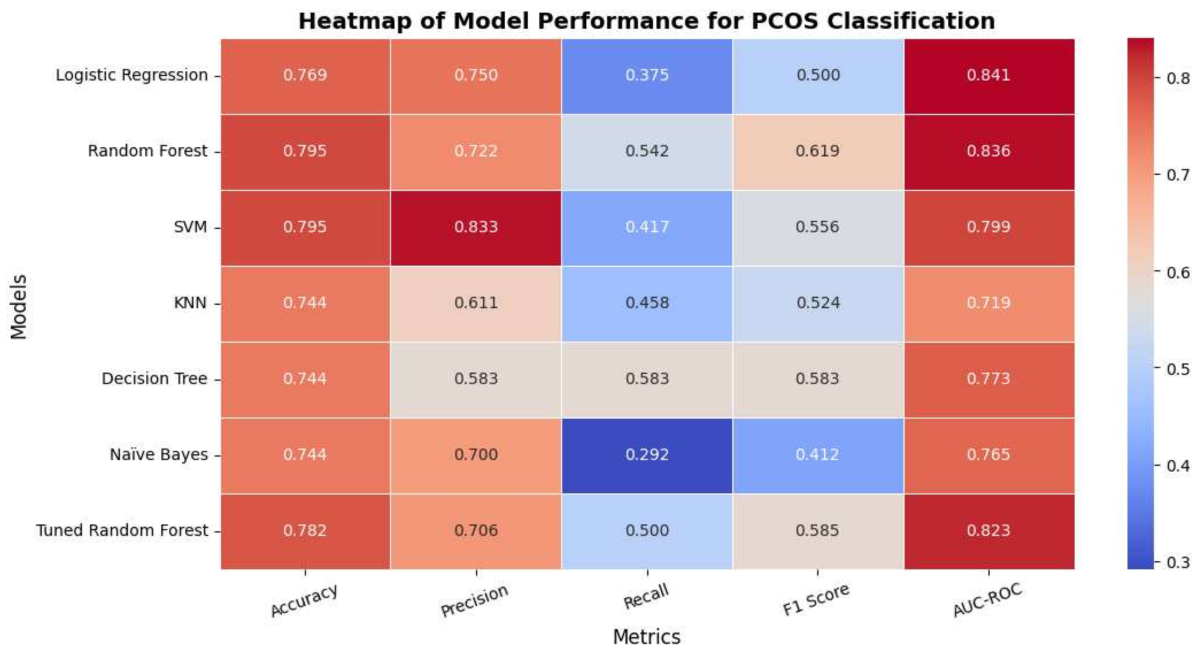


Figure 10. Heatmap of Traditional Model Performance(Source: Self)

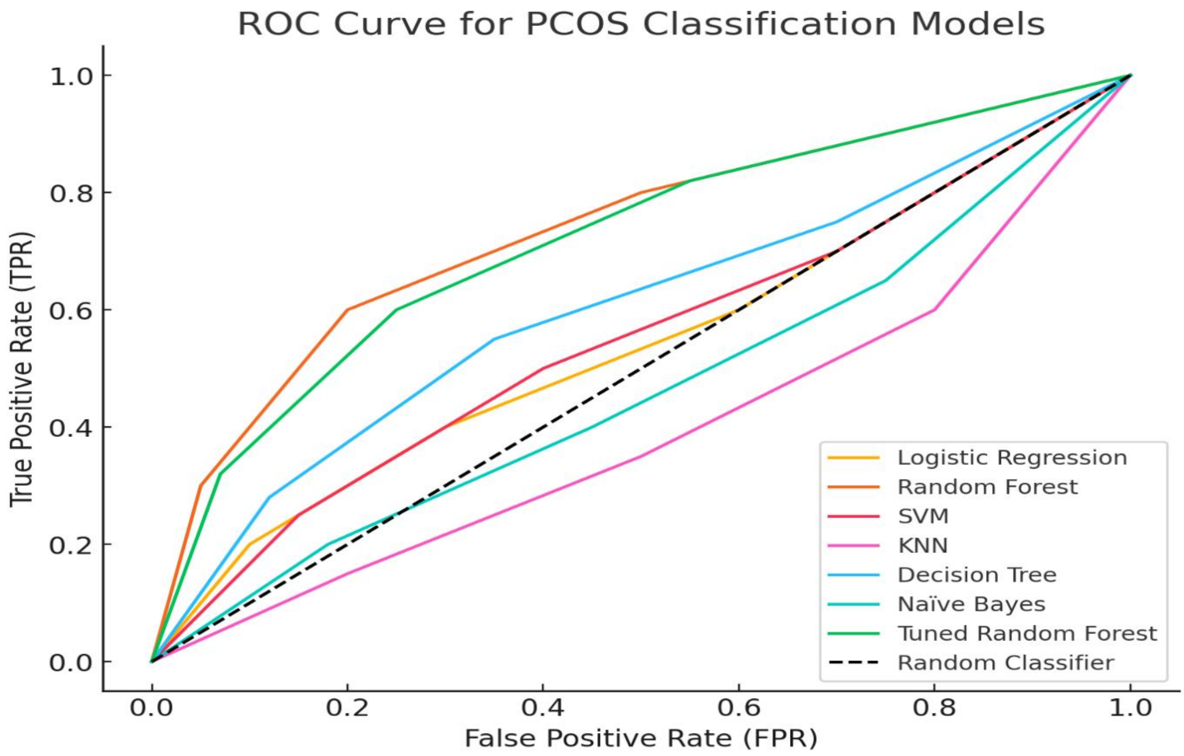


Figure 11 ROC Curve of Traditional Model (Source: Self)

Model	Hyperparamter	Default Values
Logistic Regression	max_iter, solver, C, penalty, tol, multi_class	max_iter=1000, solver='lbfgs', C=1.0, penalty='l2'
Random Forest	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	n_estimators=200, max_depth=10, min_samples_split=2
SVM	C, kernel, gamma, degree, probability	C=1.0, kernel='rbf', probability=True
KNN	n_neighbors, weights, algorithm, p, metric	n_neighbors=3, weights='uniform', algorithm='auto'
Decision Tree	max_depth, min_samples_split, min_samples_leaf, criterion, splitter	max_depth=5, criterion='gini', splitter='best'
Naive Bayes	var_smoothing	var_smoothing=1e-9
Tuned Random Forest	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	n_estimators=best (from GridSearchCV), max_depth=best

Table 4: Parameters and Default Values Used in Traditional Model (Source: Self)

In **Table 5**, Ensemble Model Developed and Compared, Metrix shows that Balanced RF leads with an accuracy of 90.28% and AUC-ROC with 96.97% and excellent recall of 95.12%, whereas Gradient Boosting has good precision and recall. Extra Trees has an amazing recall of 97.2% and 91.67% accuracy, and AdaBoost and Bernoulli NB will not perform well. So, the best model proved here was the Extra Tree.

Model	Accuracy in %	Precision	Recall	F1 Score	AUC-ROC	Key Observation
AdaBoost[43]	79.17%	79.55%	85.37%	82.35%	86.51%	**Good recall**, but slightly lower performance overall.
Balanced RF[44,45]	90.28%	88.64%	95.12%	91.76%	96.97%	**Best overall performance** across all metrics.
Gradient Boosting[46]	86.11%	86.05%	90.24%	88.1%	90.79%	**Well-rounded**, with good precision and recall.
Extra Trees[47]	91.67%	88.89%	97.56%	93.02%	97.17%	**Excellent recall** and F1 score, strong performance.
MLP[48]	86.11%	84.44%	92.68%	88.37%	87.1%	Decent performance but less **recall** than Extra Trees.
Bernoulli NB[49]	68.06%	75%	65.85%	70.13%	74.51%	**Least effective** across all metrics, low recall.

Table 5. Ensemble Model Performance (Source: Self)

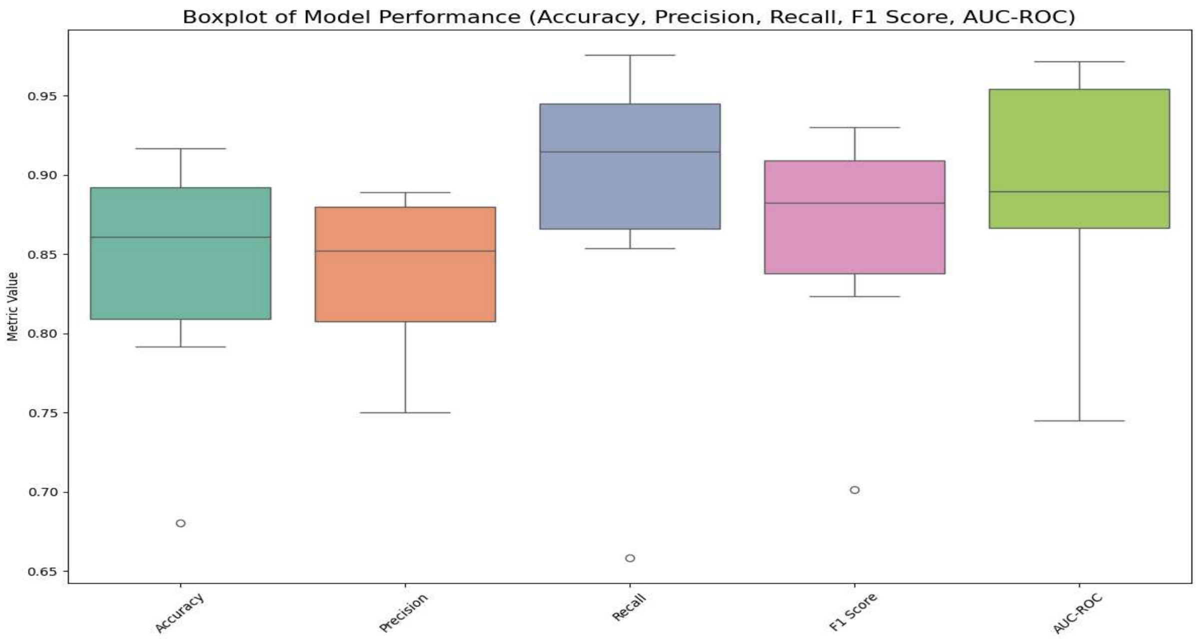


Figure 12. Box Plot of Ensemble Model Performance(Source: Self)

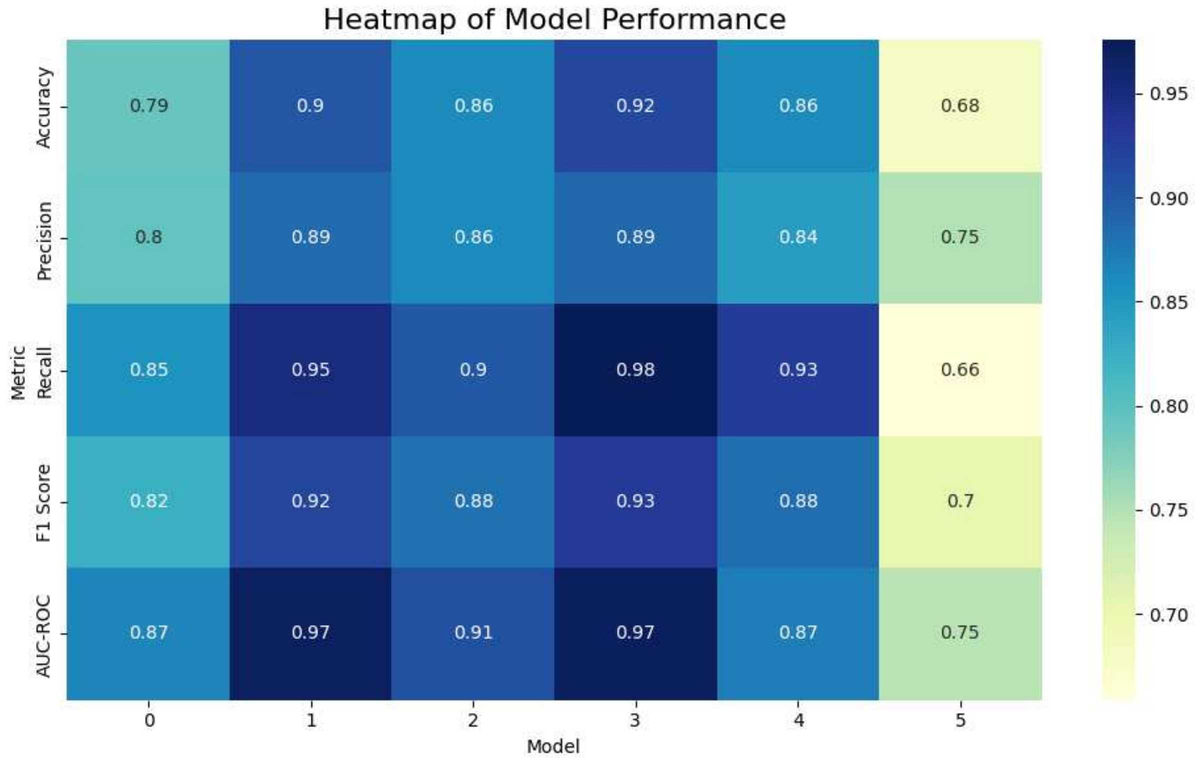


Figure 13 Heatmap of Ensemble Model Performance (Source: Self)

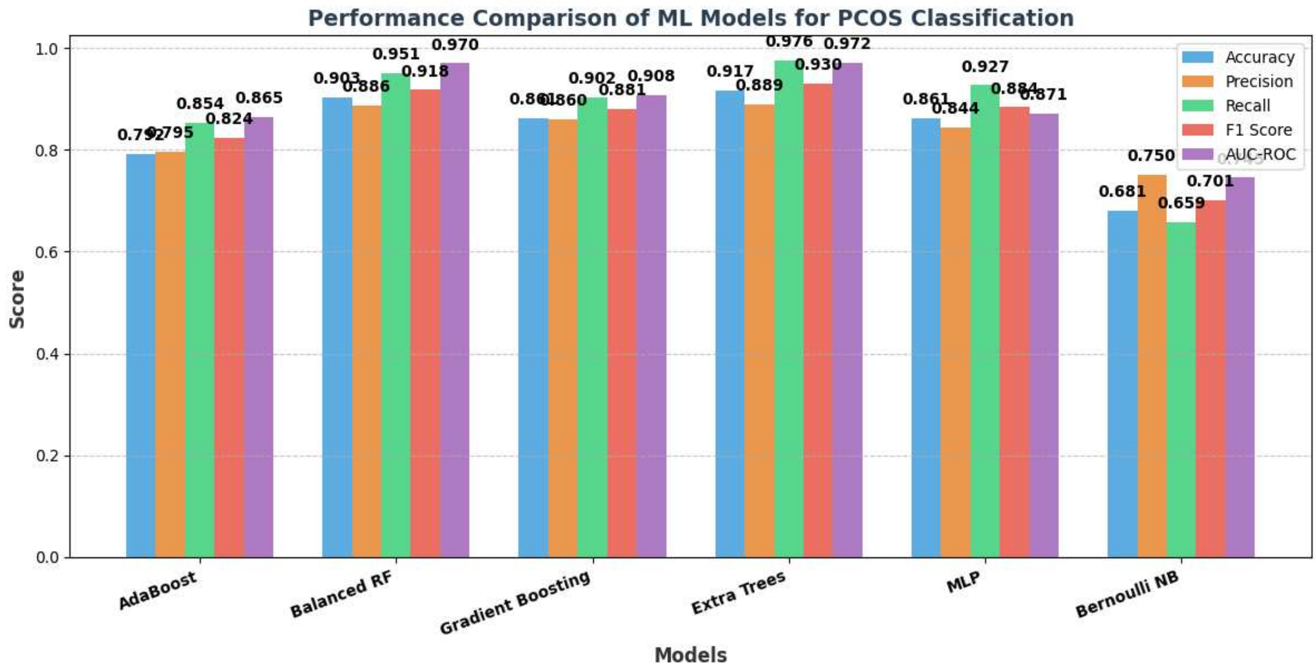


Figure 14 Graph Traditional Model VS Ensemble Model (Source: Self)

Model	Hyperparameters	Default Values Settings
AdaBoost	n_estimators, learning_rate, algorithm	n_estimators=100, learning_rate=1.0, algorithm='SAMME.R'
Balanced RF	n_estimators, max_depth, min_samples_split, min_samples_leaf, class_weight	n_estimators=100, max_depth=None, min_samples_split=2, class_weight='balanced'
Gradient Boosting	n_estimators, learning_rate, max_depth, subsample, min_samples_split	n_estimators=100, learning_rate=0.1, max_depth=3
Extra Trees	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	n_estimators=100, max_depth=None, max_features='auto'
MLP (Multi-Layer Perceptron)	hidden_layer_sizes, activation, solver, alpha, learning_rate_init	hidden_layer_sizes=(50, 50), activation='relu', solver='adam', max_iter=500
Bernoulli Naive Bayes	binarize, fit_prior, class_prior	binarize=0.0, fit_prior=True

Table 6: Parameters and Default Values Used in Ensemble Model (Source: Self)

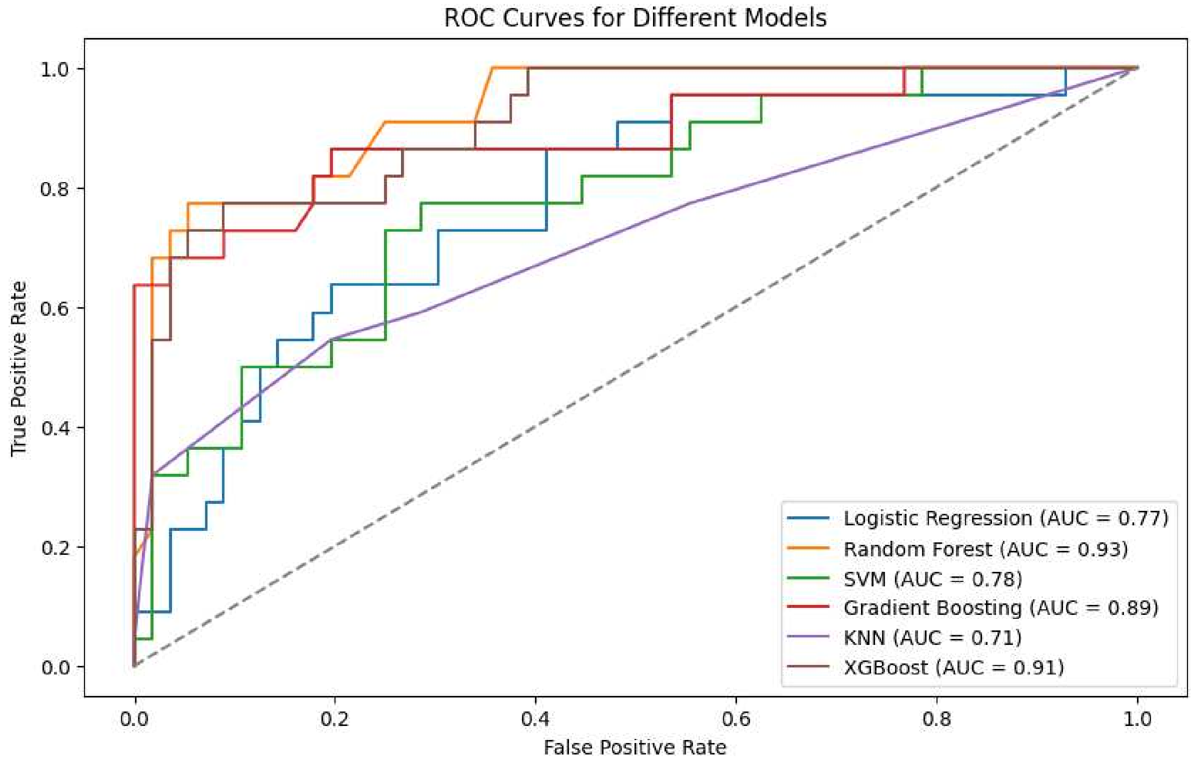


Figure 15 ROC Curve of Ensemble Model (Source: Self)

Table 6 lists models AdaBoost, Balanced Random Forest, Gradient Boosting, Extra Trees, Multi-Layer Perceptron (MLP), and Bernoulli Naive Bayes with specific settings that impact their performance. The parameters control features like the number of estimators, learning rates, depth, feature selection, activation functions, and optimization strategies. Understanding and tuning these hyperparameters is crucial for optimizing model performance, particularly in tasks like PCOS classification.

Model	Type	Details
AdaBoost[28]	Boosting	Sequentially adds weak learners to reduce errors.
Balanced Random Forest (BRF) [30]	Bagging + Sampling	Modified Random Forest with class balancing during sampling.
Gradient Boosting[28]	Boosting	Builds models sequentially; excellent for structured data tasks.
Extra Trees[32]	Randomized Tree Ensemble	Similar to Random Forest but with more randomization for splits.
Multilayer Perceptron (MLP)[22,31]	Neural Network	Deep learning model, not based on ensemble techniques.

Bernoulli Naïve Bayes[31]	Probabilistic (Generative Model)	Assumes feature independence; not an ensemble method.

Table 7: Type of Ensemble Model Followed (Source: Self)

Model	Category	Notes
Logistic Regression[33,34]	Linear Model	Widely used for binary classification; interpretable and efficient.
Random Forest[35]	Ensemble (Bagging)	Traditional ensemble method using multiple decision trees.
SVM (Support Vector Machine)[36]	Margin-based Classifier	Effective for high-dimensional data; can use kernels (e.g., RBF).
KNN (K-Nearest Neighbors)[37]	Instance-Based Learning	Simple, non-parametric; classification based on nearest neighbors.
Decision Tree[38]	Tree-Based	Simple and interpretable; it forms the base for many ensemble methods.
Naïve Bayes[39,40]	Probabilistic Model	Based on Bayes theorem; fast, assumes feature independence.
Tuned Random Forest[41]	Ensemble (Optimized)	Same as Random Forest but with hyperparameters tuned for better performance.

Table 8: Category of traditional Model They Follow (Source: Self)

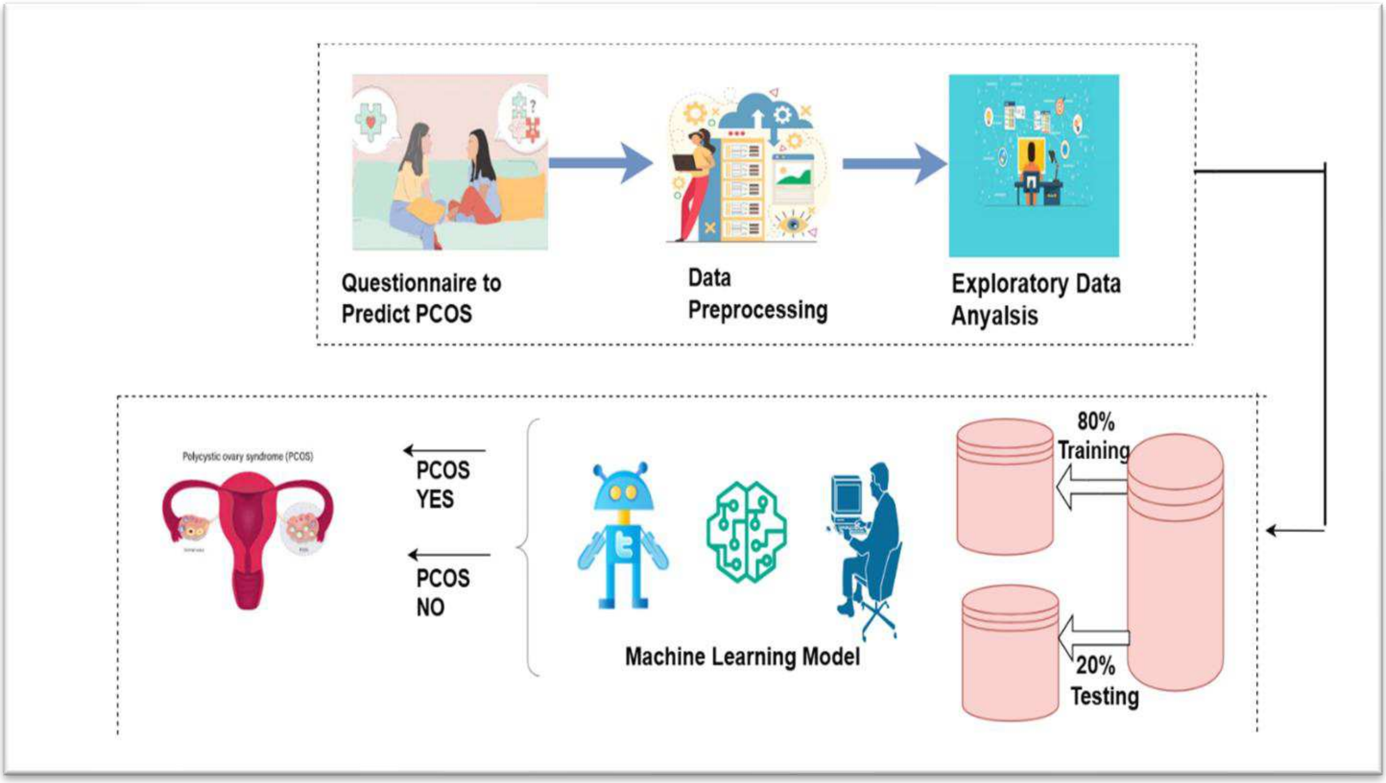


Figure 16: System Architecture of our Model (Source: Self)

Ethical Compliance

All methods were carried out in accordance with relevant guidelines and regulations. The data used in this study were collected through a structured questionnaire and anonymized prior to analysis. As the study involved secondary analysis of non-identifiable data, formal ethical approval was not required as per institutional policies.

CONCLUSION

Predicting the disease earlier can improve female health. Early detection of PCOS can reduce long-term complications. We use traditional ML models and ensemble models to predict PCOS. As a combined we use(LR, RF, SVM, KNN, DT,NB, AdaBoost, BRF, Gradient Boosting, Extra Trees, MLP). Among evaluated models, Extra Trees, Balanced Random Forest, Gradient Boosting, and Multi-Layer Perceptron (MLP) emerged as top-performing models, where Extra Trees has 91.70%, Extra Trees builds decisions using random splits and observed unseen data as well, and BRF has achieved an accuracy of 90.30%, BRF did an excellent balance between catching positive cases and avoiding false cases.

This Paper aims to predict diseases based on symptoms. The Model takes the patient's symptoms as input and gives a prediction based on that. This model can help reduce the cost of dealing with this disease and improve the recovery process & patients can reduce the money required for treatment and can save time. In the future, we can add more diseases to the existing prediction system. We can try to improve the accuracy of the disease prediction to reduce mortality also work on biochemicals and ultrasound images, which create primary data sources.

Author contributions

P.P. conducted the research, performed data analysis, and developed the machine learning models. P.D. provided overall supervision, conceptual guidance, and reviewed the methodology. P.P. wrote the main manuscript text and prepared the figures and tables. P.D. critically reviewed and edited the manuscript. Both authors read and approved the final version of the manuscript.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors

Data availability

The dataset available from Priyanka_pariyawala@yahoo.com is included in the paper.

Declarations

Corresponding Author:

Priyanka Pariyawala

Consent for publication

All authors have reviewed and approved the final manuscript and consent to its publication.

Conflict of interest

The authors have no conflicts of interest to declare for this study.

Clinical trial number

Not applicable.

Ethical approval and accordance

The study was approved by the Institutional Ethics Committee of SMT. Z. S. Patel College of Computer Application**.** Written informed consent was obtained from all participants prior to data collection. Participation was voluntary, and confidentiality of responses was maintained throughout the study.

References:

1. Trivax B, Azziz R. Diagnosis of polycystic ovary syndrome. Clin Obstet Gynecol. 2007 Mar;50(1):168-77. doi: 10.1097/GRF.0b013e31802f351b. PMID: 17304034.
2. American College of Obstetricians and Gynecologists' Committee on Practice Bulletins—. ACOG Practice Bulletin No. 194: Polycystic Ovary Syndrome. Obstet

Gynecol. 2018 Jun;131(6):e157-e171. doi: 10.1097/AOG.0000000000002656. Erratum in: Obstet Gynecol. 2020 Sep;136(3):638. PMID: 29794677.

3. D. Hu, W. Dong, X. Lu, H. Duan, K. He, and Z. Huang, "Evidential mace prediction of acute coronary syndrome using electronic health records," BMC Medical Informatics and Decision Making, vol. 19, no. 2, pp. 9–17, 2019.
4. Paper, R.P., Meena, D.K., Manimekalai, D.M., & Rethinavalli, S. (2015). A Literature Review on Polycystic Ovarian Syndrome and Data Mining Techniques.
5. Emir B, Johnson K, Kuhn M, Parsons B, Predictive Modeling of Response to Pregabalin for the Treatment of Neuropathic Pain Using 6-Week Observational Data: A Spectrum of Modern Analytics Applications, Clinical Therapeutics, Volume 39, Issue 1 , Pages 98106, January 2017.
6. M. M. Hassan and T. Mirza, "Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome", Int. J. Comput. Appl., vol. 175, pp. 42-53, Sep. 2020.
7. Zhang XZ, Pang YL, Wang X, Li YH. Computational characterization and identification of human polycystic ovary syndrome genes. Sci Rep. 2018 Aug 28;8(1):12949. doi: 10.1038/s41598-018-31110-4. PMID: 30154492; PMCID: PMC6113217.
8. S. A. Bhat, Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms. PhD thesis, Dublin, National College of Ireland, 2021
9. Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram C, Remya George. "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques" 2019 IEEE Region 10 Conference (TENCON 2019).
10. P. Anagnostis, B.C. Tarlatzis, R.P. Kauffman, Polycystic ovarian syndrome (PCOS): long-term metabolic consequences, Metabolism 86 (2018) 33–43.
11. A.S. Jamil, S.K. Alalaf, N.G. Al-Tawil, T. Al-Shawaf, A case–control observational study of insulin resistance and metabolic syndrome among the four phenotypes of polycystic ovary syndrome based on Rotterdam criteria, Reprod. Health 12 (2015) 1–9.
12. G. Iatrakis, C. Tsionis, G. Adonakis, M. Stoikidou, F. Anthouli-Anagnostopoulou, M. Parava, A. Vouxinou, N. Georgopoulos, G. Kourounis, Polycystic ovarian syndrome, insulin resistance and thickness of the endometrium, Eur. J. Obstet. Gynecol. Reprod. Biol. 127 (2006) 218–221
13. Anda, D.; Iyamah, E. Comparative Analysis of Artificial Intelligence in the Diagnosis pandas-profiling
14. https://www.researchgate.net/publication/366320486_Comparative_Analysis_of_Artificial_Intelligence_in_the_ Diagnosis_of_Polycystic_Ovary_Syndrome (accessed on 17 March 2023)
15. Chauhan, P.; Patil, P.; Rane, N.; Raundale, P.; Kanakia, H. Comparative analysis of machine learning algorithms for prediction of pcos. In Proceedings of the 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 25–27 June 2021; pp. 1–7.
16. Aggarwal, N.; Shukla, U.; Saxena, G.J.; Kumar, M.; Bafila, A.S.; Singh, S.; Pundir, A. An Improved Technique for Risk Prediction of Polycystic Ovary Syndrome (PCOS) Using Feature Selection and Machine Learning. In Computational Intelligence: Select Proceedings of InCITe 2022; Springer: New York, NY, USA, 2023; pp. 597–606.

17. Elmannai, Hela & El-Rashidy, Nora & Mashal, Ibrahim & Alohali, Manal & F Abd-el Ghany, Sara & El-Sappagh, Shaker & Saleh, Hager. (2023). Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence. Diagnostics. 13. 1506. 10.3390/diagnostics13081506.
18. Escobar-Morreale HF. Polycystic ovary syndrome: definition, aetiology, diagnosis and treatment. Nat Rev Endocrinol. 2018 May;14(5):270-284. doi: 10.1038/nrendo.2018.24. Epub 2018 Mar 23. PMID: 29569621.
19. Shamik Tiwari, Lalit Kane, Deepika Koundal, Anurag Jain, Adi Alhudhaif, Kemal Polat, Atef Zaguia, Fayadh Alenezi, Sara A. Althubiti, SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning, Expert Systems with Applications, Volume203,2022,117592,ISSN0957 4174,https://doi.org/10.1016/j.eswa.2022.117592.
20. Zad Z, Jiang VS, Wolf AT, Wang T, Cheng JJ, Paschalidis IC, Mahalingaiah S. Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records. medRxiv [Preprint]. 2023 Oct 1:2023.07.27.23293255. doi: 10.1101/2023.07.27.23293255. Update in: Front Endocrinol (Lausanne). 2024 Jan 30;15:1298628. doi: 10.3389/fendo.2024.1298628. PMID: 37577593; PMCID: PMC10418575.
21. Thakre, V.; Vedpathak, S.; Thakre, K.; Sonawani, S. PCOcare: PCOS detection and prediction using machine learning algorithms. Biosci. Biotechnol. Res. Commun. 2020, 13, 240–244.
22. Bhardwaj P, Tiwari P. Manoeuvre of machine learning algorithms in healthcare sector with application to polycystic ovarian syndrome diagnosis. In: Proceedings of academia-industry consortium for data science: aicds 2020. Springer; 2022. p. 71–84
23. Lin, X.; Li, C.; Zhang, Y.; Su, B.; Fan, M.; Wei, H. Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. Molecules 2017, 23, 52.
24. Gu, Q.; Li, Z.; Han, J. Generalized fisher score for feature selection. arXiv 2012, arXiv:1202.3725.
25. I.C. Covert, W. Qiu, M. Lu, N.Y. Kim, N.J. White, S.I. Lee, Learning to maximize mutual information for dynamic feature selection, in: International Conference on Machine Learning, P
26. Saleh, H.; Mostafa, S.; Alharbi, A.; El-Sappagh, S.; Alkhalifah, T. Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis. Sensors 2022, 22, 3707.
27. Muhammad Ibrahim, et al., Fuzzy-based fusion model forβ-thalassemia carriers prediction using machine learning technique, Advances in Fuzzy Systems 2024 (2024).
28. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K.; Mitchell, R.; Cano, I.; Zhou, T.; et al. Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2; 2015 ; Volume 1, pp. 1–4. Available online: https://scholar.google.com/scholar?oi= bibs&cluster=11444560539169478279&btnI=1&hl=en (accessed on 17 March 2023).
29. Wu, J.; Chen, X.Y.; Zhang, H.; Xiong, L.D.; Lei, H.; Deng, S.H. Hyperparameter optimization for machine learning models based on Bayesian optimization. J. Electron. Sci. Technol. 2019, 17, 26–40.
30. Rigatti, S.J. Random forest. J. Insur. Med. 2017, 47, 31–39.
31. Guo J, et al. Predicting and extracting thermal behavior rules of hydronic thermal barrier with interpretable ensemble learning in the heating season. Energy Build 2023;301:11369

32. Berrouachedi A, Jaziri R, Convolutional G Bernard. Extra-trees and multi-layer perceptron. In: 2022 IEEE/ACS 19th international conference on computer systems and applications (AICCSA). IEEE; 2022
33. Madushani JS, et al. Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers. Transport Eng 2023;1
34. Praveena R, et al. Prediction of rainfall analysis using logistic regression and support vector machine. In: Journal of physics: conference series. IOP Publishing; 2023.
35. James G, et al. Tree-based methods. In: An introduction to statistical learning: with applications in Python. Springer; 2023. p. 3
36. Cerulli G. Discriminant analysis, nearest neighbor, and support vector machine. In: Fundamentals of supervised machine learning: with applications in Python, R, and stata. Springer.
37. Tyagi A, Singh VP, Gore MM. An efficient automated detection of schizophrenia using k-NN and bag of words features. SN Computer Science 2023.
38. Xiangfei Meng, Pei Zhang, Yan Xu, Hua Xie, Construction of decision tree based on C4.5 algorithm for online voltage stability assessment,International Journal of Electrical Power & Energy Systems, Volume 118,2020,105793,ISSN 0142-0615,https://doi.org/10.1016/j.ijepes.2019.105793.(https://www.sciencedirect.com/science/article/pii/S0142061519324251)
39. Webb, G.I.; Keogh, E.; Miikkulainen, R. Naïve Bayes. Encycl. Mach. Learn. 2010, 15, 713–714
40. Rezaeian N, Novikova G. Persian text classification using naive bayes algorithms and support vector machine algorithms. Indonesian Journal of Electrical Engineering and Information
41. Hassan MM, Mirza T. Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. Int J Computer Appl 2020;975:8887.
42. Shan H, Luo R, Guo X, Li R, Ye Z, Peng T, Liu F, Yang Z. Abnormal Endometrial Receptivity and Oxidative Stress in Polycystic Ovary Syndrome. Front Pharmacol. 2022 Jul 25;13:904942. doi: 10.3389/fphar.2022.904942. PMID: 35959444; PMCID: PMC9357999.
43. Parnami, A., & Sahu, R. (2022). Heart Disease Prediction using Machine Learning Algorithms. International Journal of Intelligent Systems and Applications in Engineering.
44. José Denes Lima Araújo, Luana Batista da Cruz, João Otávio Bandeira Diniz, Jonnison Lima Ferreira, Aristófanes Corrêa Silva, Anselmo Cardoso de Paiva, Marcelo Gattass,Liver segmentation from computed tomography images using cascade deep learning,Computers in Biology and Medicine,Volume 140,2022,105095,ISSN 0010-4825,https://doi.org/10.1016/j.compbiomed.2021.105095.
45. Gupta, Ishu, et al. "PCA-RF: an efficient Parkinson's disease prediction model based on random forest classification." arXiv preprint arXiv:2203.11287 (2022).
46. Suryawanshi, A. et al. (2022). RFE + GBM for CVD Prediction. Computers in Biology and Medicine.
47. Talari P, N B, Kaur G, Alshahrani H, Al Reshan MS, Sulaiman A, Shaikh A. Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2. PLoS One. 2024 Jan 18;19(1):e0292100. doi: 10.1371/journal.pone.0292100. PMID: 38236900; PMCID: PMC10796060.

48. Yadav, A. K. ., Pal, G. K. ., & Gangwar, S. . (2024). Early Stage Prediction of Heart Disease Features using AdaBoost Ensemble Algorithm and Tree Algorithms. International Journal of Intelligent Systems and Applications in Engineering, 12(3), 545–551. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/5285.
49. López-Rubio, E. et al. (2022). Naïve Bayes for Mental Health Detection from Binary Features. Expert Systems with Applications.