SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V10I9P106

Volume 10 Issue 9, 61-75, September 2023
© 2023 Seventh Sense Research Group®

Original Article

Machine Learning-Driven PCOS Prediction for Early Detection and Tailored Interventions

B.Yamini[1], Venkata Ramana Kaneti[2], Prema.P[3], Ambhika C[4], M.Nalini[5], Siva Subramanian.R[6]

[1]*Department of Networking and Communications, School of Computing, College of Engineering and Technology, SRM Institute of Science and Technology (SRMIST), Chengalpattu, Tamilnadu, India.*
[2]*Department of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India.*
[3]*Department of Information Technology, Painmalar Engineering College, Chennai, Tamilnadu, India.*
[4]*Department of Information Technology, Velammal Institute of Technology, Thiruvallur, Tamilnadu, India.*
[5]*Department of Computer Science and Engineering, S.A.Engineering College, Tamilnadu, India.*
[6]*Department of Computer Science and Engineering, RMK College of Engineering and Technology, Tamilnadu, India.*

[6]*Corresponding Autor : sivasubramanian12@yahoo.com*

Received: 09 July 2023 Revised: 14 August 2023 Accepted: 07 September 2023 Published: 30 September 2023

Abstract - *This PCOS is a hormonal disorder that leads to the overproduction of androgens, resulting in symptoms such as interrupted periods, ovarian follicles, excess body hair, weight gain, and infertility. Hormonal imbalances and abnormal male hormone production characterize it. The precise aetiology of Polycystic Ovary Syndrome is unknown, but insulin resistance and genetic factors may play a role. Due to its prevalence and potential long-term health effects, understanding and predicting PCOS is most important in healthcare. The study of PCOS in women aids in identifying the condition earlier and protecting women from life-threatening medical complications. This research uses ML algorithms to develop a novel predictive modelling strategy to identify individuals at risk of developing PCOS. In the field of reproductive health, ML has the potential to revolutionize healthcare by enhancing detection and prediction. Multiple ML models, such as LR, RF, SVM, NB, K NN, and XGBoost, were used to predict PCOS. The examination uses a PCOS dataset containing clinical, hormonal, and biological information from women with and without PCOS issues. The acquired experimental results are projected using various validity metrics, including precision, recall, accuracy and F1-Score. The outcome indicates that Machine Learning models have promising predictive ability, and the random forest model has a 90% accuracy rate, which is higher than any other model. PCOS research is essential for encouraging early diagnosis, effective treatment, and improved reproductive health outcomes for those affected. Using Machine Learning algorithms, our proposed method provides a promising approach to PCOS prediction, enabling physicians to rapidly identify at-risk patients and perform tailored therapies. By treating PCOS early, healthcare practitioners may help women with this complicated endocrine illness avoid problems and enhance their quality of life*.

Keywords - *Machine Learning, Polycystic Ovary Syndrome, Prediction, Random forest, Women healthcare.*

1. Introduction

Polycystic Ovary Syndrome is referred to as PCOS. It is a hormonal condition that primarily affects females between the ages of 15 and 45 who are of reproductive age. One of the most prevalent hormonal conditions in women is PCOS, yet its specific aetiology is still unknown. However, it is thought to result from hereditary and environmental influences [1]. The main characteristics of PCOS include:

- Infrequent, protracted, or nonexistent periods are common in women with PCOS.
- Hyperandrogenism: This is the condition in which the female body has increased amounts of androgens (male hormones), which may cause symptoms including hirsutism (excessive hair growth on the face, chest, belly, or back), acne, and male-pattern baldness.
- Polycystic ovaries: Despite the term, not all PCOS patients have ovarian cysts [2].

The word "polycystic" describes how the ovaries look on ultrasound when there may be many tiny follicles present. Along with these fundamental characteristics, PCOS may also be linked to other health conditions, such as insulin resistance, which raises the risk of type 2 diabetes. Additionally, weight gain or difficulty decreasing weight may be problematic for women with PCOS. Not all women with PCOS will exhibit all of the symptoms above, and PCOS could appear in various ways in different people.

A medical history, physical examination, and blood tests to measure hormone levels are often used to make a diagnosis. PCOS may be controlled with dietary and activity modifications and medication that targets particular symptoms and hormonal imbalances [3]. Management is crucial since untreated PCOS may create long-term health issues and problems. See a healthcare expert for a diagnosis and proper treatment if you think you may have PCOS or are exhibiting symptoms. The study of PCOS prediction helps to decrease early medical complications and helps women's lives be better. One crucial reason to study PCOS analysis is that PCOS (Polycystic Ovary Syndrome) research is crucial for many reasons.

- Prevalence and effects on women's health: PCOS is thought to affect 5–10% of women globally. Its significant incidence significantly affects women's general well-being, quality of life, and health.
- Recognizing the root causes: Although the precise origin of PCOS is not yet entirely known, research has helped identify the underlying processes and aggravating elements. Improved diagnostic techniques and more focused therapies may result from a better understanding of the aetiology of PCOS.
- Reproductive health and fertility: PCOS may result in irregular menstrual cycles and anovulation (lack of ovulation), making it challenging to conceive [4]. Studying PCOS makes it easier to find fertility medications and interventions that work, improving the odds of becoming pregnant for those who need them.
- Metabolic and cardiovascular health: Insulin resistance, dyslipidemia, and type 2 diabetes are common metabolic problems in PCOS-affected women. Knowing these connections may help develop early diagnosis and prevention methods.
- Due to its physical symptoms (acne hirsutism) and the emotional stress associated with reproductive issues, PCOS may significantly negatively affect a woman's mental health. Researchers can detect the psychological effects of PCOS and provide choices for supportive care and therapy for those impacted by it.
- Personalized medicine: PCOS appears differently in each person since it is a complex and diverse disorder. Personalized treatment plans based on a person's unique symptoms, hormone levels, and risk factors may be developed due to PCOS research.
- Implications for public health: Healthcare systems may better monitor and treat PCOS by allocating resources and being aware of the condition's prevalence and effects. Additionally, it educates decision-makers on the value of women's health and the need for early identification and care.
- Future generations: Studies reveal a relationship between maternal PCOS and an increased risk of certain health disorders in children, suggesting that PCOS may affect offspring's health [26]. An understanding of this possible transgenerational influence may guide preconception and prenatal treatment.

Further, research on PCOS is essential for improving the diagnosis, treatment, and general well-being of women with the condition. It may result in improvements in medical therapies, lifestyle modifications, and public health measures to manage this widespread and complicated problem successfully.

The efficient analysis of the PCOS prediction using a traditional statistical approach is much different and challenging. Using a Machine Learning approach is encouraged to overcome and perform efficient analysis. Techniques based on Machine Learning have significantly contributed to the study and prediction of Polycystic Ovary Syndrome (PCOS) by exploiting large datasets and complex patterns that may be challenging to analyse using standard statistical methods.

The analysis of PCOS prediction using Machine Learning is standard and has shown encouraging results in several studies. Using a range of variables and risk factors, Machine Learning algorithms can efficiently collect and evaluate complicated data to predict whether a person would develop PCOS. Machine Learning is crucial for PCOS analysis since it enhances improved diagnosis, personalized predictions, risk stratification, feature selection, multimodal data integration, early detection and intervention, clinical decision support, predictive biomarkers, model optimization, and big data handling.

To evaluate PCOS prediction, various ML techniques may be used. Features of the data, the size of the dataset, the demands for interpretability, and the precise goals of the analysis all influence the method used. This work applies six various ML techniques: RF, LR, SVM, NB, KNN and XGB classifier. The efficacy of these algorithms in prior research and the characteristics of the PCOS dataset utilized for analysis impact the decision to employ them for PCOS prediction.

The experimental process uses the PCOS dataset obtained from the Kaggle Repository. The six models' outputs are anticipated and contrasted using performance metrics ratings. The research of PCOS using Machine Learning models offers significant benefits to the medical industry and aids in treating PCOS problems. The remainder of the paper is organized as follows: 2. Literature review, 3. Methodology, 4. Results & conclusion.

2. Literature Survey

[5] the article examines the effects of Polycystic Ovarian Syndrome (PCOS) on women's reproductive health, such as infertility and miscarriage. PCOS is a syndrome characterized by hormonal abnormalities and abnormal male hormone production. Due to the variety of symptoms and related gynaecological problems, it is not easy to diagnose. The research presents a technique for early identification and prediction of PCOS using clinical and metabolic characteristics.

Data from 541 women was evaluated, and eight possible attributes were discovered. Many machine-learning approaches were tested for classification, with the RF model being the most accurate method. The study emphasizes the significance of ML in healthcare improvement and examines its applications in detection and prediction in the medical field.

[6] an automated approach for the identification prognosis of PCOS-related problems in women is presented in this paper. The model employs a fuzzy approach to account for the linguistic aspect of symptoms and diagnoses. The SVM algorithm and TOPSIS approach are both analyzed and contrasted. The findings indicate that compared to SVM's accuracy of 94.01%, the Fuzzy TOPSIS technique obtains a superior accuracy of 98.20%. Preventive actions may be aided by early identification and treatment of PCOS and mental health conditions. Women's psychological health is also assessed and might be considered while developing PCOS treatment plans.

[7, 8] the development of PCOS prediction using ML methods is explored in this article. In women of childbearing age, PCOS is a prevalent endocrine issue that may cause infertility and other health problems. PCOS prediction uses classification techniques like KNN, NB, DT Classifier, SVM, and LR. The most reliable model for predicting PCOS in this research was the Decision Tree Classifier.

[9] the research uses a Kaggle dataset with 541 women, 177 of whom have PCOS. Univariate feature selection is used to determine the most relevant features in the dataset. Different ML Models like GB, RF, LR, hybrid RF and RFLR are applied for PCOS prediction. The findings indicate that the top 10 features adequately predict PCOS and utilise 40-fold cross-validation. Results indicate that RFLR performs better.

[10] PCOS is A frequent hormonal issue affecting women and may result in infertility. PCOS may be correctly identified by utilizing ultrasound scans that identify multiple cysts. This work, 594 ultrasound pictures are used to predict PCOS using Machine Learning. The researchers employed transfer learning and a CNN to extract information from the images. To divide PCOS non-PCOS ovaries, they employed a stacking ensemble model using Machine Learning models as base learners and bagging or boosting as a meta-learner. Comparing the suggested method to previous Machine Learning approaches, accuracy increases while training time minimises. The VGGNet16 pre-trained model using a CNN as the feature extractor and XGBoost as the image classifier, attaining 99.89% accuracy, achieved the best results.

[11] this article focuses on identifying PCOS in female patients utilizing data-driven techniques. The research uses a Kaggle dataset with 177 PCOS-afflicted women and 43 distinct features. The most precise attributes for predicting PCOS are found via univariate feature selection and deletion. The most important element is discovered to be the ratio of LH to FSH. The dataset is experimented with using ensemble ML techniques, such as CatBoost, voting hard, and voting soft. The findings demonstrate that the top 13 risk variables may precisely forecast the beginning of PCOS. The soft voting method obtains the best accuracy of 91.12% when using cross-validation.

[12] researchers recommend using XGBoost for early identification of PCOS. They resampled the data using SMOTE and ENN to resolve class disparities and data outliers. The ANOVA and Chi-Square tests discovered 23 relevant metabolic and clinical features for PCOS disorders. In numerical data-driven PCOS diagnosis, the Extreme Gradient Boosting classifier surpassed all other models with a 98% recall rate in identifying individuals without PCOS.

[13] a common endocrine system condition called PCOS affects 5–10% of teenagers. Early detection and intervention may lower the risk. RapidMiner and Python-Scikit Learn are used in this work to predict PCOS. Random Forest gets the maximum accuracy with the whole dataset, whereas KNN and SVM exhibit comparable performance with ten chosen features. RapidMiner outperforms Python; however, performance depends on the dataset and approaches used.

[14] in recent days, people struggled with a no of illnesses, including PCOS, which affects 20% of the population in India. Effectively identifying chronic illnesses may be aided by Machine Learning approaches. This study uses SVM, logistic regression, and random forests to create a diagnostic and prompt preventative system. The strategy for choosing features based on statistical data is correlation-based. On the PCOS dataset created by Prasoon Kottarathil, the system was examined, and SVM showed an accuracy of 70.55%, while logistic regression and random forest showed an accuracy of 90.18% and 92.024%, respectively. Gynaecologists may get a second opinion and diagnose PCOS using the accepted CAD approach.

[15] PCOS is a problem that results in impotence, gynecomastia, and hirsutism. To tackle this problem,

ultrasound pictures may be analyzed. However, a lack of unbiased diagnostics makes identifying and comprehending PCOS difficult. Time spent manually tracing and measuring follicles might be saved with an automated diagnosis tool. The proposed approach improved the time spent detecting PCOS. It decreased the danger of fatal consequences brought on by delayed diagnosis by achieving an accuracy of over 97% using a KNN classifier.

Table 1. Summary of literature survey

References	Methodology
5	PCA and different ML applied NB, LR, KNN, SVM, RF, CART applied. RF shows superior results compared to others. PCOS dataset with 541 instances applied.
6	The TOPSIS approach and SVM model were experimented with. TOPSIS shows superior results. PCOS dataset used.
7	Different ML applied KNN, NB, DT, SVM, LR.
8	Different ML applied GB, RF, LR, RFLR.
9	A Machine Learning approach predicts PCOS using ultrasound images, CNN, transfer learning, and stacking ensemble models, achieving 99.89% accuracy and reducing training time.
10	Univariate feature selection, CatBoost, voting hard, and voting soft applied.
11	SMOTE & ENN for resampling, ANOVA & Chi-Square Test for features selection, XGBoost perform superior to other classifiers.
12	Random Forest, K-NN, and Support Vector Machine are applied.
13	Correlation-based FS for choosing features. SVM, LR & RF.
14	Results compared with KNN with DT, NB and SVM. Results show that KNN is Superior.

3. Methodology

Algorithm

I-Input, O-Output, D-Dataset, X-features, Y-Class Label

```
Step 1 : Data Collection.
    I  : PCOS data D with attributes X and class variable Y (PCOS diagnosis labels)
Step 2 : Data Pre-processing.
    I  : D, X, Y
    O  : D_Processed dataset
  1. Clean D_Processed for missing values & outliers.
  2. Encode categorical variables in D.
  3. Normalize numerical attributes in D.
Step 3 : Splitting the Dataset.
    I  : D, X, Y
    O  : D_Processed dataset
  1. Split D to D_train and D-test dataset.
Step 4 : ML Selection.
    I  : D_train, X_selected, Y
    O  : Choisen model M
  1. Select the most suitable classification Machine Learning algorithm.
  2. Initialize the chosen model M.
Step 5 : Model Training.
    I  : D_train, X_selected, Y, M
    O  : Trained model M_traioned
  1. On the training dataset D-train, train the model M using the target variable y and the chosen features X_selected.
  2. Store the trained model in M_trained.
Step 6 : Model Evaluation.
    I  :D-test, X_selected, Y, M-trained
    O  : Model performance scores (recall, F1-score)
  1. Use X_selected and Y to assess the performance of the trained model M_trained on the testing dataset D_test.
  2. Using performance criteria to evaluate the ML model's efficiency.
```

The overall methodology is indicated in Figure 1.

3.1. Data Collection

The research aims to predict the PCOS condition, and the PCOS dataset is collected for the objective. The source of the dataset was obtained from the Kaggle repository. The dataset collected consists of 44 features and 541 instances.

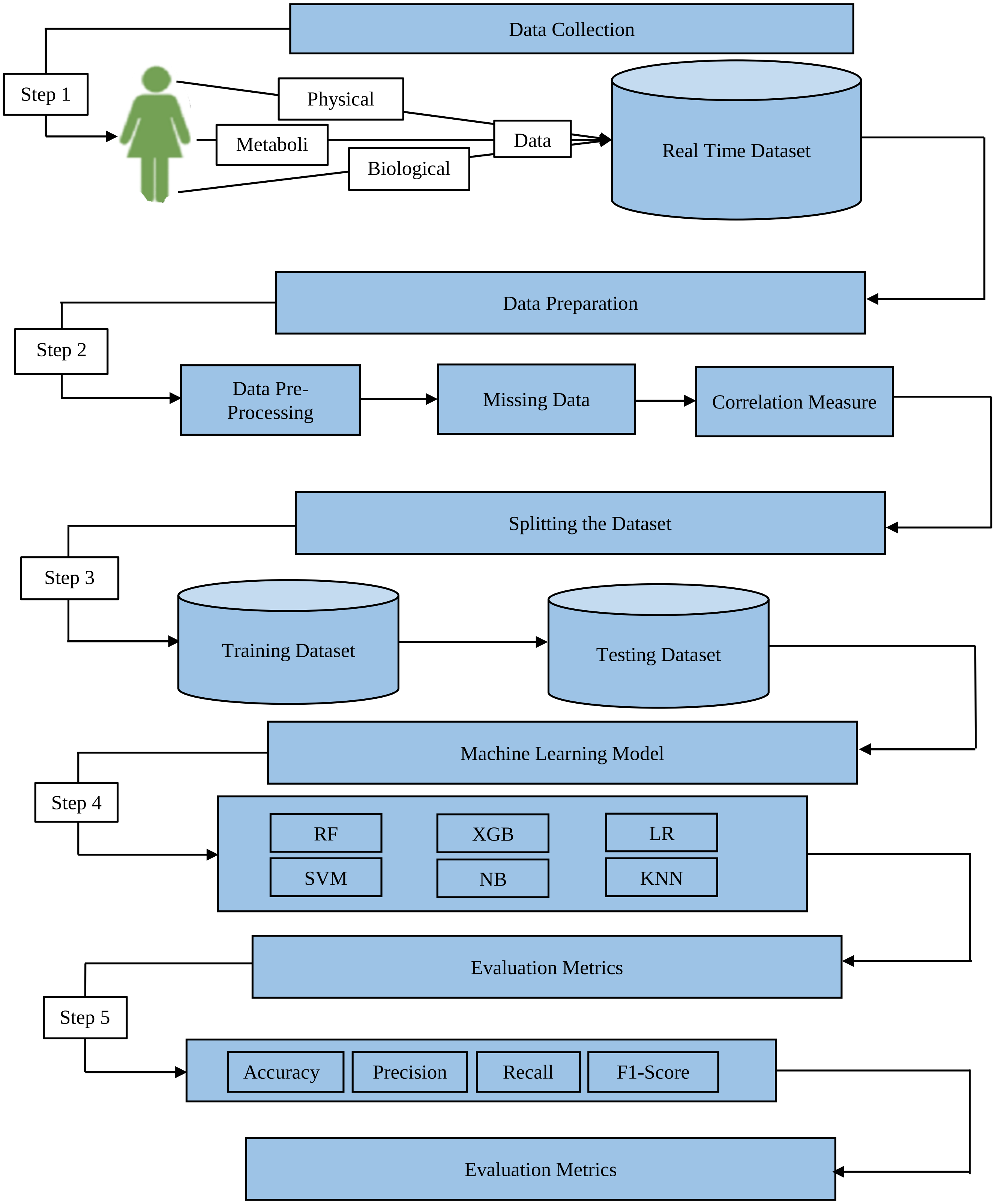

Fig. 1 Overall methodology of PCOS prediction

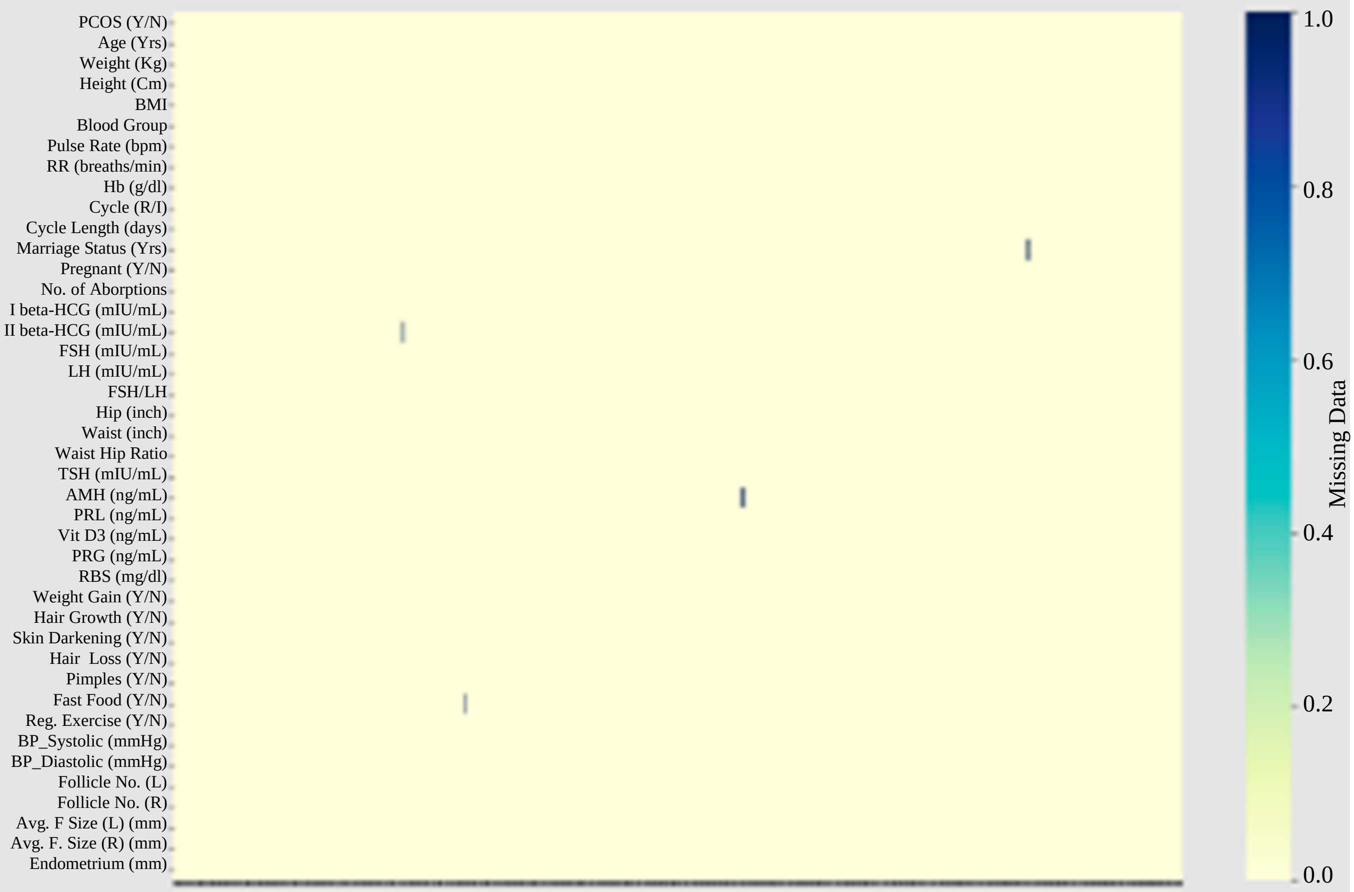

Fig. 2 Plotting of missing values in the PCOS dataset

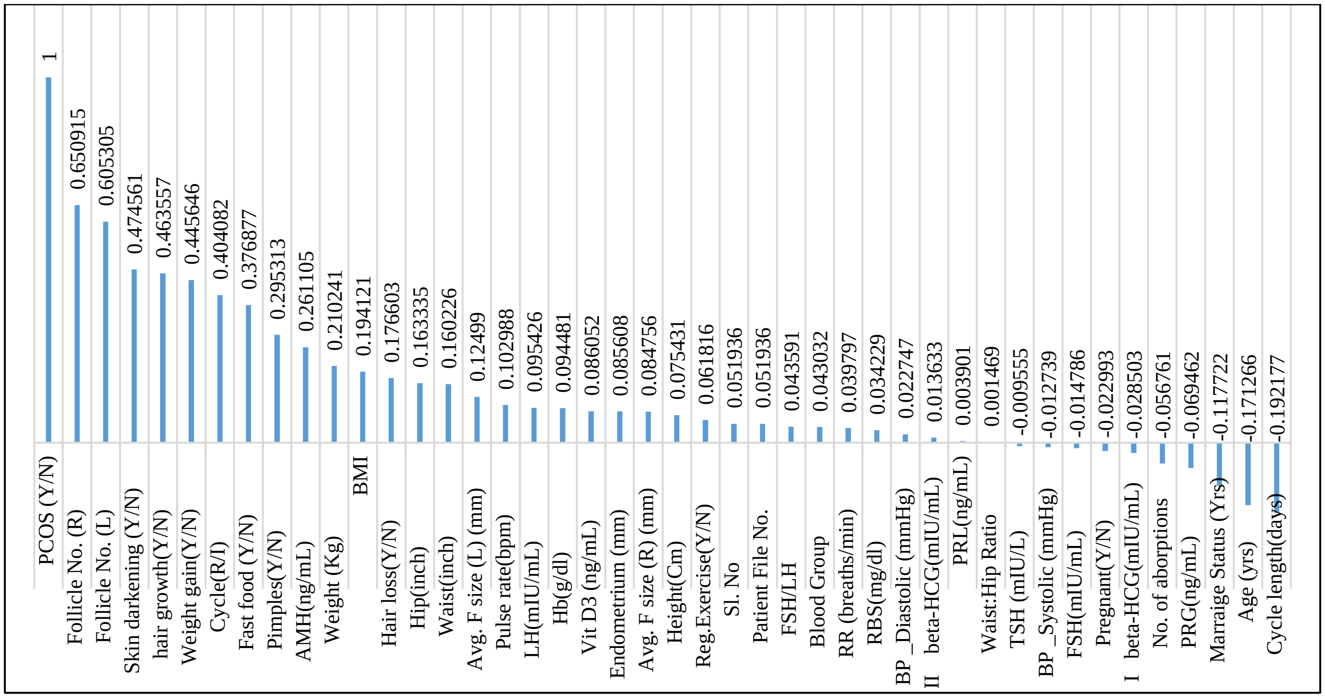

Fig. 3 Correlation matrix for PCOS dataset

3.2. Data Pre-Processing

Next, collected PCOS data was evaluated for the missing and noisy values. First, the dataset is marked using the heatmap to assess the missing values represented in Figure 2. From the Figure 2, there are some missing values in some features. Next, the missing values will be filed using the median values. The median value is taken and replaced by the missing values by computing the entire attribute column. Further the correlation with each feature in the class label is computed and projected in Figure 3.

3.3. Split the Dataset

The D_Dataset split into training testing data. Using the training data, the ML Model is built and using the testing data, the performance of the model is evaluated. Here, 70% is taken for the training, and 30 % is taken for the testing the model.

3.4. Machine Learning Model

The most essential step in the building of ML models is model selection. It entails selecting the correct algorithm or model design that solves the current issue and produces the best results on the available dataset. Model selection aims to find a model that can successfully generalize to novel, unexplored data and provide precise predictions or classifications. The research applies six ML models: RF, LR, SVM, NB, K-NN, and XGB.

3.4.1. Random Forest

RF is a form of ensemble learning that combines many DTs to produce a more robust, accurate model. It is commonly used for regression classification problems. A high-level approach for developing a Random Forest model for PCOS prediction is as follows:

Algorithm

Input : PCOS dataset with features (X) and labels (y)
Output : Trained Random Forest model

```
1. Data Preparation: Pre-process the dataset and split the dataset into features (X) and labels (y).
2. Random Forest Parameters:
   • Choose the number of trees (n_estimators) in the Random Forest.
   • Choose the number of features to consider at each split (m) in each tree.
3. For each tree (t = 1 to n_estimators):
   • Sample the training data with replacement using bootstrapping to create a new training dataset (X_train_t, y_train_t).
   • Randomly select a subset of features of size ‘m’ from the total set of features (X) at each split.
4. Train the Decision Tree:
   • Train a decision tree using the new training dataset (X_train_t, y_train_t) and the selected subset of features.
   • Use a criterion to determine the best split at each split node.
5. Combine Predictions:
   • For binary classification (PCOS prediction):
   • During prediction, get class predictions from each tree (t_pred) for a given test sample.
   • Take the majority vote of the class predictions as the final predicted class for the test sample.
6. Model Evaluation:
   • Evaluate the random forest model’s performance
```

3.4.2. Logistic Regression

Logistic regression is a well-known and commonly used Machine Learning approach for binary classification applications, making it ideal for PCOS prediction. It is used to simulate the association between a group of independent factors (predictor variables) and a binary outcome variable (target variable), for example, PCOS existence (1) or absence (0). The Logistic Regression model aims to discover the best-fitting parameters that maximize the probability of the observed data [16].

Algorithm

Input : PCOS dataset with features (X) and labels (y)
Output : Trained Random Forest model

```
1. Data Preparation: Pre-process the dataset and split the dataset, training_set, test_set = split_data(data)
2. Fit a logistic regression model to the training set.
   model = LogisticRegression ()
   model.fit (training_set, training_set[“PCOS”])
3. Predict the probability of PCOS for the test set.
   predictions = model.predict_proba(test_set)
   Calculate the different validity scores of the model.
4. Validity score = validity _score (test_set [“PCOS”], predictions)
   return validity score
```

3.4.3. Support Vector Machine

SVM is a sophisticated, adaptable Machine Learning technique widely used for binary classification applications such as PCOS prediction. SVM seeks the ideal hyperplane for separating data points of various classes (PCOS and non-PCOS) while maximizing the margin between them. It is effective in both linearly and non-linearly separable datasets [17].

Algorithm

Input : PCOS dataset with features (X) and binary labels (y)
Output : Trained Support Vector Machine model

```
1. Data Preparation: Pre-process the dataset and split the dataset, training_set, test_set = split_data(data)
2. SVM Hyperparameters: Choose the SVM
```

hyperparameters, such as the kernel type, the kernel parameters and the regularization parameter (C).

3. Train the SVM model: Initialize the SVM model with the chosen hyperparameters.
   For linear SVM:
   svm_model = LinearSVC(C=C)

$Train\ the\ SVM\ model\ on\ the\ training\ data$:

$$svm\_model.fit(X\_train, y\_train)$$

4. Model Prediction:
   For each test sample x_test in X_test:
   - Predict the class label y_pred for x_test using the trained SVM model:
     y_pred = svm_model.predict(x_test)
5. Model Evaluation: Compute the SVM model on the test set using appropriate metrics like accuracy, precision, recall, F1-score, etc.

3.4.4. Naive Bayes Model

A popular ML technique for classification and text categorization tasks is called Naive Bayes. It is simple yet very effective. It is based on Thomas Bayes' 18th-century Bayes' theorem, a probabilistic framework. The feature independence assumption, which makes the calculation more accessible and efficient, particularly for high-dimensional data, gives the model its "naive" quality. Due to its efficiency and simplicity of usage, the Naive Bayes model is often employed in various applications, including spam detection, sentiment analysis, document categorization, and recommendation systems [18, 22].

Algorithm

Input **:** PCOS dataset with features (X) and binary labels (y)

Output **:** Trained Naive Bayes model

1. Data Preparation: Pre-process the dataset and split the dataset, training_set, test_set = split_data(data)
2. Class Prior Probabilities:
   Calculate the prior probabilities for each class (PCOS = 1 and non-PCOS = 0) based on the training data:
   - p(PCOS) = (Number of PCOS samples) / (Total number of samples)
   - p(non-PCOS) = (Number of non-PCOS samples) / (Total number of samples)
3. Feature Likelihoods:
   For each feature 'F' in the dataset:
   For each class 'C' (PCOS and non-PCOS):
   Calculate the likelihood of observing feature 'F' given class 'C':
   - For continuous features: Assume a Gaussian (normal) distribution and calculate each class's mean and standard deviation.
   - For discrete features: Calculate the frequency of each unique value in each class.
4. Model Training:
   - The Naive Bayes model does not require an explicit training step, as the likelihoods and priors are computed directly from the data.
5. Model Prediction:
   For each test sample x_test in X_test:
   - Calculate the posterior probability of each $class\ 'C'\ given\ the\ test\ sample$:
   - For each class 'C':
   - Calculate the likelihood of the $test\ sample$ given class 'C' using the feature likelihoods.
   - Calculate the posterior probability $p(C|x\_test) = p(C) * p(x\_test|C)$ (using the Naive Bayes assumption).
   - Assign the test sample to $the\ class\ with\ the\ highest$ posterior probability:
   - If p(PCOS|x_test) > p(non-PCOS|x_test), predict class 1 (PCOS).
   - Otherwise, predict class 0 (non-PCOS).
6. Model Evaluation: Compute the Naive Bayes model on the test set using appropriate metrics like accuracy, precision, recall, and F1-score.

3.4.5. KNN

KNN is a simple but effective machine-learning technique that may be used for several tasks, including predicting PCOS. The KNN method predicts the label of the test sample based on the labels of the k closest neighbours by locating the k training samples most similar to the current test sample. K's value is a hyperparameter that the user must choose. The KNN algorithm may be used in PCOS prediction to determine if a patient has PCOS based on their medical history, symptoms, and other variables. The algorithm would first identify the k patients in the training set who were most similar to the new patient. The algorithm would predict that the new patient also has PCOS if most of the k closest neighbours had the condition. The KNN approach makes no assumptions about the data's underlying distribution since it is a non-parametric algorithm. This makes it a flexible approach that can be used for many sorts of data. The KNN approach, however, may be computationally costly, mainly when there are many training examples. Furthermore, the selection of the hyperparameter k may impact the algorithm's accuracy [19].

Algorithm

Step 1 : Calculate distances between test_sample and all PCOS_samples

Step 2 : Sort distances in ascending order

Step 3 : Select the top k samples with the smallest distances

Step 4 : Determine the majority class label among the k nearest neighbour

Step 5 : Calculate the different validity scores of the model

3.4.6. XGB Classifier

The popular Machine Learning technique Extreme Gradient Boosting (XGBoost) is efficient for several applications, including PCOS prediction. Since XGBoost is a gradient-boosting technique, it creates a group of decision trees. Every tree in the ensemble is taught to fix the mistakes made by the earlier trees. As a result, XGBoost may understand intricate associations in the data and provide precise forecasts. XGBoost may determine a patient's likelihood of having PCOS based on their medical history, symptoms, and other details. An ensemble of decision trees would initially be constructed using the algorithm. Based on a subset of the attributes, each tree would be trained to determine if a patient has PCOS. The final prediction would be generated by averaging the forecasts of each tree. Numerous studies have demonstrated that XGBoost helps predict PCOS [20, 21].

Algorithm

Input : PCOS dataset with features (X) and labels (y)
Output : Trained Random Forest model

```
1. Data Preparation: Pre-process the dataset and split the
   dataset into features (X) and labels (y).
2. Create an XGB classifier.
   classifier = xgb.XGBClassifier()
3. Train the classifier on the training set.
   classifier.fit(training_set, training_set[“PCOS”])
4. Predict the probability of PCOS for the test set.
   predictions = classifier.predict(test_set)
5. Calculate the accuracy of the model.
   accuracy = accuracy_score(test_set[“PCOS”],
   predictions)
```

3.5. Evaluation Metrics

Evaluation metrics are used in PCOS (Polycystic Ovary Syndrome) prediction, just like in any other classification problem, to rate the effectiveness of the predictive model. These measures assist in estimating the model's accuracy in formulating predictions and its ability to discriminate between positive (PCOS) and negative occurrences. These evaluation criteria for PCOS prediction used are:

3.5.1. Accuracy

The most fundamental evaluation statistic is accuracy. The formula is as described below:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

3.5.2. Precision

Precision, sometimes called positive predictive value, analyzes how well the model predicts the future. It determines the ratio of accurate forecasts to all accurate predictions (including both true and false positives):

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (2)$$

3.5.3. Recall (Sensitivity)

The model's capacity to accurately identify positive occurrences among all the actual positive cases in the dataset is measured by recall, also known as sensitivity or true positive rate. It determines the ratio of correctly predicted positive outcomes to all positive instances [23]:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

3.5.4. F1-Score

A high F1-score indicates a precision/recall ratio that is well balanced [24, 25]:

F1-Score = 2 * (precision * recall) / (precision + recall) (4)

4. Experimental Results

The research aims to analyse whether women with PCOS problems are not and if they take the necessary medical treatment to cure it in the initial stage. For this, different ML algorithm models are considered to perform prediction.

4.1. Experimental Procedure

Step 1 : DataCollection
$D = \{(x_1, y_1), (x_2, y_2), \ldots \ldots, (x_n, y_n)\}$
Step 2 : DataPre-processing.
$D^P = \text{Preprocess Data}(D)$
Step 3 : Model Selection
Choose the ML Model
$\text{Model} = \text{ChooseModel}()$
Step 4 : Data Split
Split the D^P into training & testing group
$(X_{train}, y_{train}), (X_{test}, y_{test}) = \text{TrainTestSplit}(D^P)$
Step 5 : Model Training
Now, with the training set, train the Chosen ML Model.
$\text{Model.train}(X_{train}, y_{train})$`
Step 6 : Model Evaluation
Assess the Chosen ML model performance on testing group data.
y_pred = Model.predict(X_test)
accuracy = CalculateAccuracy(y_test,y_pred) Precision
= CalculatePecision(y_test,y_pred)
Recall = CalculateRecall(y_test,y_pred)
F1_Score = CalculateF1_Score(y_test,y_pred)
Step 7 : Compare the performance of the ML model and project the best model

4.2. Results of PCOS Analysis Using Different Machine Learning Models

The highest performance in Figure 4 was Random Forest, which achieved an excellent accuracy of 0.90. This demonstrates Random Forest's great predictive skills for PCOS diagnosis and shows that it correctly predicted 90% of the occurrences in the sample.

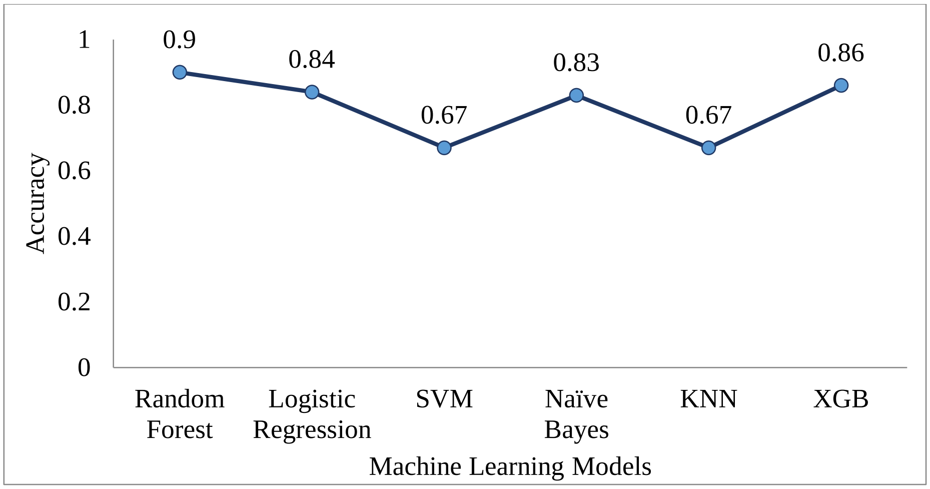

Fig. 4 Results of PCOS prediction using different Machine Learning models using accuracy parameters

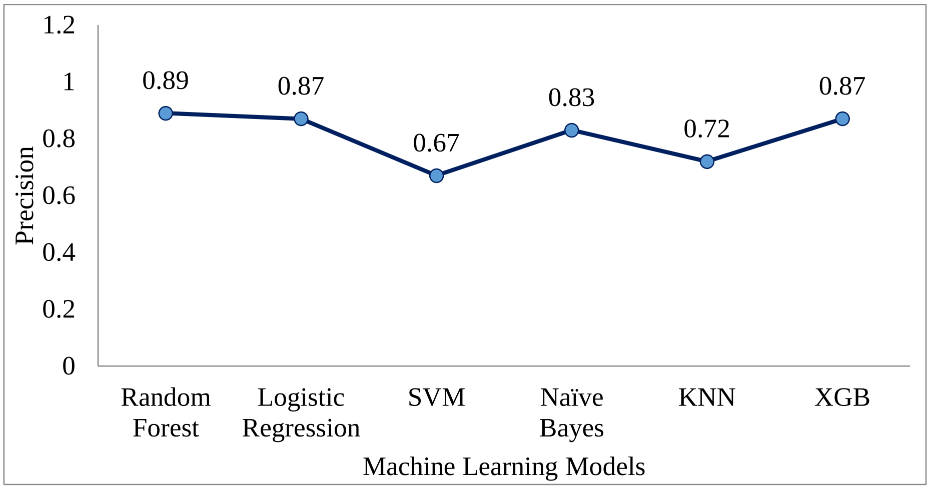

Fig. 5 Results of PCOS prediction using different Machine Learning models using precision parameters

The ensemble aspect of random forest, which combines numerous decision trees, probably made it possible to identify intricate patterns and generalize effectively to new data. The accuracy of 0.86 shown by XGB, which trailed closely behind, was competitive and further supported its efficacy in PCOS prediction. XGB is a reliable classifier for this challenge because of its gradient-boosting methodology and capacity for dealing with non-linear interactions. Naive bayes and logistic regression performed well, with accuracy values of 0.84 and 0.83, respectively. These models are well-known for being straightforward to understand, which makes them good options for binary classification problems like PCOS prediction. SVM and KNN, on the other hand, produced lesser accuracies of 0.67.

Even though SVM's strong recall (ability to recognize all positive examples) would be advantageous in certain circumstances, its accuracy and precision deteriorated, indicating that it might incorrectly categorize some negative occurrences. Despite being straightforward and understandable, KNN may not be the best option for this dataset due to its poor accuracy. The study results based on precision scores from several Machine Learning Models for PCOS prediction provide essential insights into their capacity to categorize positive instances accurately. With both models reaching a precision of 0.87 among the models tested, Random Forest and XGB stood out as the best performers.

This shows their efficiency in reducing false positives and correctly detecting PCOS patients by showing that 87% of the instances these models predicted as positive were real positives. With a precision of 0.87, Logistic Regression came in second place, demonstrating its effectiveness as a trustworthy classifier.

Naive Bayes' competitive ability in precisely categorizing affirmative instances is shown by its accuracy score 0.83. Despite having a perfect recall, SVM's accuracy of 0.67 indicates that it may often mistakenly identify negative cases as positive, which would reduce its total precision. KNN demonstrated a respectable capacity for identifying affirmative instances, with a precision of 0.72.

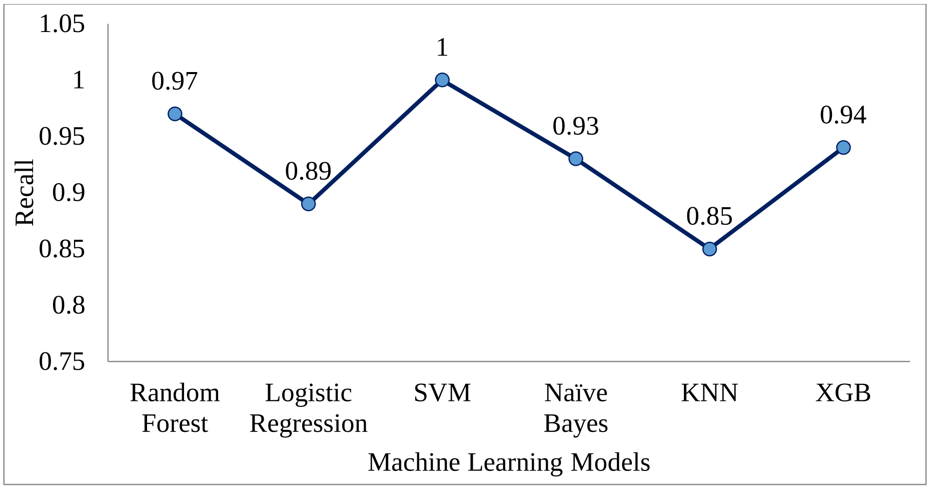

Fig. 6 Results of PCOS prediction using different Machine Learning models using recall parameters

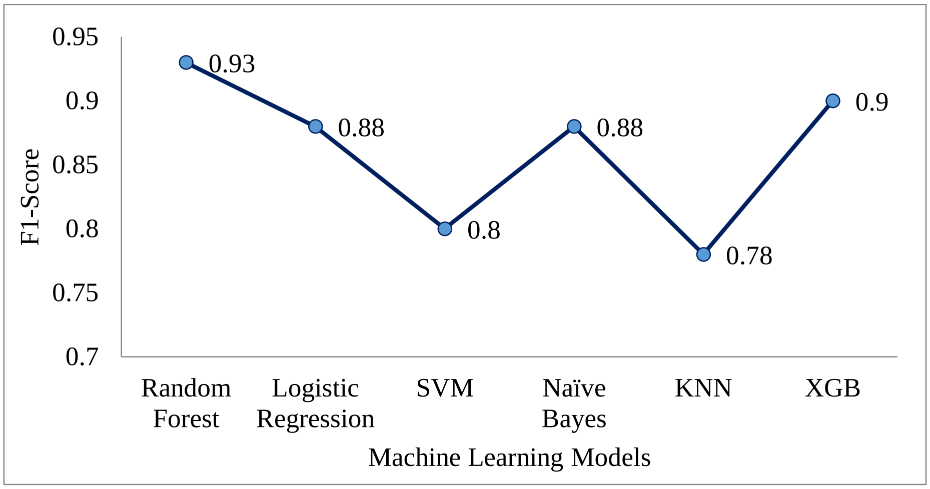

Fig. 7 Results of PCOS prediction using different Machine Learning models using F1-score parameters

The study results based on the recall scores of several Machine Learning Models for PCOS prediction provide essential insights into their capacity to identify positive instances accurately. SVM (Support Vector Machine) stood out among the analyzed models with a flawless recall score of 1.00, demonstrating that it accurately recognized all genuine positive PCOS cases in the dataset. This remarkable recall is encouraging, but it is essential to consider other factors like precision and accuracy since SVM may have a more significant percentage of false positives.

Strong competitors for accurate PCOS identification, random forest and XGB, also revealed high recall scores of 0.97 and 0.94, respectively. With a recall of 0.93, Nave Bayes has shown a solid capacity to recognize good occurrences. Although the recall scores for Logistic Regression and KNN were significantly lower, at 0.89 and 0.85, respectively, they showed respectable competence in collecting positive instances. Overall, the results indicate that owing to their high recall scores, SVM, random forest, XGB, and nave bayes are viable models for PCOS prediction.

The F1-score findings provide in-depth perceptions of the general effectiveness of several Machine Learning models for PCOS prediction. Of all the models tested, random forest has the highest F1-Score (0.93), demonstrating its exceptional capacity to find an ideal balance between accurately recognizing positive PCOS cases and reducing false positives. This exceptional result places random forest as the best option for precise and trustworthy PCOS prediction. In a close second place, XGB showed an F1-Score of 0.90, confirming its efficiency as a potent classifier for this job. As evidence of their balanced performance in reliably collecting positive cases while reducing false positives, logistic regression and naive bayes demonstrated competitive F1 scores of 0.88. These models demonstrate that they are strong PCOS prediction competitors.

SVM had a flawless recall with an F1-score of 0.80, demonstrating its capacity to recognize all instances of positivity. Its accuracy, however, may affect the total F1-score, indicating potential misclassification of negative cases as positive. This demonstrates the importance of considering accuracy and recall when assessing SVM performance. KNN obtained an F1-score of 0.78, indicating a little uneven performance compared to other models. KNN efficiently finds positive cases, although it may produce more false positives than true ones.

4.3. Result Discussion

Figure 8 shows the outcomes of PCOS prediction using several ML algorithms and validity scores used to assess the effectiveness of the models. Upon analysing the results, the random forest model had the most remarkable accuracy score (0.90), correctly classifying 90% of the instances. Additionally, it has a high recall of 0.97, indicating that many real PCOS cases were detected. The F1-score of 0.93 shows that it successfully balances recall and accuracy. The accuracy of the Logistic Regression model was 0.84. It effectively avoids misclassifying non-PCOS patients because of its high accuracy of 0.87. The recall of 0.89, however, raises the possibility that some PCOS cases may have been overlooked.

The SVM model’s accuracy score of 0.67 is the lowest. Although it has a perfect recall of 1.00, which means it recognizes all actual PCOS instances, the accuracy is just 0.67, which is not very high. This suggests that there are a lot of false positives in it. With an accuracy of 0.83, precision of 0.83, and recall of 0.93, the Naive Bayes model performed well. Its F1-score of 0.88 indicates a fair balance between recall and accuracy. Like the SVM model, the KNN model had an accuracy of 0.67. Its F1-score was 0.78 because of its accuracy of 0.72 and recall of 0.85. With a precision of 0.87, recall of 0.94, and accuracy of 0.86, the XGB model performed well. Like random forest, it performs well overall, as seen by its F1-score of 0.90.

It is clear that in practically every category, the Random Forest model fared better than the other methods. It accurately predicted 90% of the instances in the dataset, an outstanding accuracy of 0.90. Additionally, the model showed excellent recall (0.97) and accuracy (0.89), demonstrating that it could correctly identify PCOS-positive cases while limiting false positives. The model’s high performance and ability to achieve a harmonic trade-off between recall and accuracy were further supported by the balanced F1-score of 0.93. Logistic Regression also demonstrated respectable performance with an accuracy of 0.84 and a well-balanced F1-Score of 0.88, indicating its capacity to produce reliable predictions.

SVM, however, distinguished out because of its flawless recall (1.00), which indicates that it accurately detected all actual positive instances. However, the SVM’s lower precision (0.67) and accuracy (0.67) suggest that it could have mistakenly categorized negative examples as positive. With F1 scores of 0.88 and 0.90, respectively, reflecting their balanced performance, Nave Bayes and XGB demonstrated encouraging results. However, KNN and SVM demonstrated lower accuracies than other algorithms, indicating that they may not be the best option for this prediction job.

In conclusion, the models with the highest accuracy, precision, recall, and F1 score were Random Forest and XGB. While KNN had a balanced precision and recall but significantly lesser accuracy, SVM had a flawless recall but struggled with poor precision. Regarding overall performance, Random Forest and XGB surpassed Logistic Regression and Naive Bayes, which did only reasonably well. These results may be a starting point for developing an effective PCOS prediction system to aid medical practitioners in early diagnosis and intervention.

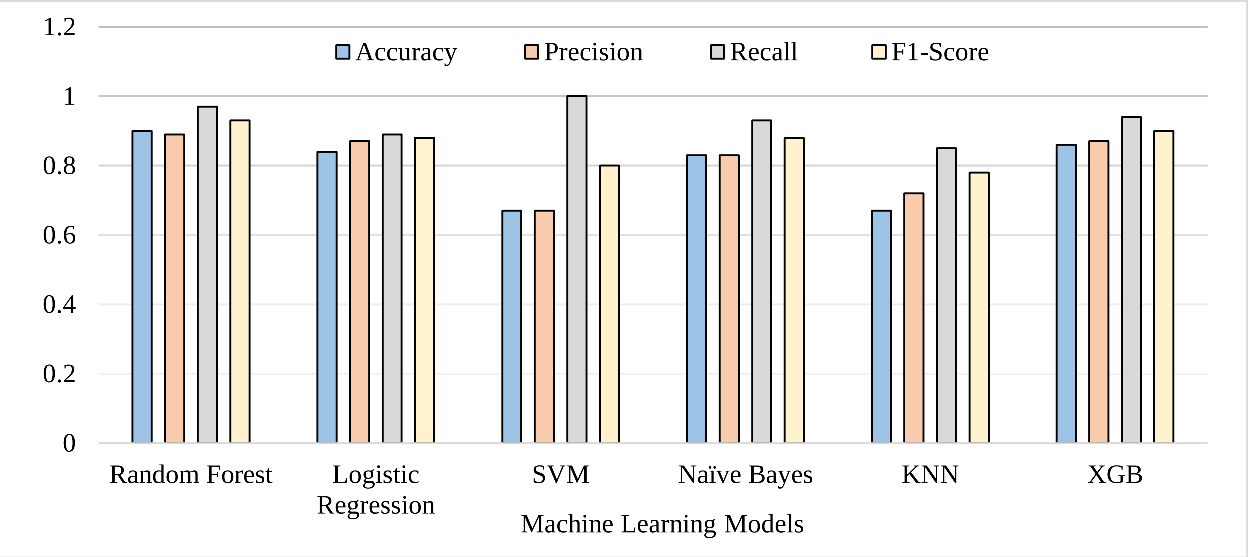

Fig. 8 Results of PCOS prediction using different Machine Learning algorithm and projected using different validity scores

4.4. Research Findings

1. The study evaluated several Machine Learning methods for PCOS prediction to determine the most efficient model for precise categorisation.
2. The Random Forest algorithm came out on top, scoring the best accuracy (0.90) and a balanced F1-Score (0.93), demonstrating its potent prediction ability.
3. Random Forest showed superior recall (0.97) and accuracy (0.89), displaying its capacity to correctly identify PCOS-positive patients while reducing false positives.
4. With respectable accuracy and balanced F1-Scores (0.88 for Logistic Regression and Nave Bayes, and 0.90 for XGBoost), Logistic Regression, Nave Bayes, and XGBoost have shown promise as workable alternatives.
5. Despite the SVM's perfect recall (1.00), which demonstrated its capacity to recognize all positive examples, its lower accuracy (0.67) and precision (0.67) raised questions about its tendency to categorize negative occurrences as positive mistakenly.
6. KNN's accuracy score was lower (0.67) than other models, suggesting that it may not be the best option for predicting PCOS on this dataset.
7. The results highlight the relevance of selecting the best Machine Learning algorithm for a particular job to provide an accurate PCOS prediction.

5. Conclusion

Many women of reproductive age are affected by the prevalent endocrine condition known as PCOS. Numerous health problems, such as infertility, insulin resistance, obesity, and an elevated risk of cardiovascular disease, may be brought on by PCOS. Prompt diagnosis and action are essential to treat PCOS and avoid possible consequences successfully. Accurately diagnosing PCOS may be difficult for medical experts due to the intricacy of the disorder and the wide range of symptoms across people. Machine Learning (ML) methods have recently shown remarkable promise to help diagnose and predict PCOS. Large volumes of data may be processed by ML models, which can also spot trends that would be difficult to see with conventional diagnostic techniques.

We did a research study to assess the performance of several ML algorithms to address the significance of accurate PCOS prediction. The main goal was to find the model that best discriminates between positive PCOS and non-PCOS instances. Our dataset was valuable for creating reliable prediction models since it included pertinent clinical and biological characteristics from various individuals. Random Forest, XGB (Extreme Gradient Boosting), Logistic Regression, SVM (Support Vector Machine), Nave Bayes, and KNN (K-Nearest Neighbours) are the six well-known ML methods we used for our study. Each model was trained and evaluated on the dataset to ensure a fair assessment using a cross-validation procedure.

The experimental findings provided insightful information about how well the various models performed. The top-performing models were Random Forest and XGB, which obtained remarkable F1 scores of 0.93 and 0.90, respectively. These algorithms accurately identified positive PCOS patients while reducing false positives. They are strong candidates for PCOS prediction because of their ensemble-based and gradient-boosting techniques, which enable them to capture complicated correlations in the data. With an F1-Score of 0.88, Logistic Regression and Naive Bayes also showed their viability as alternatives. When decision-making openness is crucial, their more accessible nature and interpretability make them appealing options. SVM earned a flawless recall score of 1.00, demonstrating its ability to recognise every instance of PCOS.

Nevertheless, given that it has an overall F1-Score of 0.80, it may have a greater risk of false positives, necessitating cautious evaluation in real-world applications. With an F1-Score of 0.78, KNN performed well in detecting affirmative examples, although it showed a little precision/recall imbalance that suggested the potential for improvement.

Finally, our study demonstrates the potential of ML models for precise PCOS prediction. The results emphasize the significance of choosing suitable models based on the particular specifications of clinical applications. Top-performing models include Random Forest and XGB, effectively detecting positive instances while reducing false positives. SVM and KNN perform well, while Naive Bayes and Logistic Regression offer competitive options. Successfully deploying accurate PCOS prediction algorithms may enable medical personnel to provide prompt therapies, improve patient outcomes, and raise the standard of care for PCOS patients.

Future studies may improve model interpretability and applicability to other patient demographics, eventually resulting in more individualized and efficient PCOS treatment approaches. Overall, applying ML approaches to healthcare holds enormous promise and has the potential to significantly influence the identification and management of complicated medical diseases like PCOS.

References

[1] S. Minooee et al., "Prediction of Age at Menopause in Women with Polycystic Ovary Syndrome," *Climacteric*, vol. 21, no. 1, pp. 29-34, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[2] Angela Zigarelli, Ziyang Jia, and Hyunsun Lee, "Machine-Aided Self-Diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study," *JMIR Formative Research*, vol. 6, no. 3, pp. 1-23, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Miaoxian Ou et al., "AMH is a Good Predictor of Metabolic Risk in Women with PCOS: A Cross-Sectional Study," *International Journal of Endocrinology*, vol. 2021, pp. 1-7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Jacob P. Christ, and Tommaso Falcone, "Bariatric Surgery Improves Hyperandrogenism, Menstrual Irregularities, and Metabolic Dysfunction among Women with Polycystic Ovary Syndrome (PCOS)," *Obesity Surgery*, vol. 28, pp. 2171-2177, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[5] Amsy Denny et al., "i-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," *2019 IEEE Region 10 Conference (TENCON)*, Kochi, India, pp. 673-678, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ashwini Kodipalli, and Susheela Devi, "Prediction of PCOS and Mental Health Using Fuzzy Inference and SVM," *Frontiers in Public Health*, vol. 9, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Preeti Chauhan et al., "Comparative Analysis of Machine Learning Algorithms for Prediction of PCOS," *2021 International Conference on Communication Information and Computing Technology*, Mumbai, India, pp. 1-7, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Dipo Theophilus Akomolafe, Oluwatoyin Mary Yerokun, and Ayo Fasakin, "Resolving Some Critical Issues in the Prevention, Diagnosis, Treatment and Management of Covid-19 Using Machine Learning," *International Journal of Computer and Organization Trends*, vol. 10, no. 4, pp. 1-8, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Subrato Bharati, Prajoy Podder, and M. Rubaiyat Hossain Mondal, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms," *2020 IEEE Region 10 Symposium*, Dhaka, Bangladesh, pp. 1486-1489, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[10] Sayma Alam Suha, and Muhammad Nazrul Islam, "An Extended Machine Learning Technique for Polycystic Ovary Syndrome Detection Using Ovary Ultrasound Image," *Scientific Reports*, vol. 12, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Subrato Bharati et al., "Ensemble Learning for Data-Driven Diagnosis of Polycystic Ovary Syndrome," *International Conference on Intelligent Systems Design and Applications*, pp. 1250-1259, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Muhammad Sakib Khan Inan et al., "Improved Sampling and Feature Selection to Support Extreme Gradient Boosting For PCOS Diagnosis," *2021 IEEE 11th Annual Computing and Communication Workshop and Conference*, USA, pp. 1046-1050, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Satish C.R. Nandipati, Chew Xin Ying, and Khaw Khai Wah, "Polycystic Ovarian Syndrome (PCOS) Classification and Feature Selection by Machine Learning Techniques," *Applied Mathematics and Computational Intelligence*, vol. 9, pp. 65- 74, 2020. [Google Scholar] [Publisher Link]

[14] Harsita Batra et al., "Machine Learning Techniques for Data-Driven Computer-Aided Diagnostic Method of Polycystic Ovary Syndrome (PCOS) Resulting from Functional Ovarian Hyperandrogenism (FOH)," *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, Greater Noida, India, pp. 195-201, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] B. Rachana et al., "Detection of Polycystic Ovarian Syndrome Using Follicle Recognition Technique," *Global Transitions Proceedings*, vol. 2, no. 2, pp. 304-308, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Kanish Shah et al., "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augmented Human Research*, vol. 5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Derek A. Pisner, and David M. Schnyer, *Support Vector Machine*, Machine Learning, pp. 101-121, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Smriti Gupta, Harsh Kumar Verma, and Divyansh Bhardwaj, "Classification of Diabetes Using Naive Bayes and Support Vector Machine as a Technique," *Operations Management and Systems Engineering*, Singapore, pp. 365–376, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[19] M.N.A.H. Sha'abani et al., "KNN and SVM Classification for EEG: A Review," *Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering*, Kuantan, Pahang, Malaysia, pp. 555-565, 2020. [Google Scholar] [Publisher Link]

[20] Nayan Kumar Sinha et al., "Developing a Web-Based System for Breast Cancer Prediction Using Xgboost Classifier," *International Journal of Engineering Research and Technology*, vol. 9, no. 6, pp. 852-856, 2020. [Google Scholar] [Publisher Link]

[21] Zulaiha Parveen A., and T. Senthil Kumar, "The Deep Learning Methodology for Improved Breast Cancer Diagnosis in MRI," *International Journal of Computer and Organization Trends*, vol. 11, no. 3, pp. 11-14, 2021. [CrossRef] [Publisher Link]

[22] D. Prabha et al., "Performance Evaluation of Naive Bayes Classifier with and without Filter-Based Feature Selection," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 10, pp. 2154-2158, 2019. [Google Scholar] [Publisher Link]

[23] R. Gurumoorthy, and M. Kamarasan, “Wildebeest Habit Optimizer with Deep Learning-Based Histopathological Image Analysis for Breast Cancer Diagnosis,” *International Journal of Engineering Trends and Technology*, vol. 71, no. 7, pp. 233-243, 2023. [CrossRef] [Publisher Link]

[24] Mohammed S. Atoum et al., “A Fog-Enabled Framework for Ensemble Machine Learning-Based Real-Time Heart Patient Diagnosis,” *International Journal of Engineering Trends and Technology*, vol. 71, no. 8, pp. 39-47, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] S. Praveena Rachel Kamala et al., “Predictive Analytics for Heart Disease Detection: A Machine Learning Approach,” *4th International Conference on Electronics and Sustainable Communication Systems*, Coimbatore, India, pp. 1583-1589, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Varada Vivek Khanna et al., “A Distinctive Explainable Machine Learning Framework for the Detection of Polycystic Ovary Syndrome,” *Applied System Innovation*, vol. 6, no. 2, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]