A Comparative Study on Different Machine Learning Algorithms to Detect PCOS

Rithwick Sethi
Department of Electronics and Communications Engineering
Delhi Technological University
New Delhi, India
rithwick11111@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Delhi Technological University
New Delhi, India
dinesh@dtu.ac.in

Sreetam Ganguly
SystemOnSilicon Corporation
Mumbai, India
thomas.ganguly@systemonsilicon.com

Raj Ray
SystemOnSilicon Corporation
Mumbai, India
raj.ray@systemonsilicon.com

2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-3509-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCCNT56998.2023.10307174

***Abstract*—PCOS, or Polycystic Ovary Syndrome, is a frequent hormonal condition affecting women during their reproductive years. This disorder results in irregular and infrequent menstrual cycles, which can cause infertility and other related health issues. PCOS can present with various symptoms such as irregular periods, acne, obesity, and excessive hair growth, among others. However, some symptoms may not be visibly apparent and can only be detected through testing. This paper focuses on different techniques and algorithms and compares them to suggest which model is best suited to accurately classifying whether a woman has PCOS or not. The algorithms Random Forest Classifier (RFC), Support Vector Machine (SVM), XGBoost, and Ensemble Learning are applied on the basis of dimensionality reduction and Principal Component Analysis (PCA) to datasets available on Kaggle. This dataset consists of 43 attributes for 541 women, among whom 177 have PCOS. Based on the accuracy, it was found that the XGBoost Classifier performed better than the other models and ended up giving an overall accuracy of 89.63% before applying PCA and that RFC and SVC performed similarly and better than the other models and ended up giving an overall accuracy of 91.11% after applying PCA and retaining 99% of its variance.**

Keywords—PCOS, algorithms, Random Forest Classifier, Support Vector Machine, XGBoost, Ensemble Learning, Principal Component Analysis, models, accuracy, F1 Score.

I. Introduction

For a long, long time now females have been prone to a regular endocrine disorder called polycystic ovary syndrome or PCOS, caused by elevated levels of androgens (male sex hormones), leading to cysts in ovaries. Stein and Leventhal described this for the first time in 1935 [1, 2]. PCOS is most commonly found in reproductive-aged women, with a prevalence estimated to be between 2-18% and hence the leading contributor to infertility in women. Recent research has shown that PCOS has risk factors for breast cancer. It is highly heritable and complex with unfamiliar underlying genetic factors [3, 4]. PCOS is a potential lead to various disorders and shows several symptoms. In addition to symptoms such as irregular periods, excessive hair growth, and pain, PCOS is often associated with metabolic abnormalities like insulin resistance and high insulin levels, as well as mood disorders like depression [5]. It is very important that everyone is aware of this seemingly innocent disorder and all the possible ways this can be treated. We are in an era of women empowerment and if all females know how to detect PCOS and in fact any other disorders, it will be a big win for all of us. In this paper, we explore the use of different machine learning algorithms for PCOS detection and determine the most appropriate model that can delivery accurate and precise results, thereby contributing to the early diagnosis of this disease.

The objective of this study is to statistically evaluate the metabolic and clinical features, using dimensionality reduction and principal component analysis, and compare the accuracy and F1-Score of different machine learning models. The models being compared are:

1) Random Forest Classifier (RFC)
2) Support Vector Classifier (SVC)
3) XGradient Boost (XGBoost)
4) Ensemble Model combining RFC and XGB (RFCXGB)
5) Ensemble Model combining RFC and SVC (RFCSVC)
6) Ensemble Model combining SVC and XGB (SVCXGB)

These models are further worked on by applying dimensionality reduction, and Principal Component Analysis as per different percentages of variance retained. These are:

1) Without applying PCA
2) Applying PCA and 85% variance retained
3) Applying PCA and 90% variance retained
4) Applying PCA and 95% variance retained
5) Applying PCA and 99% variance retained

The paper is structured as follows: Section II covers related works, Section III explains the methodology used in this study, Section IV presents the results obtained from all the methods, and finally, Section V provides the conclusion.

II. Related Works

Due to their significance in today's society, among the countless issues that face us, those that pertain to women's health were chosen as the focus of our attention. This section shines a light on the literature work on PCOS. It discusses all similar works done to detect PCOS using machine learning algorithms. Table I summarizes all the methods used for similar research objectives and their respective results in terms of accuracies of the model used to detect PCOS successfully.

In the literature on PCOS, there has been limited research on the early detection of PCOS. A few of the best works have been discussed and are compared here. In [1], a novel

algorithm RFLR (hybrid random forest and logistic regression) has been applied to metabolical and clinical features to detect PCOS and provides an accuracy of 89.27%. Similarly, in [6, 7, 8], different known machine learning models help determine the presence of PCOS and develop smart systems to detect PCOS on similar metabolic and

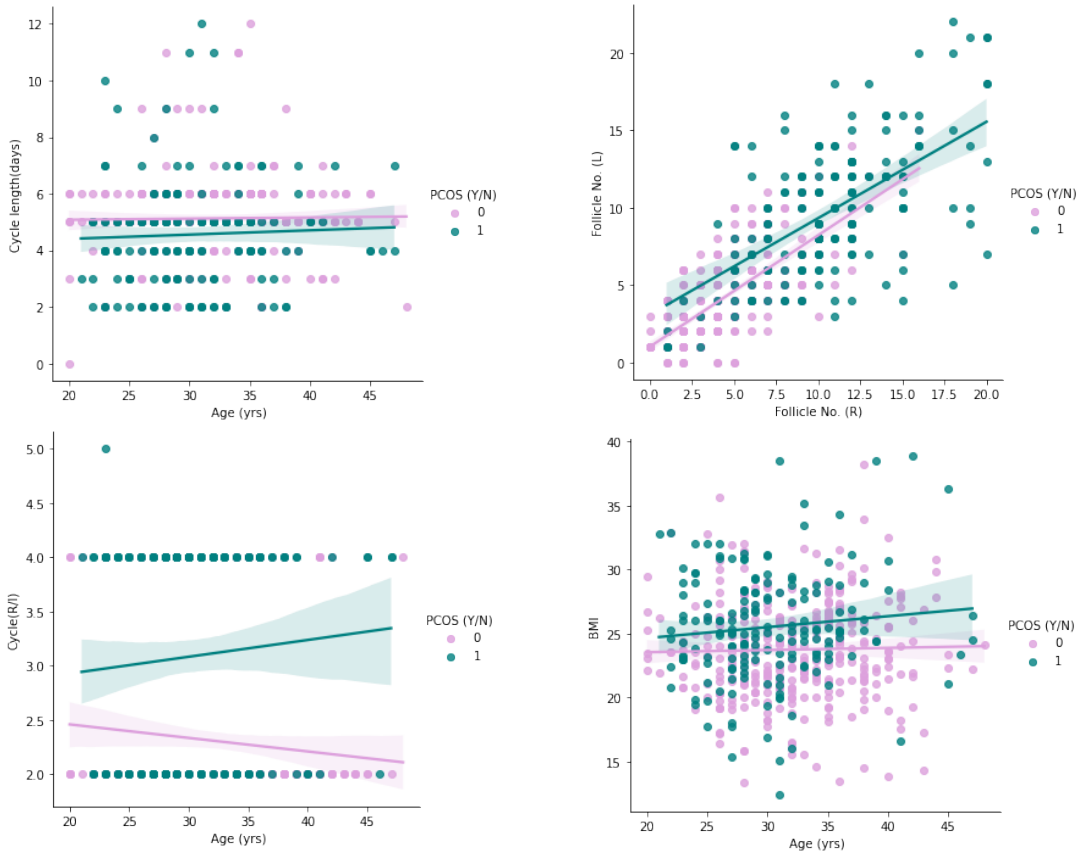

Figure 1: LMPlot used to understand the effect of (a) Length of Menstrual Phase, (b) Pattern of BMI over years, (c) Distribution of follicles in both ovaries, (d) Cycle of periods on PCOS

clinical features, namely, Logistic Regression, Bayesian Classifier, Linear Discriminant Analysis (LDA), Linear Support Vector Machines (SVMs), and Gaussian Naive Bayes with accuracies 91.04%, 93.93%, 90.70%, 87.10%, and 84.14% respectively. On the other hand, in [9], the segmentation of the follicles is done to determine the presence of PCOS using K-Nearest Neighbors (KNNs) with an accuracy of 80.73%. In [10], ovary ultrasonography (USG) scans are used to detect multiple cysts to diagnose PCOS accurately. Therefore, in this study, a proposed machine learning classification method for predicting PCOS has been expanded where a Convolutional Neural Network (CNN) model was built but gave the least accuracy of 74.79%.

This paper mainly focuses on comparing the different techniques and algorithms to detect PCOS sooner and more accurately.

TABLE I. AN OVERVIEW OF LITERATURE WORK ON PCOS

Ref No.	Algorithm	Objective of Research	Best Accuracy
[1]	Hybrid Random Forest Logistic Regression (RFLR)	Determining PCOS	89.27%
[6]	K Nearest Neighbors (KNN)	PCOS Follicle detection through ultrasound images	80.73%
[7]	1. Logistic Regression **2. Bayesian Classifier**	**PCOS Detection**	91.04% **93.93%**
[8]	Gaussian Naïve Bayes	PCOS Detection and Prediction System	84.14%
[9]	Convolutional Neural Network (CNN) Model	Detection of PCOS over ultrasound images	74.79%
[10]	1. Linear Discriminant Analysis (LDA) 2. Linear Support Vector Machine (SVM)	Smart PCOS Detection System	87.10% 90.70%

III. METHODOLOGY

A brief overview of machine learning classifier research and experiments for PCOS prediction is explained in this section. All experiments are conducted in Google Colaboratory using Python as a machine-learning tool and numerous other libraries like Scikit-learn, Pandas, NumPy, etc. are used for the deployment of Python.

A. Data Selection and Preprocessing

For this, a dataset consisting of 539 women each having 43 attributes was taken from Kaggle's repository [11]. Out of these 539 instances, 363 involve women who do not suffer from PCOS, and the remaining 176 involve women who have PCOS. A comprehensive dataset has been compiled, which

encompasses a range of physical and clinical parameters relevant to the diagnosis of PCOS and infertility-related conditions. This data has been collected from ten distinct healthcare facilities situated throughout the state of Kerala, India.

After selecting the appropriate data, data cleaning and preprocessing are performed. Data is labeled properly, encoding of non-numeric features is performed and all null values are accounted for. For normalization of the data, the MinMax scaler is used as follows:

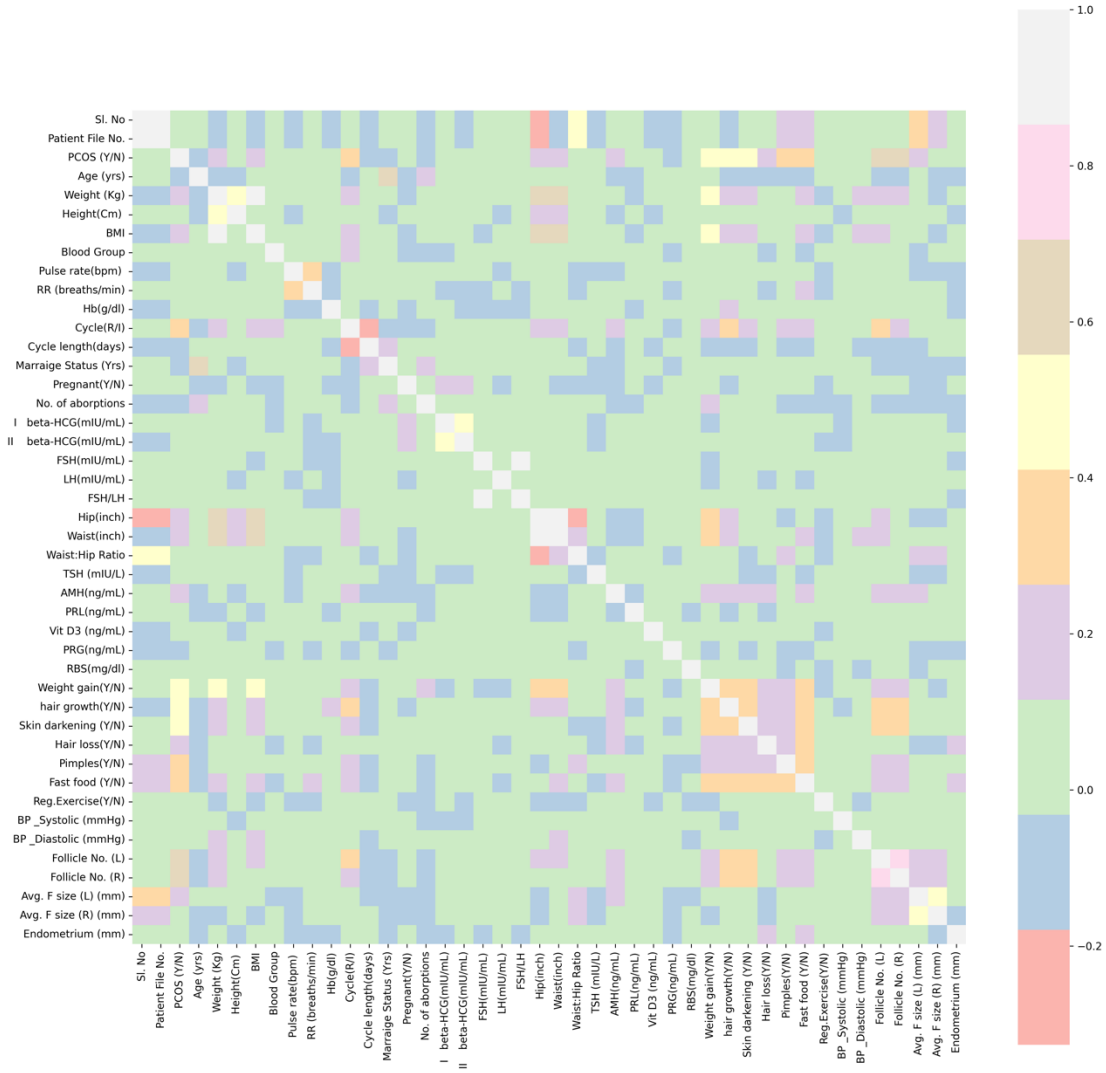

Figure 2: Heatmap consisting of 33 positive attributes and 10 negative attributes

$$p' = \frac{p - p_{min}}{p_{max} - p_{min}} \quad (1)$$

p = original feature value
p' = normalized value of p
p_{min} = minimum value of p
p_{max} = maximum value of p

The MinMax scaler scales and confines each feature into a specific range. We performed a 75-25 train-test split to conduct all experiments.

B. *Feature Analysis*

The process of determining and choosing important features from a multitude of features is referred to as feature selection. In machine learning, feature selection serves primarily as a means of constructing a subset of useful information to train the classifier and make it more understandable. This helps to analyze the data better and produce an improved learning process. Throughout the feature selection step, key variables are chosen to obtain strong predictive accuracy for the model [12].

Initially, to understand the specifics of the features, we plotted a few of the features to compare how important they are and to understand the data better. Fig. 1 shows the data visualization behind the features and provides an insight to understand the effect of different features on PCOS.

For further analysis and more in-depth understanding of the entire data and relevance of each feature, a heatmap is plotted and each feature's correlation coefficient is calculated and added to determine which feature can be neglected for better accuracy and F1-score, as shown in Fig. 2.

The Pearson correlation coefficient is computed using the formula:

$$r = \frac{\sum(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum(A_i - A)^2 \sum(B_i - B)^2}} \quad (2)$$

r = Pearson's coefficient of correlation
A_i = A-variable's values in a sample
$\bar{A}$ = mean of A-variable's values
B_i = B-variable's values in a sample
$\bar{B}$ = mean of B-variable's values

The correlation coefficients of each feature is taken and arranged in ascending order. According to this list, the optimal features are chosen based on 'r' values. The parameters BMI, Age, Left, and Right Follicles, Systolic Blood Pressure, Fasting Blood Sugar, Diastolic Blood Pressure, Post Prandial Blood Sugar, and Cycle Length have high values of 'r' and are amongst the best features taken for input patients.

C. Machine Learning Classifiers

- **Random Forest Classifier**: An ensemble of traditional decision tree classifiers makes up this effective machine learning algorithm. The training subsets for each tree are created using a bootstrap bagging technique. The classifier's output is typically determined by the majority voting method for each tree and is taken into account as a cumulative judgment of all trees [13].
- **Support Vector Machine**: The supervised machine learning approach SVM can be used for both Classification and Regression Problems. Datasets are divided into classes by SVMs so that a maximally marginal hyperplane can be found. The kernel method is used to transform the data and depending on the modifications, it creates an optimum border between the possible outputs. The kernel is a collection of mathematical operations utilized in SVM methods [14].
- **XGBoost Classifier**: Extreme Gradient Boosting or XGBoost is an accurate and scalable gradient boosting ensemble machine learning technique based on decision trees, suitable for classification and regression predictive modeling problems. It uses the stochastic gradient boosting method quickly and with a variety of hyperparameters for fine-grained control during model training [15].
- **Ensemble Learning**: Ensemble learning represents a machine learning methodology that enhances classification accuracy through the integration of multiple base classifiers. This approach departs from conventional machine learning techniques by building multiple hypotheses, which are then combined, as opposed to solely learning a single hypothesis from the training data [16].

D. Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate technique that analyzes a data table with interrelated dependent variables. The most significant information is extracted from the data and presented as a set of new orthogonal variables known as principal components, which are displayed on maps to illustrate the patterns of similarity between variables and observations [17]. PCA objectives are to

- extract crucial information from the data
- compress the data's size by retaining only the most significant data
- simplify the description of the dataset
- examine the arrangement of variables and the observations

For this paper, we have applied PCA to these features to see its effect on the accuracy of each of the models. The number of new principal components is dependent on how much variance is retained by the data after dataset compression.

First, we find the 'Covariance Matrix' or 'Sigma', using the formula:

$$Sigma = \frac{1}{m} * (\sum_{i=1}^{n} (x^{(i)})(x^{(i)})^T) \quad (3)$$

m = total no. of samples
$x^{(i)}$ = ith original data sample

After, finding the eigenvectors of Sigma, Z is calculated through the formula:

$$Z = U_{reduce}^T * X \quad (4)$$

U_{reduce}^T = U matrix of size (k x n)
X = Input vector of size (n x 1)

After applying PCA, the total variation V is calculated for the data, using the formula:

$$V = (\frac{1}{m} * \sum_{i=1}^{m} ||x^{(i)}||^2) \quad (5)$$

Then, the average squared projection error in the data is calculated, through the formula:

$$\varepsilon = \frac{1}{m} * \sum_{i=1}^{m} ||x^{(i)} - x_{approx}^{(i)}||^2 \quad (6)$$

$x_{approx}^{(i)}$ = projected data

Using equations (5) and (6), check for v:

$$\frac{V}{\varepsilon} = \frac{\frac{1}{m} * \sum_{i=1}^{m} ||x^{(i)} - x_{approx}^{(i)}||^2}{\frac{1}{m} * \sum_{i=1}^{m} ||x^{(i)}||^2} \leq \frac{v}{100} \quad (7)$$

v = variance retained

To determine the number of principal components for a given percentage of variance retained, Algorithm 1 is used.

Algorithm 1. Attain the number of principal components

Input: Percentage of variance retained
Output: Number of principal components
Process:

1. Calculate the shape of the dataset and store it in the form [m,n].
2. Assign U and S to an array of zeroes having length n
3. Calculate Sigma.
4. Find the 'Singular Value Decomposition' or 'svd' of Sigma and store it in the form [U, S, V] to find the eigenvectors of Sigma.
5. Take the first 'k' columns of U and assign them to U_{reduce}
6. Calculate Z
7. Try PCA with k = 1
8. Find the total variation in data.
9. Calculate the average squared projection error.

10. If equation (7) holds true:
    Increment the value of k.
11. Repeat steps 7-11 till ideal value of v is achieved.

TABLE II. THE NUMBER OF PRINCIPAL COMPONENTS W.R.T. VARIANCE RETAINED

S.No.	Variance Retained (v)	No. of Principal Components (k)
1.	85%	4
2.	88%	5
3.	90%	6
4.	91%	7
5.	93%	8
6.	94%	9
7.	95%	10
8.	96%	11
9.	97%	13
10.	98%	16
11.	99%	22

IV. RESULTS AND DISCUSSIONS

In this section, the findings obtained from applying various machine learning algorithms on the PCOS dataset are reported and discussed. This research evaluated multiple machine learning models for PCOS detection and aimed to identify the most accurate model while examining the impact of PCA on each of these models. The number of principal components was established using Algorithm 1, based on the retained percentage of variance. The data was compressed from 43 features to each of these numbers of principal components and the accuracies were compared for each of the six models discussed in section I.

First, let's compare the accuracies of all models without performing PCA.

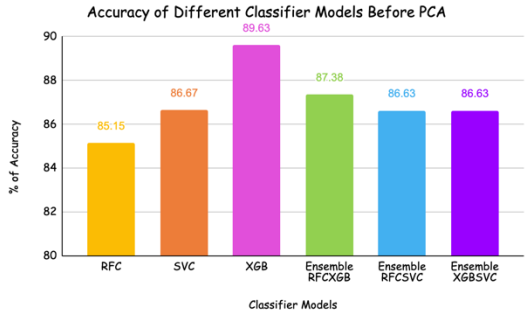

Figure 3: Accuracy of different classifier models before PCA

Fig. 3 clearly shows that the XGBoost outperforms the RFC, SVC, Ensemble RFCXGB, Ensemble RFCSVC, and Ensemble SVCXGB with an accuracy of 89.63% and an F1-score of 0.93. Here the RFC performs the worst with an accuracy of 85.19% and an F1-score of 0.90.

Now, let's compare all the models after applying PCA with varying numbers of principal component analysis.

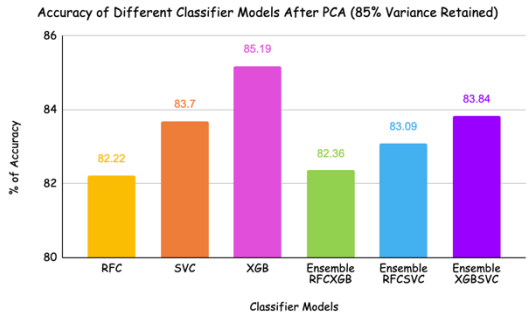

Figure 4: Accuracy of different classifier models after PCA (85% variance retained)

Fig. 4, clearly shows that when 85% variance is retained, or the number of principal components is equal to 4, XGBoost again outperforms the RFC, SVC, Ensemble RFCXGB, Ensemble RFCSVC, and Ensemble SVCXGB with an accuracy of 85.19%. Here the RFC again performs the worst with an accuracy of 82.22%.

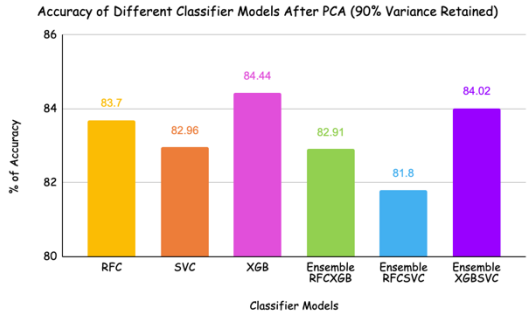

Figure 5: Accuracy of different classifier models after PCA (90% variance retained)

Fig. 5, clearly shows that when 90% variance is retained, or the number of principal components is equal to 6, XGBoost again outperforms the RFC, SVC, Ensemble RFCXGB, Ensemble RFCSVC, and Ensemble SVCXGB with an accuracy of 84.44%. Here the Ensemble RFCSVC performs the worst with an accuracy of 81.80%.

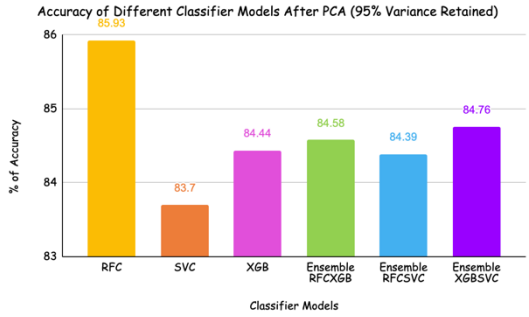

Figure 6: Accuracy of different classifier models after PCA (95% variance retained)

Fig. 6, clearly shows that when 95% variance is retained, or the number of principal components is equal to 10, RFC outperforms the SVC, XGBoost, Ensemble RFCXGB, Ensemble RFCSVC, and Ensemble SVCXGB with an accuracy of 85.93%. Here the SVC performs the worst with an accuracy of 83.70%.

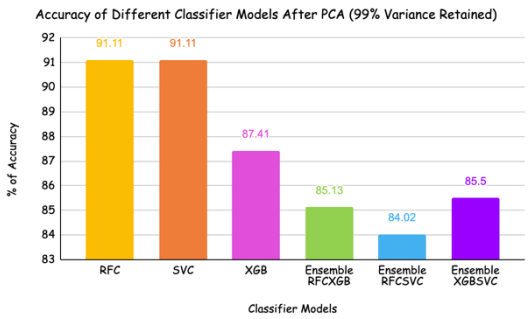

Figure 7: Accuracy of different classifier models after PCA (99% variance retained)

Fig. 7, clearly shows that when 99% variance is retained, or the number of principal components is equal to 22, RFC and SVC outperform the XGBoost, Ensemble RFCXGB, Ensemble RFCSVC, and Ensemble SVCXGB with an accuracy of 91.11%. Here the Ensemble RFCSVC performs the worst with an accuracy of 84.02%. Table III summarizes the above results. We can also determine which model performs the best with and without PCA.

From the table below, we can conclude that overall, the best models are RFC and SVC both with an accuracy of 91.11% and they are obtained after performing PCA and retaining 99% of the original data. This confirms two things for us. First, Principal Component Analysis does improve the accuracy of our models, as shown by this research as compared to not applying PCA at all. Second, XGBoost Classifier performs the best when no PCA is applied, and RFC and SVC perform similarly and better than the rest when PCA is applied with 99% of its variance retained.

Comparing our results with the ones in section II, we can clearly see that our model outperforms seven of the models and is definitely comparable to the Bayesian Classifier model with it leading just by an accuracy difference of 2.82%.

TABLE III. COMPARISON BETWEEN ALL MODELS ON THE BASIS OF PCA EVALUATED BY ACCURACY

PCA	% Variance Retained	Accuracy (%)					
No		RFC	SVC	XGB	Ensemble RFCXGB	Ensemble RFCSVC	Ensemble SVCXGB
		85.15	86.67	89.63	87.38	86.63	86.63
Yes	85	82.22	83.7	85.19	82.36	83.09	83.84
	90	83.7	82.96	84.44	82.91	81.8	84.02
	95	85.93	83.7	84.44	84.58	84.39	84.76
	99	**91.11**	**91.11**	87.41	85.13	84.02	85.5

V. CONCLUSION

The objective of this study is to identify PCOS in women as early as possible using the best of the different machine learning algorithms. To do so, the concepts of machine learning principal component analysis are applied. A dataset consisting of 539 patient records was obtained from Kaggle's data repository. According to our results, PCA was found to enhance the accuracy of selected models, and upon retaining 99% variance of the entire data, Random Forest Classifier and Support Vector Classifier perform quite well with an accuracy of 91.11%. If PCA is not applied, the XGBoost Classifier yields an accuracy of 89.63%, which is comparable to the accuracy of other models and methods discussed in Section II. This work could be used by doctors and in fact women suffering from PCOS themselves to identify patients who are more likely to develop PCOS through early screening.

The findings of this research can be validated and further compared by choosing only the visible features so that all women can self-detect PCOS accurately and do not have to be dependent on medical tests or wait till it is too late.

REFERENCES

[1] P. P. a. M. R. H. M. S. Bharati, "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms," in *IEEE Region 10 Symposium (TENSYMP),*, Dhaka, Bangladesh, 2020.

[2] R. Azziz, E. Carmina, D. D. E. Diamanti-Kandarakis, H. F.Escobar-Morreale, W. Futterweit, O. E. Janssen, R. S. Legro, R. J. Norman, A. E. Taylor and S. F. Witchel, "The androgen excess and PCOS society criteria for the polycystic ovary syndrome: The Complete Task Force Report," Fertility and Sterility, vol. 91, no. 2, p. 456–488, 2009.

[3] J. Kim, J. Mersereau, N. Khankari and e. al, "Polycystic ovarian syndrome (PCOS), related symptoms/sequelae, and breast cancer risk in a population-based case–control study," *Cancer Causes Control,* vol. 27, pp. 403-414, 2016.

[4] B. Zehra and K. AA, "Polycystic ovarian syndrome : Symptoms, treatment and diagnosis: A review," *Journal of Pharmacognosy and Phytochemistry,* vol. 7, no. 6, pp. 875-880, 2018.

[5] R. Pasquali, E. Sten-Victorin, B. O. Yildiz, A. J. Duleba, K. Hoeger, H. Mason, R. Homburg, T. Hickey, S. Franks, J. S. Tapanainen, A. Balen, D. H. Abbott, E. Diamanti-Kandarakis and R. S. Legro, "PCOS Forum: research in polycystic ovary syndrome today and tomorrow," *Clinical Endocrinology,* vol. 74, no. 4, p. 424–433, 2011.

[6] J. C. C. C. B. G. a. S. G. P. Mehrotra, "Automated screening of Polycystic Ovary Syndrome using machine learning techniques," in *Annual IEEE India Conference*, Hyderabad, 2011.

[7] A. R. A. A. C. M. R. a. R. G. A. Denny, "i-HOPE:Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," in *TENCON 2019 - 2019 IEEE Region*, Kochi,India, 2019.

[8] L. K. D. K. A. J. A. A. K. P. A. F. A. a. S. A. A. S. Tiwari, "SPOSDS: A smart polycystic ovary syndrome diagnostic system using machine learning," *Expert Systems with Applications,* vol. 203, p. 117592, 2022.

[9] U. N. W. A. F. N. B. Purnama and A. G. a. T. M. , "A classification of polycystic ovary syndrome based on follicle detection of ultrasound images," in *3rd*

International Conference on (IEEE) Information and Communication Technology (ICoICT), 2015.

[10] S. I. M. Suha, " An extended machine learning technic polycystic ovary syndrome detection using ovary ultra image," *Scientific Reports,* vol. 12, no. 1, p. 17123, 2(

[11] P. Kottarathil, "Polycystic ovary syndrome (PCOS)," Kaggle, 11 July 2020. [Online]. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos. [Accessed 29 october 2022].

[12] F. Z. L. &. A. N. Thabtah, "NBA game result prediction using feature analysis and machine learning," *Annals of Data Science,* vol. 6, no. 1, pp. 103-116, 2019.

[13] S. D. a. S. C. S. S. Roy, "Autocorrelation Aided Random Forest Classifier-Based Bearing Fault Detection Framework," *IEEE Sensors Journal,* vol. 20, no. 18, pp. 10792-10800, 2020.

[14] A. J. N. M. A. C. L. M. A. A. K. D. S. A. H. S. Tiwari, ""“Machine learning based model for prediction of power consumption in smart grid Smart way towards smart city," *Expert Systems,* vol. 39, no. 5, p. 12832, 2021.

[15] Z. L. C. S. Y. H. X. D. R. Z. X. L. Y. Y. S. X. Daping Yu, "Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier," *Thoracic cancer,* vol. 11, no. 1, pp. 95-102, 2020.

[16] G. S. J. M. J. X. K. &. G. J. Wang, " Sentiment classification: The contribution of ensemble learning," *Decision Support Systems,* vol. 57, pp. 77-93, 2014.

[17] H. &. W. L. J. Abdi, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics,* vol. 2, no. 4, pp. 433-459, 2010.