Conference Paper Title*

*Note: Sub-titles are not captured for https://ieeexplore.ieee.org and should not be used

Manasvi
Dept. of Computer Science and Engineering
Indira Gandhi Delhi Technical University for Women
manasvi110btcse23@igdtuw.ac.in

Kajal Singh
Dept. of Computer Science and Engineering
Indira Gandhi Delhi Technical University for Women
kajal082btcse23@igdtuw.ac.in

***Abstract*—PCOS(Polycystic ovarian syndrome) a multifactorial endocrine disorder. It is characterized by Irregular periods, hirsutism, weight gain. Risk factors contributing to PCOS include genetics, neuroendocrine system, sedentary lifestyle, diet, and obesity. Worldwide range of PCOS is 6-36%, and in India it lies between 9.13-36%. PCOS is a prime cause of infertility due to its prevalence in females.[1] In this paper we examine different predictive methods for PCOS in fertile females and conclude with the recommendation of a highly accurate method. The required dataset includes features such as length and irregularity of cycle, No. of abortions., Pregnancy, LSH,FSH,TH,PRL and PRG levels, RBS and Hb levels . The models implemented are Logistic Regression accurate to 87.33% , Support Vector Machine accurate to 87.11% and Random Forest accurate to 90.18%. Hence, Random Forest offers the highest accuracy.**

Keywords: PCOS, Logistic Regression, Support Vector Machine, Random Forest, Machine Learning, Fertility Prediction

I. Introduction

Polycystic Ovary Syndrome (PCOS) is a hormonal imbalance condition mainly found in women of reproductive age. The occurrence of PCOS has skyrocketed due to Adulterated foods and disturbed lifestyle. PCOS distresses almost 1 out of every 10 women in their conceptive age.[2] PCOS hinders the quality of life as it affects various domains of health and causes cardiovascular diseases, failure to ovulate and infertility, late menopause, type 2 diabetes, acne, baldness, hair loss, hirsutism, obesity, anxiety, depression, and stress.[3] These long-term problems can be prevented or treated if detected well in time. Various PCOS detection methods include ultrasonography and blood test for androgen levels. The tests must be performed by an expert and the analysis of reports is time taking. In this paper we examine biochemical parameters such as hormone and sugar levels, follicle size and number and Body Mass Index to predict the presence of PCOS. The paper evaluates the reliability of this method of analysis of biochemical parameters over the conventional tests for PCOS. 35+ features and parameters have been utilised to build the predictive method with an accuracy of 90%.

II. LITERATURE REVIEW

A study on PCOS and infertility by Hart, R. (2008), suggests that Women with PCOS have higher risk of miscarriage and in pregnancy. They are at an increased risk of developing gestational diabetes, pregnancy-induced hypertension [6]. Koskinen, P., Penttilä, T. A., Anttila, L., Erkkola, R., & Irjala, K. (1996), published their research on Optimal use of hormone determinations in the biochemical diagnosis of the polycystic ovary syndrome stating that simultaneous analysis of LH,FSH levels could be effectively used for PCOS predictions.[7] Research by Agrawal, A., Ambad, R., Lahoti, R., Muley, P., & Pande, P. S. (2022), examines the use of AI in detection of PCOS. It uses different AI techniques to process ultrasound images and detect PCOS in its early stages.[8] Previously, many models have been implemented for the prediction of PCOS. Kumar, D., & Kumar, A. (2023), employed ADA Boost and dimensionality reduction algorithms for prediction of PCOS. They were able to achieve an accuracy of 84% on their dataset outperforming other models (Random Forest ,Logistic Regression) tested over the same dataset.[5] Dutta, P., Paul, S.,& Majumder, M. (2021), used Synthetic Minority Over Sampling Technique (SMOTE) before utilising the dataset. Outstanding performance was achieved by SMOTE based Logistic Regression. SMOTE based Random Forest showed exceptionally less time of seconds.[3]

III. METHODOLOGY

A. Dataset Visualization

The dataset used consists of data collected from 541 women across 10 hospitals in Kerala, India .[2]The dataset consists of a total of 41 features. Unnecessary columns such as 'Sl.No' and 'Patient file NO.' were dropped.

The pie chart plotted using the 'matplotlib' library in Python visualizes the dataset, showing that 177 women are not affected while 364 are diagnosed as diseased.

B. FEATURE EXTRACTION

The 41 features utilised for prediction include:
1. PCOS (Y/N) [TARGET]
2. Age (yrs)
3. Weight (Kg)
4. Height(Cm)
5. BMI
6. Blood Group
7. Pulse rate(bpm)
8. RR (breaths/min)

Identify applicable funding agency here. If none, delete this.

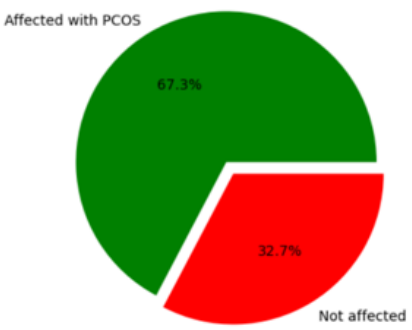

Fig. 1. PCOS IN FERTILE FEMALES

9. Hb(g/dl)
10. Cycle(R/I)
11. Cycle length(days)
12. Marraige Status (Yrs)
13. Pregnant(Y/N)
14. No. of aborptions
15. I beta-HCG(mIU/mL)
16. II beta-HCG(mIU/mL)
17. FSH(mIU/mL)
18. LH(mIU/mL)
19. FSH/LH
20. Hip(inch)
21. Waist(inch)
22. Waist:Hip Ratio
23. TSH (mIU/L)
24. AMH(ng/mL)
25. PRL(ng/mL)
26. Vit D3 (ng/mL)
27. PRG(ng/mL)
28. RBS(mg/dl)
29. Weight gain(Y/N)
30. hair growth(Y/N)
31. Skin darkening (Y/N)
32. Hair loss(Y/N)
33. Pimples(Y/N)
34. Fast food (Y/N)
35. Reg.Exercise(Y/N)
36. BP _Systolic (mmHg)
37. BP _Diastolic (mmHg)
38. Follicle No. (L)
39. Follicle No. (R)
40. Avg. F size (L) (mm)
41. Avg. F size (R) (mm)
42. Endometrium (mm)

C. DATA PREPROCESSING

1) IDENTIFYING CATEGORICAL COLUMNS: To prepare the dataset for further analysis involved identification of all categorical columns. Categorical variables are used to group qualitative outcomes. Features containing categorical data are represented as strings or objects in the dataset. The code used is: categorical_columns = data.select_dtypes(include=['object']).columns

Method select_datatypes selects the categorical columns. The feature names extracted by .columns attribute are stored in categorical_columns.

2) LABEL ENCODING: To make the dataset fit for use as training and testing data, it was required to convert all non-numeric values into a format Machine Learning models could process. We import LabelEncoder from sklearn.

from sklearn.preprocessing import LabelEncoder

LabelEncoder converts non-numeric values to unique integer values. We initialise a dictionary label_encoders to store the encoders for each column. An instance of class LabelEncoder is created for every column in categorical_columns.

.fit_transform method converts each category to its unique integer and stores the transformed data into original column. Each value is converted to string using .astype(str) before encoding to ensure compatibility of data. The labelEncoder object for each column is stored in label_encoders for future reference.

```
label_encoders = {}
for column in categorical_columns:
le = LabelEncoder()
data[column] = le.fit_transform(data[column].astype(str))
label_encoders[column] = le
```

```
print("label_encoders= ",label_encoders)
label_encoders=  {'II    beta-HCG(mIU/mL)': LabelEncoder(), 'AMH(ng/mL)': LabelEncoder(), 'Unnamed: 42': LabelEncoder()}
```

3) IMPUTATION: Features ' FSH/LH' and 'Waist:Hip Ratio' contain 532 null values. Null values or Missing values interfere with estimations as Machine learning models assume that all values are numerical. To deal with Null Values we impute the Null Values with the mean of non-missing values in the column. For this, we instantiate an object of class SimpleImpuer and specify the strategy as mean.

.fit_transform() method computes the mean and replaces all missing values with mean.

The imputer is applied to X and the result is stored back to X.

```
imputer = SimpleImputer(strategy='mean')
X = imputer.fit_transform(X)
```

4) DATA STANDARDISATION: The dataset contains values in different units mm/hg, mm, cm, kg and more. Some features represented as Y/N contain 0/1 values. The dataset contains broad ranged values which leads to wide ranged values dominating the predictions. To convert all values to comparable scales we standardise it. Standardisation transforms the data

to consistent, standard format which increases the accuracy and speed of Machine Learning models.
We employ standard scaler class provided by sklearn to process the dataset. An instance of class StandardScaler is loaded into variable scaler. Applied alongwith the .fit_transform method it standardises the data and stores the data back to original dataframe.
scaler = StandardScaler() X_train s̄caler.fit_transform(X_train) X_test = scaler.transform(X_test)

5) SPLITTING THE DATA: The dataset is divided into two parts.
• Training Data – that will be used to train the machine learning model
• Testing data- that will be used to test the accuracy of our prediction
The size of test data is 30% of the total dataset.
We employ the train_test_split function for the task.
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

6) MODEL SELECTION: We predict the existence of PCOS in fertile women, hence we require efficient binary classification model. Three models were implemented Logistic Regression, Random Forest Classifier and SVM Classifier. Logistic Regression utilises a logistic function to predict values. It gives each feature equal weightage and can be optimised by hyperparameter tuning. SVM can handle complex non linear relationships easily. The margin between classes can be maximized by optimisation. Random Forest can easily handle complex relationships. More Influential are prioritised.

7) MODEL IMPLEMENTATION:

- LOGISTIC REGRESSION
  Logistic Regression is a binary classification model which predicts the possibility of an input belonging to a particular class. It implements a logistic function to output a probability between 0 and 1, which are further approximated as binary predictions.
  Model training : The logistic regression model is initialised with max_iter=1000. This ensures sufficient number of iterations for efficient optimisation. This way, we cover inputs that require higher iterations to reach the optimal solution.
  To train the model we use the fit method. The training data X_train is then passed to the model. The model reiterates over the data and tunes its parameters accordingly.
- SVM
  SVM optimises linear discriminant models based on the perpendicular distance between classes.
  The dataset we use is linearly separable, hence SVM model is initialised with a linear kernel. Specifying Random state ensures reproducibility.
  Model Training: SVM classifier is trained on training data X_train using fit method.
  The model iterates to find the hyperplane that best separates the two classes, maximizing the margin.
- RANDOM FOREST
  Fast Tree learning model Random Forest predicts classes by averaging results from multiple trees. An instance of random forest classifier with 100 trees is loaded into model.
  Model training Using the fit method we pass the training data X_train to the model. The model creates multiple trees and gives the combined output.

IV. RESULTS

The accuracy obtained using different models is shown below:

- **Logistic Regression**

Accuracy: 0.8711656441717791
Classification Report:

	precision	recall	f1-score	support
0	0.87	0.95	0.91	110
1	0.86	0.72	0.78	53
accuracy			0.87	163
macro avg	0.87	0.83	0.85	163
weighted avg	0.87	0.87	0.87	163

Fig. 2. Accuracy of Logistic Regression

- **SVM**

Accuracy: 0.8773006134969326
Classification Report:

	precision	recall	f1-score	support
0	0.88	0.95	0.91	110
1	0.87	0.74	0.80	53
accuracy			0.88	163
macro avg	0.87	0.84	0.85	163
weighted avg	0.88	0.88	0.87	163

Fig. 3. Accuracy of SVM

- **Random Forest**

Accuracy: 0.901840490797546
Classification Report:

	precision	recall	f1-score	support
0	0.91	0.95	0.93	110
1	0.89	0.79	0.84	53
accuracy			0.90	163
macro avg	0.90	0.87	0.88	163
weighted avg	0.90	0.90	0.90	163

Fig. 4. Accuracy of Random Forest

It is evident that Random Forest is the most accurate model with an accuracy of 0.9018 .
Also, Random Forest gives the most precise results.

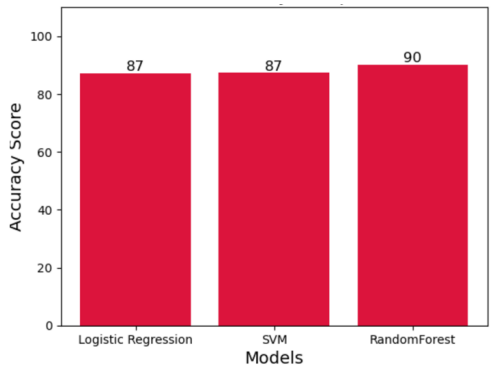

Fig. 5. Model Accuracy Comparision

V. DISCUSSION

The results obtained in the study shows the performance of different classification models in predicting PCOS. Minimum accuracy observed is 87.11% for logistic regression. The difference between accuracy for Logistic Regression(87.11%) and SVM classifier(87.73%) is almost negligible. This suggests that the features used for prediction are such that they are equally suited for linear as well as complex decision boundaries. Random Forest gives 90.11% accuracy which shows that the model was able to capture more complex relationships with its multiple decision tress. For a dataset with multiple features and wide interactions Random Forest is a good choice. We see that in previous literature [4] Data is not processed before training. The inclusion of data preprocessing steps Standardization, Imputation, and label encoding improves the performance of predictive models. It is also seen that random Forest performs best on different datasets as well[5]. Machine Learning Based systems for early detection of PCOS can prove to be of great clinical significance. Prevention of long-term problems due to PCOS can be achieved with timely detection[5].

Using multiple algorithms and more precise datasets coupled with effective hyperparameter tuning can improve the results. The use of 41 features in our dataset also highlights the possibility of presence of some irrelevant features. Future research areas include development of hybrid models and generalisation to all types of women. The evaluation of bias towards features forms an important area of future research. The development of this system as a real- time clinical detection system provides a great advantage to the society.

VI. CONCLUSION

The paper evaluates the performance of SVM, Logistic Regression, and Random Forest. Random Forest stands superior to all other models with an accuracy of 90.11%. The inclusion of data preprocessing makes the system more effective. Development of a generalised PCOS prediction system can help with early detection and treatment of the disease. It ensures improved diagnostic accuracy and enhanced decision making in real time. Overall, the study highlights the importance of machine learning in an important healthcare application and tries to develop the foundation for a well developed biochemical factor based PCOS detection system.

REFERENCES

[1] Kodipalli, A., & Devi, S. (2021). Prediction of PCOS and mental health using fuzzy inference and SVM. Frontiers in public health, 9, 789569.

[2] Inan, M. S. K., Ulfath, R. E., Alam, F. I., Bappee, F. K., & Hasan, R. (2021, January). Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis. In 2021 IEEE 11th annual computing and communication workshop and conference (CCWC) (pp. 1046-1050). IEEE.

[3] Dutta, P., Paul, S., & Majumder, M. (2021). An efficient SMOTE based machine learning classification for prediction & detection of PCOS.

[4] S. Nasim, M. S. Almutairi, K. Munir, A. Raza and F. Younas, "A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics," in IEEE Access, vol. 10, pp. 97610-97624, 2022, doi: 10.1109/ACCESS.2022.3205587.

[5] Kumar, D., & Kumar, A. (2023). PCOS Prediction Using Machine Learning Techniques. NEU Journal for Artificial Intelligence and Internet of Things, 1(2).

[6] Hart, R. (2008). PCOS and infertility. Panminerva medica, 50(4), 305-314.

[7] Koskinen, P., Penttilä, T. A., Anttila, L., Erkkola, R., & Irjala, K. (1996). Optimal use of hormone determinations in the biochemical diagnosis of the polycystic ovary syndrome. Fertility and sterility, 65(3), 517-522.

[8] Agrawal, A., Ambad, R., Lahoti, R., Muley, P., & Pande, P. S. (2022). Role of artificial intelligence in PCOS detection. Journal of Datta Meghe Institute of Medical Sciences University, 17(2), 491-494.