PCOS detection using Machine Learning Algorithms

Shameen Shrestha, Shreya Basnet

Abstract— **Polycystic ovary syndrome (PCOS), also known as polycystic ovarian syndrome, is a common health problem caused by an imbalance of reproductive hormones. Women with PCOS have a hormonal imbalance. Machine learning (ML) can deliver critical insight to clinicians at the point of decision making and replace manual processes, such as reviewing a patient's lab history. Machine learning is used to discover patterns from medical data sources and provide excellent capabilities to predict diseases. The purpose of this paper is to understand one of the most common health issues in women and put forward a solution to detect and predict the health problem based on various factors using different machine learning algorithms. We have explored and compared different algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, Random Forest and Naive Bayes for the prediction of the disease. It was found that Logistic regression gave the best result with accuracy of 88.89%.**

Key Words— **Decision Tree, Logistic Regression, Naïve Bayes, Polycystic Ovary Syndrome, Random Forest, Support Vector Classifier, Support Vector Machine,**

I. INTRODUCTION

Polycystic ovary syndrome (PCOS) is known as one of the most common disorders among reproductive-aged women, affecting 6%–20% of premenopausal women worldwide [1].According to a 2017 research at Kathmandu Diabetes and Thyroid Centre Pvt. Ltd. that included the clinical, biochemical, and hormonal profiles of PCOS patients, it was found that this is the most common gynecological disorder prevalent amongst females of this day and age[2]It is a syndrome where the ovaries create excessive levels of androgens(male sex hormones), which are typically present in women in tiny amounts. The hormones that play a role in PCOS are Androgens (like testosterone and androstenedione), Luteinizing hormone (LH), Follicle-stimulating hormone (FSH), Estrogen, Progesterone, and Insulin The most Common symptoms of PCOS irregular periods or no periods at all, difficulty getting pregnant, excessive hair growth (hirsutism) – usually on the face, chest, back or buttocks, weight gain thinning hair and hair loss from the head. Besides reproductive abnormalities, a number of metabolic conditions, including hepatic steatosis, glucose intolerance, dyslipidemia, diabetes mellitus type II (T2DM), and hypertension, are significantly linked to PCOS [3].

Many women do not even realize they have PCOS. In one study, up to 70 per cent of women with PCOS hadn't been diagnosed [4].Nobody has a treatment for it or a certain way to diagnose it. To be properly diagnosed women have to take innumerable medical tests which is a burden for both the patients and the doctors. Machine learning (ML) can take the place of manual operations like analyzing a patient's test results and provide clinicians with crucial knowledge at the time of decision-making. The goal of this research is to study one of the most prevalent health problems in women and to provide a method to identify and predict it using various machine learning algorithms.

II. LITERATURE REVIEW

One of the prime but treatable causes of infertility in women is polycystic ovary syndrome (PCOS). The prevalence of polycystic ovarian syndrome was reported to be 35 (9.18%) out of 381 individuals (undergraduate medical students in Nepal)[2] .The main diagnosis includes scanning follicles, their number and size using ultrasound imaging. PCOS symptoms differ in every patient. A lot of women have PCOS, but do not get diagnosed with it at an earlier stage. In a study, 69 to 70 percent of women did not have a pre-existing diagnosis [5]

Different methods have been used by researchers to detect PCOS early on. Palak et al. used Bayesian and Logistic Regression to automate PCOS based on clinical and metabolic markers and the Bayesian classifier had an accuracy of 93.93% [6]. A. L. Liu et. al used K-nearest neighbor (KNN), decision trees, SVMs with different kernels function to predict PCOS from identification of new genes [7]. Denny et al. found that the best and accurate model as the random Forest with 89% accuracy [8]. Subrato et al. used gradient boosting, random forest, logistic regression, and RFLR. Using 40-fold cross validation, the results also show that RFLR has the greatest testing accuracy (91.01%) and recall value (90%) of any method [9]. Pijush et al. used the novel The Synthetic Minority Oversampling Technique (SMOTE), which automates the early diagnosis of PCOS.They used five machine learning algorithms, including Logistic Regression, Random Forest, Decision Tree Support Vector Machine, and K-Nearest Neighbour (KNN) model. The best model achieved F1 score of 0.010sec, Recall of 98% and Precision of 98% [10].

III. METHODOLOGY

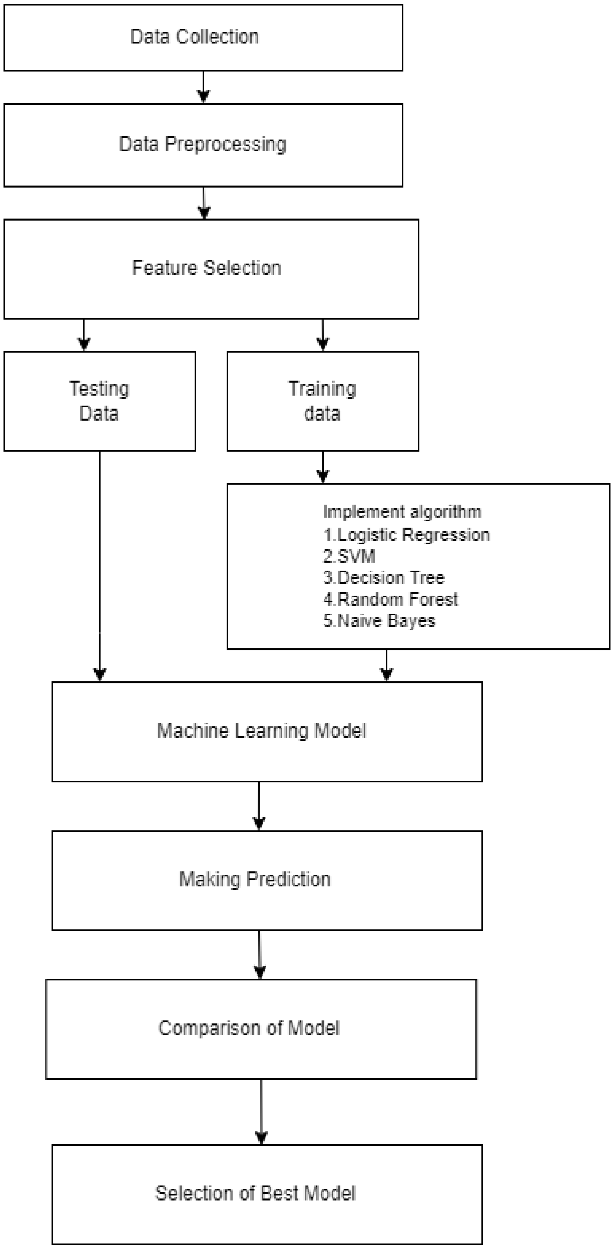

Fig. 1. Methodology of using different algorithms for PCOS detection

First the PCOS dataset is acquired from the Kaggle repository [11]. On studying the dataset, we found that there are 541 instances of data, and the dataset has 43 attributes. In that dataset we found that 364 are normal and 177 are PCOS affected patients.

The data acquired from the real world needs to be processed before it can be further analyzed. So following steps were to clean the data:

1. Finding if null values exist and if they exist replacing them with appropriate values.
2. Determining the outliers and replacing them with appropriate values. For example the prolactin blood level is found to be 25 µg/L in non-pregnant women and 80 to 400 µg/L in pregnant women [12]. Hence any value outside this range was replaced.

The next step is feature selection. For this we take some informative and significant features among a vast array of features to reduce overfitting and improve accuracy. This was done with the help of a correlation matrix.

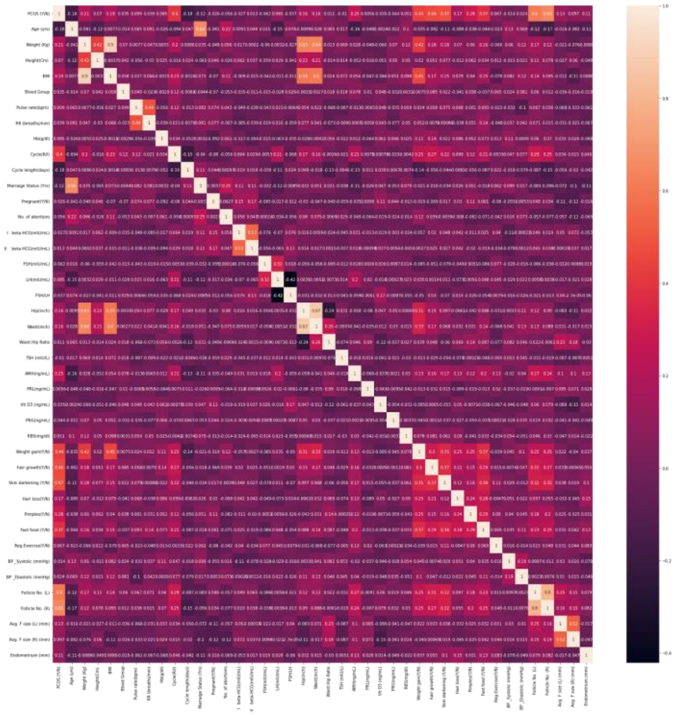

Fig. 2. Correlation matrix depicts the correlation between all the possible pairs of values in a table

The cleaned data was split into two parts i.e. testing dataset and training dataset.

The following algorithm were implemented to train the Machine learning model:

1.Logistics Regression

Logistic Regression is a supervised learning classification algorithm, it is a predictive analysis algorithm based on the concept of probability. Logistic Regression uses a more complex cost function, this cost function can be defined as the ‘Sigmoid function’ or also known as the ‘logistic function’.

Logistic function = 1/ 1 + e ^ -x

where,

e = base of natural logarithms

x = numerical value one wishes to transform

2. SVM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed a Support Vector Machine.

3. Decision Tree

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

This algorithm compares the values of the root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

4. Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. The random forest algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees.

5. Naive Bayes:

Naive Bayes is a simple supervised machine learning algorithm that uses the Bayes' theorem with strong independence assumptions between the features to procure results. That means that the algorithm just assumes that each input variable is independent.

$P(A|B) = P(B|A) * P(A) / P(B)$

Where

$P(A|B)$ = Posterior Probability
$P(B|A)$= Likelihood
$P(A)$= Class Prior Probability
$P(B)$= Prediction Prior Probability

Where the probability that we are interested in calculating P(A|B) is called the posterior probability and the marginal probability of the event P(A) is called the prior.

These trained models were applied on a test dataset and the cross-validation accuracy was found. Cross validation accuracy in machine learning means percentage of correct classification derived through the method of cross validation. This is usually done on a training dataset.

IV. Result

The experiment is carried out on the dataset created and machine learning algorithms were applied. The objective of using various algorithms is to identify the most suitable algorithms for classification of the dataset created. The Machine Learning algorithms like Logistic Regression,SVM, Decision Tree ,Random Forest and Naïve Bayes used for classification and performance is analyzed statistically. To compare the model confusion matrix and accuracy have been used. The result is summarized in the table below

Model	**Accuracy**
Logistic Regression	88.89
SVM	87.03
Decision Tree	81.48
Random Forest	82.40
Naive Bayes	85.18

Table 1. Accuracy of different algorithm

Model	**Precision** (class 0, class1)	**Recall** (class 0, class1)	**F1-Score** (class 0, class1)
Logistic Regression	(0.92,0.83)	(0.92,0.83)	(0.92,0.83)
SVM	(0.92,0.78)	(0.89,0.83)	(0.90,0.81)
Decision Tree	(0.84,0.76)	(0.90,0.63)	(0.87,0.69)
Random Forest	(0.87,0.81)	(0.92,0.71)	(0.89,0.76)
Naive Bayes	(0.94,0.72)	(0.84,0.89)	(0.88,0.79)

Table 2. F1 score table of different algorithm

As seen from the above tables the best accuracy of 88.89% and F1 score of (0.92,0.83) using Logistic Regression algorithm.

V. CONCLUSION

In this paper, various machine learning models were trained for detection of PCOS. We have applied Logistic Regression,SVM, Decision Tree ,Random Forest and Naïve Bayes Dataset, and did a lot of feature manipulation and extraction. We got the best Accuracy of 88.89% and F1 score of (0.92,0.83) using Logistic Regression algorithm.

REFERENCES

[1] H. F. Escobar-Morreale, “Polycystic ovary syndrome: Definition, aetiology, diagnosis and treatment,” *Nature Reviews Endocrinology*, vol. 14, no. 5. Nature Publishing Group, pp. 270–284, May 01, 2018. doi: 10.1038/nrendo.2018.24.

[2] M. ’Joshi Ansu, P. ’Yonzon, and S. ’ ’Tandukar, “Clinical profile of patients with polycystic ovarian syndrome in Nepal,” *Endocrinology & Metabolism International Journal (EMIJ)* , 2017.

[3] A. L. Liu *et al.*, “Association between fat mass and obesity associated (FTO) gene rs9939609 A/T polymorphism and polycystic ovary syndrome: A systematic review and meta-analysis,” *BMC Med Genet*, vol. 18, no. 1, Aug. 2017, doi: 10.1186/s12881-017-0452-1.

[4] W. A. March, V. M. Moore, K. J. Willson, D. I. W. Phillips, R. J. Norman, and M.J. Davies, “The prevalence of polycystic ovary syndrome in a community sample assessed under contrasting diagnostic criteria,” *Human Reproduction*, vol. 25, no. 2, pp. 544–551, 2010, doi: 10.1093/humrep/dep399.

[5] D. Dewailly *et al.*, “Definition and significance of polycystic ovarian morphology: A task force report from the androgen excess and polycystic ovary syndrome society,” *Hum Reprod Update*, vol. 20, no. 3, pp. 334–352, 2014, doi: 10.1093/humupd/dmt061.

[6] P. Mehrotra, J. Chatterjee, C. Chakraborty, and S. Ghoshdastidar, “Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques.”

[7] Z. Na, W. Guo, J. Song, D. Feng, Y. Fang, and D. Li, “Identification of novel candidate biomarkers and immune infiltration in polycystic ovary syndrome,” *J Ovarian Res*, vol. 15, no. 1, p. 80, Dec. 2022, doi: 10.1186/S13048-022-01013-0/FIGURES/7.

[8] A. Denny, A. Raj, A. Ashok, C. M. Ram, and R. George, “I-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques,” *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, vol. 2019-October, pp. 673–678, Oct. 2019, doi: 10.1109/TENCON.2019.8929674.

[9] S. Bharati, P. Podder, and M. R. Hossain Mondal, “Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms,” *undefined*, pp. 1486–1489, Jun. 2020, doi: 10.1109/TENSYMP50017.2020.9230932.

[10] P. Dutta, S. Paul, and M. Majumder, “An Ecient SMOTE Based Machine Learning classication for Prediction & Detection of PCOS,” 2021, doi: 10.21203/rs.3.rs-1043852/v1.

[11] “PCOS Dataset | Kaggle.” https://www.kaggle.com/datasets/shreyasvedpathak/pcos-dataset (accessed Oct. 01, 2022).

[12] “Prolactin blood test.” https://www.ucsfhealth.org/medical-tests/prolactin-blood-test (accessed Oct. 01, 2022).