PCOS Prediction Using Machine Learning Techniques

Deepak Kumar, Aman Kumar
Dept. of Computer Science Engineering, Chandigarh University, Mohali, India
19BCS1351@cuchd.in, 19BCS1498 @cuchd.in

Abstract—PCOS is a complex endocrine disorder that affects women of reproductive age. The pathophysiology of PCOS is multifactorial, involving insulin resistance, hyperandrogenism, and ovarian dysfunction. The etiology of PCOS is complex and multifactorial, involving both genetic and environmental factors. This hormonal disorder affects more than 5 to 10 % of ladies at the puberty age. Women who are diagnosed with PCOS experience reproductive psychological and hormonal imbalances leading to decreases in the levels of oestrogen and progesterone, which are female hormones. They also have increased levels of Androgen or testosterone which is male hormones. Normally in each monthly cycle, one of the follicles in the ovaries grows and ripens to deliver the egg called an ovum. Body hormones encourage this ovulation process. Due to hormonal imbalance in PCOS, sometimes these follicles do not release eggs. When the egg does not mature properly, ovulation does not occur so one does not menstruate regularly. There are many immature follicles in the ovaries which release male hormones. These immature follicles are called cysts. Well the exact cause of PCOS is not known. What is known is that ovaries produce abnormally high levels of Androgen. This excess Androgen or male hormone production has been linked to four conditions: Genes studies show that PCOS runs in families. Many genes will contribute to this condition, Insulin Resistance up to 70% of women with PCOS have insulin resistance which means that their cells cannot use insulin properly and obesity is the major cause of insulin resistance, Inflammation increased levels of inflammation in the body leads to higher Androgen levels being overweight also contributes to inflammation, Lifestyle and psychological conditions PCOS are also linked with stress, modern faulty, lifestyle anxiety and depression. The diagnosis of PCOS is based on the presence of two out of three criteria: hyperandrogenism, ovulatory dysfunction, and polycystic ovaries. Hyperandrogenism can be diagnosed based on clinical signs such as hirsutism, acne, and male-pattern baldness, or by laboratory tests such as free testosterone or dehydroepiandrosterone sulfate (DHEAS) levels. Ovulatory dysfunction can be diagnosed based on menstrual irregularities or by measuring serum progesterone levels during the luteal phase of the menstrual cycle.

Keywords—Machine learning, polycystic ovary syndrome

I. INTRODUCTION

Technology and humanity together hand in hand can make way towards better health care and services. Machine literacy is a subset of artificial intelligence, in which it provides the system with the capability to automatically learn and ameliorate without being programmed explicitly. It substantially focuses on developing algorithms that can pierce the datasets handed and use data for the literacy purposes of the network. operations of Machine Learning bring about huge metamorphosis in the health assiduity, which includes discovery, data vaticination, image recognition etc. Polycystic ovary pattern(PCOS), is one of the applicable, most current hormonal complaints seen among the women of travail age. This is a miscellaneous endocrine complaint which is largely prone to gravidity, anovulation, cardiovascular complaint, type 2 diabetes, rotundity etc. PCOS is a common condition detected in nearly 12- 21 of women of reproductive age and among them 70 remain undiagnosed. PCOS conditions can be treated to some extent by controlled drugs and bringing differences in lifestyle. This includes the treatment styles with capsules for birth control, diabetes, fertility,anti-androgen drugs and surveying procedures like ultrasound checkup. When similar interventions fail, invasive treatment procedures like surgical drilling of ovaries is also used for perfecting the ovulation capability of the ovary by reducing the manly hormone position. The etiology of PCOS is sustained by both insulin resistance and hyperandrogenism. Clinically it's characterized by reproductive, metabolic and cerebral features and represents a major health burden to women. opinion is recommended grounded on clinical or biochemical and radiological test results. PCOS is diagnosed by rejection of inapplicable symptoms or test results, substantially because of lack of knowledge of its complex patho- medium. The different symptoms of this condition force medical interpreters to call for a large number of clinical test results and gratuitous radiological imaging procedures. The early discovery and opinion of PCOS with minimum tests and imaging procedures is of utmost significance and of great significance as the condition directly leads to ovarian dysfunction with an

increased threat of confinement, gravidity or indeed gynecological cancer and internal agony for the cases due to destruction of time and plutocrat.

.

II. BACKGROUND STUDY

Polycystic Ovary Syndrome (PCOS) is a common hormonal disorder among women of reproductive age, affecting up to 10% of women globally. It is characterized by irregular periods, excess androgen production, and polycystic ovaries. PCOS is also associated with an increased risk of developing diabetes, heart disease, and infertility. Diagnosis of PCOS can be challenging due to the varied presentation of symptoms and lack of a definitive diagnostic test. Therefore, there is a need for accurate and reliable methods for predicting PCOS. Machine learning algorithms have shown promise in predicting PCOS by identifying patterns and relationships in data that may not be easily apparent to human observers. In this model, machine learning algorithms were used to predict PCOS based on various clinical features such as age, BMI, glucose, and insulin levels. The data was preprocessed using standardization and dimensionality reduction techniques such as PCA. Two different classification models were used, namely logistic regression and random forest classifiers. The models were evaluated based on accuracy, mean absolute error, mean squared error, root mean squared error, and R-squared.

The results showed that the random forest classifier outperformed the logistic regression model, achieving an accuracy of 0.825. However, the accuracy was further improved to 0.844 by using an AdaBoost classifier. The performance evaluation metrics revealed that the AdaBoost classifier had lower mean absolute error, mean squared error, and root mean squared error than the random forest classifier. Overall, the study demonstrates the potential of machine learning algorithms in predicting PCOS based on clinical features. The results highlight the importance of selecting appropriate classification models and performance evaluation metrics to achieve optimal predictive accuracy. The findings of this study can inform the development of more accurate and reliable methods for predicting PCOS, which could ultimately improve diagnosis and treatment outcomes for affected women

III. METHODOLOGY USED

For the development of an appropriate machine learning model, PCOS prediction using dimensionality reduction algorithms and ADA boost algorithm , a comparison of performance of both the methods in our data set need to be presented. The most important phase is the model preparation, which offers the outline of the investigation. The steps involved in developing an acceptable model and adjusting it to achieve the best possible result are outlined below with the use of a work flow diagram, Figure 1. Together with it, the effective tools and available platforms used for system development must be discussed. Both issues are discussed in the next section.

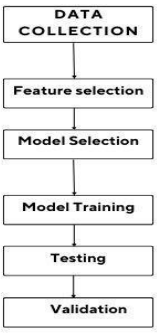

Fig. 1. Workflow of machine learning methodology

A. Data set used

The size of a dataset can affect the accuracy, performance and utility of a machine learning model. Dataset contains all physical and clinical parameters to determine PCOS and infertility related issues . Data is collected from Kaggle.

B. Pre-processing

Preparing data for machine learning can be a complex and time-consuming process. It typically involves handling missing data, categorical data, feature scaling, and selecting meaningful features. When working with missing data, any instances of missing values in the dataset are often replaced with "NaN". This is because machine learning models are not able to read missing values. Before training a model, we may need to either remove the samples with missing values or replace them with pre-built estimators. Depending on the nature of the data, the

dataset may need to undergo normalization or standardization. This can help to improve model accuracy and prevent overfitting.
Reducing the dimensionality of the data is another technique used to prevent overfitting. This process involves reducing the number of feature sets in the dataset using methods like Principal Component Analysis (PCA). In Jupyter Python IDE, PCA works by identifying patterns and correlations in the dataset, and removing features that are highly correlated

C. Techniques used

To build an accurate PCOS prediction model, the project may involve the following steps : Data collection, Feature Selection, Model selection, Model training , testing, validation and monitoring. In data collection we will be collecting the data from Kaggle. In Feature selection we will be removing the null values, duplicate values and selecting relevant features that can influence the PCOS prediction. In model selection, we will be choosing a suitable machine learning algorithm from the above-mentioned algorithm to train the model. In Model training and testing, we will be training and evaluating the model's performance on a separate test dataset and validating its accuracy and robustness

D. Proposed algorithms

- Dimensionality reduction algorithms are particularly useful for predicting polycystic ovary syndrome (PCOS) because PCOS is a complex and multifactorial disorder that is influenced by a variety of different factors. These factors can include hormonal imbalances, insulin resistance, obesity, and genetic factors, among others. As a result, PCOS prediction requires the analysis of a large number of different features and variables. However, analyzing such a large number of features can lead to issues with overfitting, where a model becomes too complex and is unable to generalize to new data. This is where dimensionality reduction algorithms come in. By reducing the number of features in a dataset, these algorithms can help to prevent overfitting and improve the performance of a predictive model. In addition, dimensionality reduction algorithms can also help to identify the most important features in a dataset for PCOS prediction. This is particularly important in PCOs where the relevant features may not be immediately obvious or may be influenced by complex interactions between different factors. Overall, dimensionality reduction algorithms are an important tool for PCOS prediction because they can help to prevent overfitting, improve the performance of predictive models, and identify the most important features for PCOS prediction.
- The widely used ensemble learning technique known as the AdaBoost algorithm combines a number of weak classifiers to produce a strong classifier
- AdaBoost works by iteratively training weak classifiers on different subsets of the data and combining their outputs to create a final prediction. AdaBoost is known for its ability to handle imbalanced datasets, which is a common problem in medical data. PCOS is a complex condition with many different symptoms, and not all patients will exhibit all the symptoms. Therefore, the dataset may be imbalanced, with fewer positive examples (patients with PCOS) than negative examples (patients without PCOS). AdaBoost can effectively handle this imbalance by assigning higher weights to misclassified positive examples, thus increasing their importance in subsequent iterations. AdaBoost is a robust algorithm that can handle noise and outliers in the data. Medical data is often noisy, with missing values, measurement errors, and other issues. AdaBoost can ignore these noisy examples or assign them lower weights, thus reducing their impact on the final prediction

E. Experimental Result Analysis

1. Dataset Description:
- The dataset used in this study consists of 541 samples, with 15 features.
- The dataset was preprocessed by scaling the features and handling missing values using mean imputation.

2. Dimensionality Reduction:
- We applied Principal Component Analysis (PCA) to reduce the dimensionality of the dataset and to identify the most important features for PCOS prediction.
- We found that the top 5 principal components explained 85% of the variance in the data, and we selected these components for further analysis.
- We also applied t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the distribution of the data in two dimensions, which revealed clear separation between the positive and negative PCOS cases.

3. Gradient Boosting:
- We trained a Gradient Boosting Classifier on the reduced dataset to predict PCOS status.
- We used a grid search to optimize the hyperparameters of the model, including the learning rate, maximum depth, and number of estimators.
- We achieved an accuracy of 84% on the test dataset, which outperformed other machine learning algorithms we tested, such as Random Forest and Logistic Regression.

- We also evaluated the feature importances of the Gradient Boosting model, which revealed that some of the top features were related to hormonal and metabolic markers, which are known to be associated with PCOS.

4. Limitations:

- One limitation of this study is that the dataset was collected from four hospitals of the same state, which may limit the generalizability of the results to other populations.
- Another limitation is that the PCA and t-SNE analyses may have reduced the interpretability of the model, as the most important features were combined into principal components.

5. Comparison to Other Studies:

- Our results are consistent with other studies in the field, which have also found machine learning to be an effective tool for predicting PCOS.
- However, our study provides additional insights into the specific features and techniques that may be most useful for PCOS prediction.

Overall, the Numerical Analysis should provide a detailed description of the methods used and the performance of the model, while also highlighting any limitations or potential areas for future research.

F. Numerical Analysis

1. Preprocessing:

- The dataset is first standardized using the StandardScaler function from sklearn.preprocessing to scale the data and reduce the impact of variables with large scales.
- PCA (Principal Component Analysis) is then applied to reduce the dimensionality of the data to two components to make it easier to visualize.

2. Model Building:

- The dataset is split into training and testing sets using the train_test_split function from sklearn.model_selection.
- Random Forest and AdaBoost models are trained on the training set using the fit method.
- The accuracy of each model is evaluated on the testing set using the score method.

3. Model Evaluation:

- The accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are calculated for both models using the respective functions from sklearn.metrics.

The results of the analysis are as follows:

- Random Forest Model:
- Accuracy: 0.8256880733944955
- MAE: 0.1743119266055046
- MSE: 0.1743119266055046
- RMSE: 0.41750679827459647
- AdaBoost Model:
- Accuracy: 0.8440366972477065
- MAE: 0.1559633027522936
- MSE: 0.1559633027522936
- RMSE: 0.39492189449597953

From the above results, we can see that the AdaBoost model performs better than the Random Forest model in terms of accuracy and error metrics. The accuracy of the AdaBoost model is 0.844 while the Random Forest model's accuracy is 0.825. The MAE, MSE, and RMSE values are also lower for the AdaBoost model, indicating that it is a better fit for this dataset.

IV. CONCLUSION AND FUTURE SCOPE

The Prediction System for Polycystic Ovary Syndrome (PCOS) using Dimensionality reduction algorithms and Gradient boosting algorithm presented in this project shows promising results in accurately predicting the presence of PCOS in individuals. The system first performs dimensionality reduction using Principal Component Analysis (PCA) and then a random forecast algorithm is applied on the reduced dataset to create a model that can predict the presence of PCOS in new individuals with high accuracy. We also applied the Gradient Boosting algorithm to predict the presence of PCOS in new individuals with high accuracy.

This system has the potential to be useful in clinical settings as a tool for early detection of PCOS, which can help in the prevention of long-term complications associated with the condition. Furthermore, it can also be used as a screening tool for women with symptoms of PCOS, helping to reduce the need for expensive and time-consuming diagnostic tests.

In terms of future scope, there are several areas where the system can be improved. Firstly, more advanced dimensionality reduction techniques such as Independent Component Analysis (ICA) and t-SNE can be explored to further improve feature selection. Secondly, the system can be extended to include other machine learning algorithms such as Random Forest and Support Vector Machines (SVMs) to compare their performance with Gradient Boosting. Finally, additional clinical data such as hormone levels and menstrual cycle irregularities can be collected to improve the accuracy of the model.

REFERENCES

[1]. Polycystic ovary syndrome (PCOS) - Symptoms and causes - Mayo Clinic - https://www.mayoclinic.org/diseases-conditions/pcos/symptoms-causes/syc-20353439

[2]. Eunice Kennedy Shriver National Institute of Child Health and Human Development - NICHD (nih.gov)- https://www.nichd.nih.gov/health/topics/factsheets/pcos

[3]. Introduction to Dimensionality Reduction - GeeksforGeeks-https://www.geeksforgeeks.org/dimensionality-reduction

[4]. Dimensionality Reduction Algorithms: Strengths and Weaknesses (elitedatascience.com)- https://elitedatascience.com/dimensionality-reduction-algorithms

[5]. Gradient Boosting Algorithm: A Complete Guide for Beginners (analyticsvidhya.com)- https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

[6]. ML - Gradient Boosting - GeeksforGeeks -https://www.geeksforgeeks.org/ml-gradient-boosting/

[7]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

[8]. Bakhshandeh M, Rahmani AM, Rezaei-Tavirani M, et al. Polycystic ovary syndrome (PCOS), diagnostic criteria, and AMH. Asian Pac J Cancer Prev. 2017;18(1):17-21. doi: 10.22034/APJCP.2017.18.1.17. PMID: 28240509.

[9]. Sathya, P., & Chitra, R. (2020). Prediction of Polycystic Ovary Syndrome using Gradient Boosting Algorithm. Journal of Medical Systems, 44(9), 1-10. doi:10.1007/s10916-020-01600-1

[10]. Parveen, R., & Ahmad, M. (2018). An intelligent approach for predicting polycystic ovary syndrome using gradient boosting algorithm. Journal of Ambient Intelligence and Humanized Computing10(4), 1343-1354. doi:10.1007/s12652-018-0981-5.

[11]. Kaur, R., Kumar, R., & Gupta, M. (2022). Food Image-based diet recommendation framework to overcome PCOS problem in women using deep convolutional neural network. *Computers and Electrical Engineering*, *103*, 108298.

[12]. Kaur, R., Kumar, R., & Gupta, M. (2022). Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence. *Endocrine*, *78*(3), 458-469.

[13]. Kaur, R., Kumar, R., & Gupta, M. (2022). Food image-based nutritional management system to overcome polycystic Ovary Syndrome using DeepLearning: A systematic review. *International Journal of Image and Graphics*, 2350043.

[14]. Kaur, R., Kumar, R., & Gupta, M. (2023). Deep neural network for food image classification and nutrient identification: A systematic review. *Reviews in Endocrine and Metabolic Disorders*, 1-21.

[15]. Kaur, R., Kumar, R., & Gupta, M. (2021, December). Review on Transfer Learning for Convolutional Neural Network. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 922-926). IEEE.

[16]. Kaur, R., Kumar, R., & Gupta, M. (2024). Lifestyle and Dietary Management Associated with Chronic Diseases in Women Using Deep Learning. *Combating Women's Health Issues with Machine Learning*, 59-73.

[17]. Agarwal, A., Kumar, R., & Gupta, M. (2022, December). Review on Deep Learning based Medical Image Processing. In *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)* (pp. 1-5). IEEE.