2023 IEEE 9th International Women in Engineering (WIE) Conference on
Electrical and Computer Engineering (WIECON-ECE)

Evaluating Machine Learning Model Performance in Predicting Polycystic Ovarian Syndrome

Fariha Jannat Ananna[1], Afsana Khan[2], MD Sadi Ashraf[3], Fatema Tuz Zohora[4],
Md Tanzim Reza[5],Md Mizanur Rahman[6]
[6]Department of Computer Science and Engineering,Purbachal American City,
Kanchon 1460, Green University of Bangladesh
[2,3,5]Department of Computer Science and Engineering, BRAC University, 66 Mohakhali, Dhaka 1212, Bangladesh
[1,4]Department of Computer Science and Engineering, Dafoddil Smart City, Birulia 1216, Dhaka
Email: [1]fjananna.98@gmail.com, [2]afsanak9090@gmail.com, [3]sadisayesgo@gmail.com, [4]shefanoor125@gmail.com,
[5]tanzim.reza@bracu.ac.bd, [6]rafi79466@gmail.com

***Abstract*—The increasing incidence of polycystic ovary syndrome (PCOS) caused significant study. Polycystic ovary syndrome (PCOS) is a hormonal disorder that disrupts the menstrual cycle and affects a significant proportion of women in reproductive age. When compared to healthy women, those who have been diagnosed with Polycystic Ovary Syndrome (PCOS) typically experience less than eight menstrual cycles each year. Significant repercussions, such as infertility and ovarian cyst development, can result from menstrual cycle irregularities. Symptoms of polycystic ovary syndrome (PCOS) include but are not limited to: infrequent menstrual periods, excess weight, skin hyperpigmentation, insulin resistance, and high blood pressure. Even after being diagnosed with Polycystic Ovary Syndrome (PCOS), many women still don't understand what it means. This means that many people are still not receiving the care they need. The current study examines this question by comparing several machine learning techniques for detecting polycystic ovarian syndrome (PCOS), a precursor to the aforementioned disease. To conduct a comparative analysis with a sample size of 1500 women, a survey instrument consisting of a patient questionnaire was developed.This study highlights the significance and utility of twelve possible points gained from the examination of nineteen medical and physiological test outcomes. Logistic regression, K-Nearest Neighbor (KNN), Gaussian Naive Bayes, Random Forest Classifier, and Support Vector Machine (SVM) are used to identify polycystic ovary syndrome (PCOS). With a 96% accuracy rate, the Radial SVM was selected as the most suited and efficient technique for predicting PCOS.**

***Index Terms*—Polycystic Ovarian Syndrome, Machine Learning, PCOS, PCOS Predict.**

I. Introduction

Better medical treatment may become available to more people as a result of the joint efforts of humans and technology. Machine learning is a subfield of artificial intelligence that enables computers to teach themselves new abilities and improve their efficiency without the use of humans or traditional statistical analysis. The fundamental focus of machine learning is on the creation of algorithms that can efficiently gather and utilise current datasets. Machine Learning systems used for detection, data prediction, and picture identification have had a substantial impact on the healthcare industry.

The hormonal disorder polycystic ovary syndrome (PCOS) is common among sexually active women. Miscarriage, polycystic ovaries, heart disease, type 2 diabetes, obesity, and other health problems are all linked to this complex endocrine issue. In the absence of ovulation, there may be changes in the levels of progesterone, oestrogen, follicle-stimulating hormone (FSH), and luteinizing hormone (LH). Approximately 12%-21% of reproductive-aged women suffer from polycystic ovarian syndrome (PCOS), a chronic medical illness for which over 70% of affected individuals go untreated. Clinical, biochemical, and radiological test results are all used in the diagnostic procedure. It has been established that as women age and enter menopause, their symptoms improve. The signs and symptoms of polycystic ovary syndrome (PCOS) may improve with the help of prescribed medication and adjustments in lifestyle.

Reproductive health therapies include things like birth control pills, insulin for diabetes, hormone replacement therapy, anti-androgen medications for males, and ultrasounds for monitoring. In many cases, novel approaches are successful when tried and true ways have failed. One potential result of surgically drilled reproductive organs is increased ovulation due to decreased testosterone levels. The wide variety of symptoms associated with this illness requires numerous medical evaluations and an excessive amount of x-ray imaging treatments. In order to effectively diagnose and treat polycystic ovarian syndrome (PCOS), it is crucial to catch the condition early on and use as few diagnostic tools as possible. This is owing to the fact that ovaries may not operate as well if you have this illness, which might increase your risk of infertility, pregnancy difficulties, and even obstetric tumours. Patients may also endure emotional discomfort as a result of the time, effort, and money wasted on unneeded medical procedures. There is a need for further improvements in terms of accuracy and precision when using medical data, despite the fact that studies have been conducted on the use of machine learning algorithms for PCOS diagnosis.

979-8-3503-1965-1/23/$31.00

II. Literature Review

A. Previous Works

[1] Classification and Regression Trees (CART), Random Forest Classifier, Naive Bayes classifier, Support Vector Machine (SVM), K-Nearest neighbour (KNN), and logistic regression were all discovered by Amsy Denny using machine learning algorithms. 541 women were consulted by experts to determine the model's diagnosis. In the end, testing on the dataset showed an accuracy of 89%. [1] . Authors [2], PCOS was detected via artificial intelligence. Tools including Support Vector Machines, Convolutional Neural Networks, Naive Bayes Classification, Random Forest, and Logistic Regression were used. Ten hospitals in Kerala provided data for this Kaggle study. With this data set, they achieved 92% accuracy. Namrata Tanwani [3],Machine learning linked obesity, insulin resistance, blood pressure, depression, and inflammation to PCOS. Identification uses Logistic Regression and K-Nearest Neighbour. KNN had the highest weights while Logistic Regression has 10 parameters. The F1 score determined the best classifier from two. Logistic Regression is used to detect PCOS because its F1 score is 0.92, compared to 0.90 for KNN. [4], PCOS was discovered using data mining techniques. The Support Vector Machine, the Naive Bayes Classifier, and the Decision Tree were used to study PCOS; adopting these methods in the future will considerably increase our ability to forecast the development of PCOS in a systematic manner.

[5],To identify additional PCOS genotypes, Researchers are developing a new machine learning approach. 233 women suspected of having PCOS participated in the study. In terms of computational properties, the PCOS genotype was compared to all others in the genome. Several kernel functions were used to test the Support Vector Machine (SVM) and K-nearest neighbour (KNN) classifiers. In terms of accuracy, the linear kernel outperformed the other SVM kernels.

They demonstrated a physiological and medical way for diagnosing PCOS. Their method creates a feature vector using medical and physiological characteristics, and a two-sample test identifies statistically significant features to differentiate between the normal and PCOS groups.The feature is defined using Bayesian and Logistic Regression (LR) approaches. Bayesian classifiers may attain an accuracy of 93.93 percent, which is much greater than logistic regression's 91.04 percent. [6]

[7],In this study, they determined a set of criteria, both maximum and minimal, to use in the initial diagnosis and treatment of PCOS. Random Forest, Support Vector Machine, Logistic Regression, Gaussian Naive Bayes, and K-Nearest Neighbours are the five machine learning classifiers that have been used to make PCOS diagnosis predictions.The Random Forest Classifier has shown to be the most effective and trustworthy classifier available.

[8],In this investigation, ultrasound pictures of the ovary were used as diagnostic tools for polycystic ovary syndrome (PCOS) by the application of image segmentation and classification techniques inspired by artificial intelligence.In addition, many studies have been conducted to detect and diagnose PCOS utilising artificial intelligence techniques such as Neural Networks (NN), Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Bayesian Classifier, Logistic Regression (LR), and k-Nearest Neighbours (kNN).When applied to PCOS data, these have the highest accuracy. To determine whether or not PCOS could be predicted by demographic variables, the existence of oligomenorrhea, the prevalence of hyperandrogenism, or both, 147 women were observed over a period of two years.Clinical and biochemical profile information, age at diagnosis, and family history of diabetes and cardiovascular disease were later documented.Based on replies to a brief questionnaire assessing symptoms, this study indicated that 4.8% of reproductive-aged women in Northern Sweden fit the criteria for polycystic ovarian syndrome (PCOS). [9]

Using the Rotterdam ESHRE/ASRM, genetics, and metabolic syndrome as criteria, this study examines the effects of PCOS on women's health and infertility. Treatments to prevent adult PCOS and related metabolic derangements from being passed on from generation to generation are suggested. [10],

III. Proposed model

A. Data Collection Procedure

There exist multiple approaches to data collecting, including real-time data collection and data collection through repository sites as [1] Kaggle and the UCI machine learning repository, which is the most frequently utilized. In order to obtain accurate data on PCOS patients during this pandemic, we faced the need to travel outside of hospitals. We have collected information on PCOS patients' symptoms using a questionnaire. We identified the patients initially, and then we enquired about their symptoms. Afterwards, we moved it to a Google Form and saved it as a CSV file.

B. Statistical Analysis

Data was collected from a sample size of 1500 individuals, encompassing a diverse range of physical states and age groups. Figure 1 illustrates the distribution of individuals with Polycystic Ovary Syndrome (PCOS) and those without PCOS within our sample.

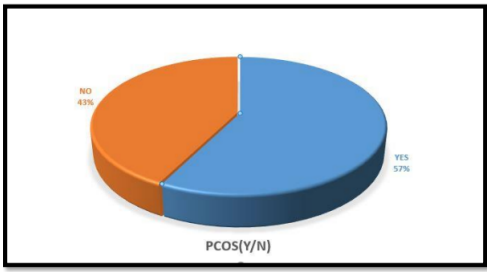

Fig. 1. Pie Chart of PCOS.

Figure 2 illustrates that a majority of the patients exhibit irregular menstrual cycles, which is identified as a contributing

factor to the manifestation of Polycystic Ovary Syndrome (PCOS).

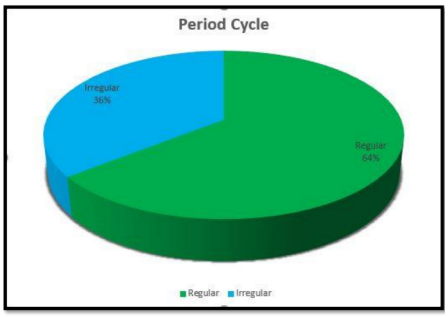

Fig. 2. Pie Chart of PCOS Cycle.

The data pertaining to undesired hair development in the patients included in the dataset is presented in Figure 3.

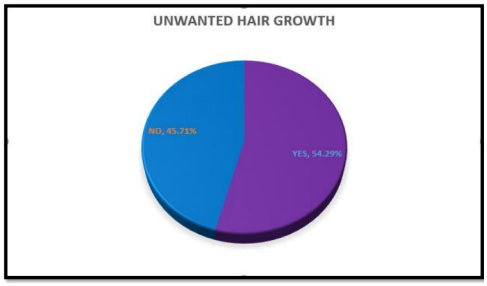

Fig. 3. Pie Chart of Unwanted Hair Growth

C. Proposed methodology

In order to fulfill the objectives of our research, it was necessary to obtain a dataset. The dataset was obtained from various hospitals and afterward, was examined for missing values.After that, the imputation approach was used to prepare the data for pre-processing. It was built for the following machine learning models: Random Forest (RF), Support Vector Machine with Linear Kernel (SVML), Support Vector Machine with Radial Basis Function Kernel (SVMR), Logistic Regression (LR), Gaussian Naive Bayes (GNB), and k-Nearest Neighbors (KNN). During the implementation of numerous algorithms, the Radial SVM technique was proven to have the best degree of accuracy, particularly 96%. The next section describes the architectural structure of the proposed technique.

In order to optimize the model's performance and reduce its computational expense, only particular sample attributes are incorporated as features. As we can see above, we have nineteen features. All of them could be useful, but there's a chance that overfitting will happen or that some won't be at all helpful. Therefore, we will use the Chi Square technique to discover significant traits. The object will be scored using the chi square technique. The calculated score indicates the significance of that feature. We'll use the top 12 essential feature. Select KBest, and chi-squared will be used to estimate

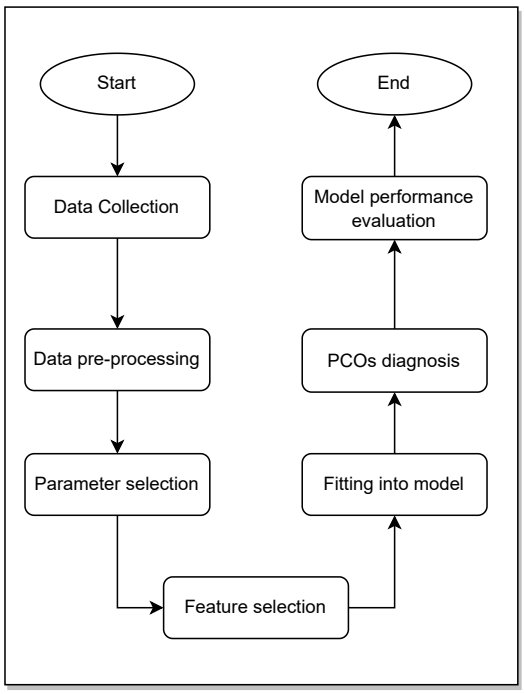

Fig. 4. Proposed Methodology.

the feature relevance. In order to predict the disorder, we constructed a predictive system and entered the values of the features chosen using the Chi square approach. When we put all six algorithms into implementation, we discovered that they could all accurately predict whether or not PCOS exists.

$$x_c^2 = \frac{\Sigma\,(O_i - E_i)^2}{E_i}$$

Fig. 5. Equation of Chi square Method

D. Dataset Splitting Methodology

In the development of our predictive model, we utilized a rigorous data splitting strategy to ensure the integrity and generalizability of our results. We partitioned the dataset into three distinct subsets: training, validation, and testing.

- **Training Set:** The training set is used to train our machine learning models. It includes 64% of the original data, which provides a substantial amount of examples for the model to learn from.
- **Validation Set:** The validation set plays a critical role in the model selection process. It represents 16% of the data and is used to fine-tune the model parameters and to provide an unbiased evaluation of model fit during the training phase.
- **Testing Set:** Finally, the testing set is kept separate from the training and validation processes. It comprises 20% of the data and is used to assess the final model's

performance, providing an unbiased evaluation of its predictive power.

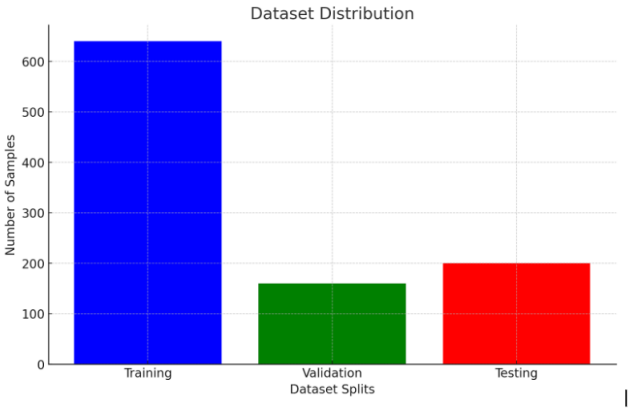

Fig. 6. Dataset Distribution.

our dataset to include 1,500 instances. The class distribution in this augmented dataset comprises 862 instances of PCOS and 638 instances of Non-PCOS.

E. Data Preprocessing

This academic study focuses on the fundamental prerequisites pertaining to inaccurate data, numerical values, function expansion, and feature extraction. The presence of invalid values in the dataset is denoted by the use of 'NaN'. Overfitting may arise when a model exhibits an inability to effectively capture the underlying patterns and relationships present in a given dataset, particularly with respect to a specific variable. One strategy to mitigate overfitting is to reduce the size of the sample space by limiting the number of feasible choices. This process is achieved by the reduction of dimensionality and the normalization of the data.

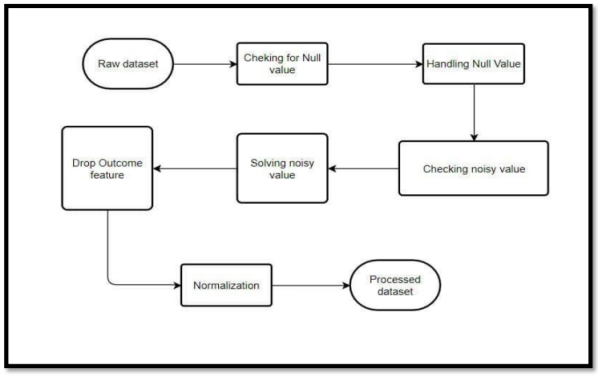

Fig. 7. Data Preprocessing.

F. Feature Selection

The model's performance is enhanced and computational costs are reduced by utilizing only specific qualities of the samples as features. The Chi-Square approach has been employed to discover significant traits. The chi-square method is utilized to compute a statistical score. The determined score provides insight into the significance of the respective attribute. A selection of the 12 most significant qualities is utilized. The feature relevance will be determined using the methods of Select KBest and chi-squared. Table I presents a comprehensive overview of the qualities and their corresponding association to the target variable, organized in a descending manner. The greater the weight of a characteristic, the greater its independent influence on the objective.

TABLE I
FEATURE SELECTION.

Feature	Score
Weight (kgs)	314.327905
BMI	188.179570
Age (yrs.)	51.003770
Weight gain(Y/N)	22.075743
Marriage Status (Y/N)	16.612233
Pregnant(Y/N)	13.272180
High BP(Y/N)	13.007881
Unwanted Hair Growth(Y/N)	8.339211
Skin Darkening (Y/N)	7.415745
Height(cm)	6.608136
Cycle(R/I)	6.393862
Anxieties(Y/N)	6.15033

G. Fitting into Model

Following the completion of data cleaning and selection procedures, the dataset is now ready for further processing through the algorithms. 'Linear Support Vector Machine (SVM),' 'Radial Support Vector Machine (SVM),' 'Logistic Regression,' 'Random Forest Classifier,' 'K Neighbours Classifier,' and 'Gaussian Naive Bayes' are the six models that will be used for supervised machine learning.

IV. RESULT AND ANALYSIS

We employed six methods to analyze our preprocessed datasets, which consisted of a total of 19 features. Subsequently, Chi-square analysis was used to identify meaningful characteristics. With the chi-square method, a statistical score is calculated to show how important a characteristic is. Six algorithms' precision is shown in Figure 8 below.

We can infer the following results from the bar chart: Logistic Regression achieved 93% accuracy, KNN achieved 93% accuracy, Linear SVM achieved 93% accuracy, Radial SVM achieved 96% accuracy, Gaussian Naive Bayes achieved 88% accuracy, and Random Forest achieved 95% accuracy. At last, we have enough evidence to conclude that Radial SVM performed at its highest level of accuracy.

The ROC Curve of Radial SVM is given in figure 9

TABLE II
COMPARATIVE EVALUATION.

Method/Work Done	Sample size	Size of Feature set	Algorithm	Accuracy
This work	1500	19	Radical SVM	96%
[1] AmsyDenny	541	23	Random Forest	89%
[2] Malik Hasan	1000	42	Random Forest	96%
[3] Namrata Tanwani	538	39	Logistic Regression	92%
[5] XingZhong Zhang	306	25	SVM linear	80%
[6] Palak Mehrotra	250	9	Bayesian classifier	93%
[7] Vaidehi Thakre	540	30	Random Forest	90%

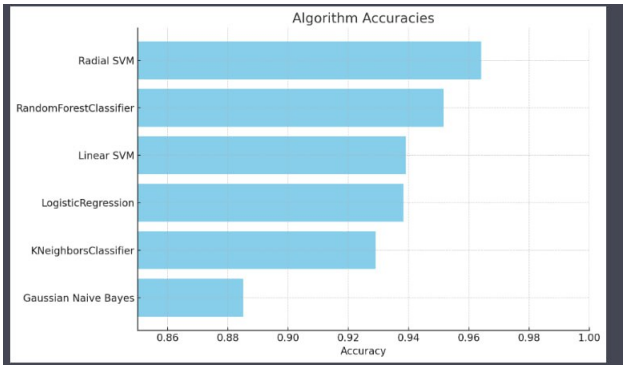

Fig. 8. : Accuracy of Six Algorithms.

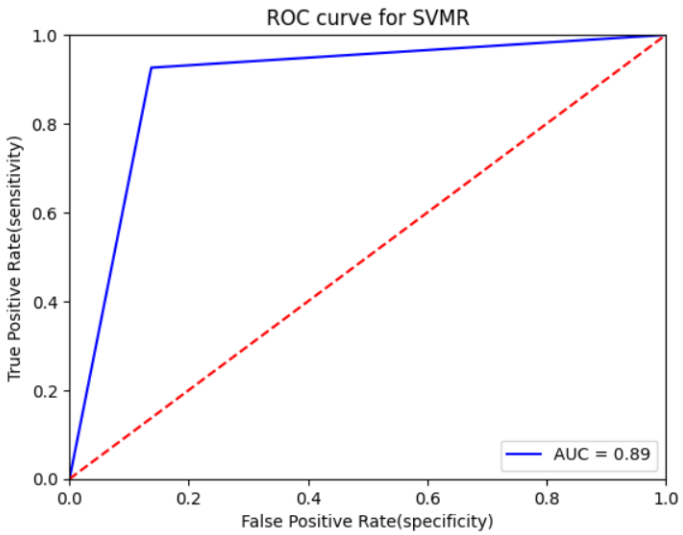

Fig. 9. ROC curve of SVMR

Finally, the models were evaluated using the parameters such as precision, recall, and F1 score. The results based on the parameters are given in III

The table II shows the comparison between our work and some of the best works in the literature. The information given in the table showcases that our work performs fairly well compared to some of the other works in the literature, even surpassing some of them. A dataset consisting of 1500 instances was utilised in the experimentation, employing a Radial Support Vector Machine (SVM) model. The resulting accuracy achieved was 96%. We have conducted an analysis on an additional seven papers. We obtained the highest level of accuracy compared to the other options.

TABLE III
MODEL RESULTS

Models	Precision	Recall	F1-Score
Random Forest Classifier	95%	96%	98%
K-Neighbors Classifier	92%	89%	92 %
Linear kernel SVM	93%	85%	90 %
Logistic Regression	93%	82%	88%
Radial kernel SVM	96 %	95 %	89 %
Gaussian Naive Bayes	88%	70%	68%

V. CONCLUSION AND FUTURE WORKS

Polycystic Ovary Syndrome (PCOS) is a prevalent endocrine condition that frequently affects women within the reproductive age range. The occurrence of infertility and anovulation may arise as a consequence of this condition. The diagnostic criteria encompass clinical and metabolic indications that function as biomarkers for the disease. A novel approach was developed to identify polycystic ovary syndrome (PCOS) by utilising a limited set of potential indicators. In this study, a range of widely employed machine learning algorithms were applied to analyse clinical data from patients diagnosed with Polycystic Ovary Syndrome (PCOS) using symptom-based criteria. The methods utilised in this investigation included Support Vector Machines (SVM) with both linear and radial kernels, Logistic Regression, K-Neighbors Classifier, Gaussian Naive Bayes, and Random Forest. The performance evaluation parameters, namely recall, accuracy, precision, and F-statistics, demonstrated that SVMR exhibited the highest level of performance. Therefore, it can be inferred that the Support Vector Machine Regression (SVMR) algorithm is the most appropriate algorithm for the diagnosis of Polycystic Ovary Syndrome (PCOS) based on the provided data.

The potential future directions of this study may involve the utilisation of diverse or extensive datasets for the purpose of disease diagnosis. In order to effectively utilise our model, it is imperative that we integrate it into various platforms

such as a product, web application, or Android application inside our local context. In the future, it will be possible to enhance the precision of our model by employing a larger dataset. In addition, the development of user-friendly graphical user interfaces (GUIs) can enhance the accessibility of the model's product to a wider range of individuals. In the future, we intend to incorporate medical diagnostic photos into the existing collection, so enhancing its informational value to the greatest extent possible.

REFERENCES

[1] Amsy Denny, Maneesh Ram C, Anita Raj, Ashi Ashok, and Remya George. i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques. In *2019 IEEE TENCON*, pages 1–6. IEEE, 2019.

[2] Malik Mubasher Hassan and Tabasum Mirza. Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. *Int. J. Comput. Appl*, 975:8887, 2020.

[3] Namrata Tanwani. Detecting pcos using machine learning. *IJMTES—International Journal of Modern Trends in Engineering and Science*, 7(01):2348–3121, 2020.

[4] Palvi Soni and Sheveta Vashisht. Exploration on polycystic ovarian syndrome and data mining techniques. In *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pages 816–820. IEEE, 2018.

[5] Xing-Zhong Zhang, Yan-Li Pang, Xian Wang, and Yan-Hui Li. Computational characterization and identification of human polycystic ovary syndrome genes. *Scientific reports*, 8(1):12949, 2018.

[6] Palak Mehrotra, Jyotirmoy Chatterjee, Chandan Chakraborty, Biswanath Ghoshdastidar, and Sudarshan Ghoshdastidar. Automated screening of polycystic ovary syndrome using machine learning techniques. In *2011 Annual IEEE India Conference*, pages 1–5. IEEE, 2011.

[7] Vaidehi Thakre, Shreyas Vedpathak, Kalpana Thakre, and Shilpa Sonawani. Pcocare: Pcos detection and prediction using machine learning algorithms. *Biosci Biotechnol Res Commun*, 13(14):240–244, 2020.

[8] V Deepika. Applications of artificial intelligence techniques in polycystic ovarian syndrome diagnosis. *J. Adv. Res. Technol. Manag. Sci*, 1(3):59–63, 2019.

[9] Åsa Lindholm, Liselott Andersson, Mats Eliasson, Marie Bixo, and Inger Sundström-Poromaa. Prevalence of symptoms associated with polycystic ovary syndrome. *International Journal of Gynecology & Obstetrics*, 102(1):39–43, 2008.

[10] Gautam N Allahbadia and Rubina Merchant. Polycystic ovary syndrome and impact on health. *Middle East Fertility Society Journal*, 16(1):19–37, 2011.