PREDICTIVE MODELING OF PCOS: A MULTI-LINEAR REGRESSION ANALYSIS INCORPORATING HORMONAL, CLINICAL, AND LIFESTYLE PARAMETERS

Kokila Ramesh [1], Kumudha H R [2*], Angel Richard [3], Anita Chaturvedi [4]
[1]Associate Professor, Faculty of Engineering and Technology Jain (Deemed-to-be University), Bangalore, India
[2] Research Scholar, Jain University, Bangalore and Assistant Professor, Bharathi College PG & RC, Mandya, India.
[3]Research Scholar, Jain (Deemed-to-be University), Bangalore, India.
[4]Professor, Faculty of Engineering and Technology Jain (Deemed-to-be University), Bangalore, India
*Corresponding Author: Kumudha H R, E-mail: kumudhamayur@gmail.com

Abstract
Polycystic Ovary Syndrome (PCOS) is a complex endocrine disorder affecting reproductive-aged women, characterized by hormonal imbalances and diverse clinical manifestations. This study aims to develop a multi-linear regression model to predict PCOS based on various parameters. Comprehensive assessments including menstrual cycle regularity, hair growth weight gain, fast food consumption, skin darkening, follicle counts (left and right ovaries), insulin levels, Anti-Mullerian Hormone (AMH) concentrations, and the presence of pimples were performed on a group of individuals presenting with PCOS symptoms. Statistical analysis involved correlation studies and the development of a multi-linear regression model to elucidate relationships between these parameters and the diagnosis of PCOS. Preliminary findings suggest significant associations between fast food intake, irregular cycles, weight gain, skin darkening, excess hair growth, follicle counts, insulin levels, AMH concentrations, and the presence of pimples with the manifestation of PCOS. The multi-linear regression model exhibited predictive capability, offering insights into the combined influence of these parameters on the likelihood of PCOS development. This Model will be able to approximately predict PCOS with the help of the symptoms.

Keywords: Analysis, Clinical, Hormonal, Infertility, Lifestyle, PCOS, Statistical model.

INTRODUCTION

Hyperandrogenic ovarian stimulation (HAA) or Stein-Leventhal syndrome are other names for polycystic ovarian syndrome (PCOS). One of the most prevalent conditions affecting women who are fertile is PCOS. They undergo an altered lifestyle due to the manifestation of the disease in the form of obesity, acne, Hirsutism, Amenorrhea and other lifestyle characteristics. They also suffer from comorbid diseases such as diabetes, cardio-vascular diseases, endometrial carcinoma, etc. Though it is a non-communicable disease, the women suffering from PCOS need to manage the symptoms and it is a lifelong process. The women with PCOS needs to be improved to have a life of quality, through early diagnosis and treatment. Most of the existing literature to predict the PCOS are based on the Rotterdam criteria, or NIH or the AES criteria. Though some of the research does not explicitly mention use the above mentioned three diagnosing criteria. The parameters considered definitely fall in the realm of the standard criteria.

PCOS is a common endocrine illness that affects women in their reproductive years and interferes with metabolic, neuroendocrine, and ovarian functions [1]. Menstrual problems, infertility, hirsutism, acne, and obesity are among the main signs of PCOS. [2]. About 8% to 13% of the women suffer from this disorder as reported by World Health Organization [3]. Recent survey shows that in India, approximately 11.34% females of the reproductive age are suffering from PCOS [4]. Predisposed conditions such as genetic background, and the environmental factors such as endocrine disruptors and lifestyle, increases the risk of PCOS [1].

The symptoms of this condition in females include irregular periods, infertility [5], hair loss, obesity, acne, hair growth (WHO, [3]). The prevalence of PCOS in women makes them susceptible to various comorbid conditions such as Type 2 diabetes, endometrial cancer, hypertension, high cholesterol, hormonal imbalance, depression, mood swings affecting them both physically and psychologically ([6], [7]).

Pathophysiology

The pathophysiology of PCOS is complex and is considered to be a vicious cycle [7]. The major factors causing this syndrome are due to the wrong interaction between defective genes and the unhealthy lifestyle leading to obesity. The first hormonal imbalance is the decrease in the ratio of follicle stimulating hormone (FSH) to luteinizing hormone (LH) decreases as a result of an increase in LH and a reduction in FSH. Androgen development is induced by an increase in the hormone LH. The increased level of androgen increases the level of estrogen. The increase in estrogen decreases the FSH. Decreasing the level of FSH leads to these endocrine abnormalities. With the decrease in FSH and increase in LH, there exists a vicious cycle. In the absence of Follicle stimulating hormones, there is improper growth of follicles. Most of the cycle becomes anovulatory due to lack of FHS and also LH is high. The hormonal disturbance causes anovulation resulting in infertility. As ovulation does not take place, the abnormal follicles are converted into follicular cysts. There exist multiple cysts in the ovaries. Another feature of anovulation is the lack of progesterone. There is an upsurge of LH. Hence these endocrine abnormalities are self-propagators.

One of the causes of PCOS is obesity. This increases the risk of Hyperinsulinemia. This favors the endocrine abnormalities. The excessive levels of androgen (male hormone), leads to the development of Acne and Hirsutism. And the increase in the estrogen can cause endometrial carcinoma. The decrease in the progesterone will cause heavy bleeding. As the Progesterone and estrogen lose their cyclic behaviour. This imbalance leads to Amenorrhea (irregular menstrual cycle).

Clinical Presentations of PCOS:

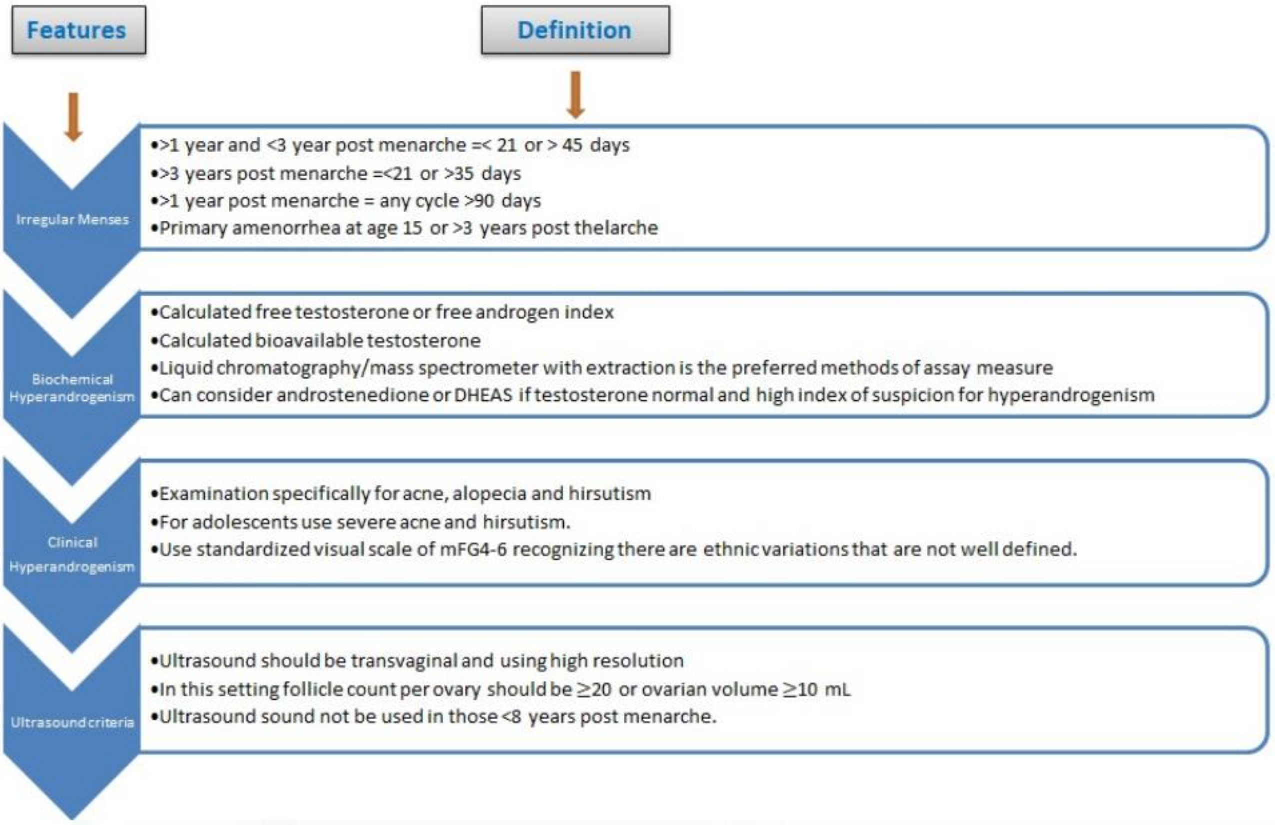
The clinical features are Hyperandrogenism, Acne, Hirsutism, Hyperinsulinemia, infertility, endometrial carcinoma, infertility, follicular cysts, heavy bleeding and amenorrhea. Approximately 60% to 80% of cases have hyperandrogenism, which is a well-established contribution to the aetiology of PCOS. 50% to 80% of women with PCOS have insulin resistance as a pathophysiological contributing factor. Obesity increases hyperandrogenism, hirsutism, infertility and exacerbates PCOS [8]. The clinical presentations differ based on the ethnicity of women [9]

Diagnostic criteria:

The clinical manifestations of PCOS are the elevated levels of androgens, insulin resistance and ovarian dysfunction. There are different methods to diagnose PCOS. The most common ones are (i) Rotterdam Criteria: any two of the three symptoms are included. polycystic ovaries, oligo-ovulation, and hyperandrogenism (ii) National Institute of health (NIH) criteria- Hyperandrogenism, Oligo-ovulation and Exclusion of other related disorders (ii) The Androgen Excess Society (AES) criteria include polycystic ovaries, oligo-ovulation, hyperandrogenism, and the elimination of other illnesses that may be associated. [10]. The exclusion of other related disorders includes thyroid dysfunction, hyperprolactinaemia, rare conditions like- Cushing syndrome, virilising tumours, and so on, [8]

Definition of the features for diagnosis:

For an evidence-based diagnosis of PCOS of women, National Health and Medical Research Council have recommended the following guidelines for the assessment of polycystic ovary syndrome, which is documented in the international evidence-based guideline documented as "International evidence- based guideline for the assessment of polycystic ovary syndrome 2018". The diagnosis recommendations for the compelling features of PCOS are presented in Figure 1.



- >1 year and <3 year post menarche =< 21 or > 45 days
- >3 years post menarche =<21 or >35 days
- >1 year post menarche = any cycle >90 days
- Primary amenorrhea at age 15 or >3 years post thelarche

- Calculated free testosterone or free androgen index
- Calculated bioavailable testosterone
- Liquid chromatography/mass spectrometer with extraction is the preferred methods of assay measure
- Can consider androstenedione or DHEAS if testosterone normal and high index of suspicion for hyperandrogenism

- Examination specifically for acne, alopecia and hirsutism
- For adolescents use severe acne and hirsutism.
- Use standardized visual scale of mFG4-6 recognizing there are ethnic variations that are not well defined.

- Ultrasound should be transvaginal and using high resolution
- In this setting follicle count per ovary should be ≥20 or ovarian volume ≥10 mL
- Ultrasound sound not be used in those <8 years post menarche.

Figure 1: Assessment of PCOS based on the guidelines provided in 2018.

REVIEW OF LITERATURE

With the development of Artificial intelligence and machine learning techniques, the prediction of health emergencies has become relatively accurate, faster diagnosis and relatively simple. A four-parameter model was developed using the LASSO logistic regression method to forecast the likelihood of PCOS occurrence in Chinese women. The parameters included anti-Müllerian hormone (AMH), menstrual cycle duration, body mass index (BMI), and testosterone. [12]. [13] defined three auxiliary variables IM (Irregular Menses')- based absence of menstruation, scanty or infrequent menstruation, irregular cycle, abnormal bleeding, and infertility, Hyperandrogenism (excess Androgen levels) based on High levels of testosterone, Hirsutism, acne, and PCOM (Ovary Morphology) using the Ultrasound reports. The machine learning models employed included Random Forest, Gradient Boosted Trees, Support Vector Machine (SVM), and Logistic Regression. The performance was assessed using the Receiver Operating Characteristic (ROC) curve, achieving a predictive accuracy of 85% prior to clinical diagnosis.

Using artificial intelligence and machine learning methods [14], datasets containing 41 attributes of women were analyzed to predict the presence of PCOS. The commonly utilized techniques for PCOS prediction include Support Vector Classification (SVM), Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), and Gaussian Naïve Bayes (GNB). The study examined five scenarios with various combinations of variables. The scenario incorporating the attributes such as Body Mass Index (BMI), Monthly Cycle length in terms of days, AMH (ng/mL), Weight gain, Hair growth, Skin darkening, Pimples, Fast food and Follicle No – left and right emerged as the best predictor. Among all the machine learning methods, Random Forest is found to be the best model. It also suggests, considering lifestyle data and other easily obtainable data for future research.

Another study [15] explores a dataset that combines clinical information from individuals diagnosed with PCOS and non-PCOS, along with ultrasound scans of the ovary. Based on this dataset, a deep learning model is proposed to detect polycystic ovarian morphology (PCOM). To aid radiologists in diagnosing PCOS using both ultrasound images and clinical data, a fusion model was developed. This model integrates image features extracted using the MobileNet deep learning architecture with clinical data, achieving an accuracy of 82.46%.

In order to detect PCOS, using AI, the Clinical Decision Support System (CDSS) has been in use in the healthcare domain for the diagnosis of the disease accurately [16]. It is imperative to select the necessary parameters to predict the disease. The process of extraction of the relevant features from a pool of various attributes is known as feature selection. The methods of feature selection include filter, wrapper, and embedded. The feature selection process employs the wrapper method. The Random Forest Classifier and the Red Deer Algorithm wrapping methods for feature selection are applied. Weight, BMI, hemoglobin, cycle duration, follicle stimulating hormone, LH, waist-hip ratio, thyroid stimulating hormone, AMH, prolactin level, fluctuating blood sugar levels, weight gain, skin darkening, hair growth, and hair loss are the twenty features that were extracted. Endometrial thickness, Follicle Number (Left), Follicle Number (Right), Average Follicle Size (Left), and Average Follicle Size (Right). This feature selection method outperforms the other conventional classifiers.

[17] put out two models. The first is a self-diagnostic prediction model for a non-invasive structure that is based on age, lifestyle characteristics, anthropometric measurements, and symptoms. The results of laboratory tests are not necessary to use the prediction tool. The variables included are acanthosis nigricans, acne, hirsutism, irregular menstrual cycles, weight gain, fast food consumption and age. Using the clinical data, a second model was created that applies all of the predictor variables to the diagnosis of PCOS. This model assessed performance using the k-fold cross validation approach and classified data using the CatBoost algorithm. For the non-invasive method, the accuracy was 81%, and for the clinical data, it was approximately 87%.

The healthcare system can be used with great benefits thanks to the development of AI and machine learning. With The vast amount of data produced with the Electronic Health Records (EHR), meaningful insights need to be generated, like observing the patterns, processing of huge information in a short span of time, efficiently. Machine learning techniques act as a support system for the healthcare professionals to make accurate, speedy and a reliable diagnosis. Some of the limitations of machine learning are the ethical concerns, loss of personal elements of healthcare and practical approaches, and the probability of error in diagnosis which might affect human life [18]. The present research work in this paper suggests a multivariate regression model with a statistical inference based on the correlation between the independent and the dependent variables and the statistical analysis.

DATA

The objective of the research is to investigate the relationship between various health indicators and the women with the presence of Polycystic Ovary Syndrome (PCOS). The dataset used for the analysis contains information about factors such as menstrual cycle length, Anti-Mullerian Hormone (AMH) levels, hair growth, weight gain, fast food consumption, pimples, skin darkening, follicle count and insulin levels. The goal is to build a multi linear regression model to predict the presence of PCOS based on these variables.

The dataset used in this research is obtained from www.kaggle.com [19]. Data is provided extensively with many features for either having PCOS or not having the same. This data gave the insight to the authors to come out with the statistical analysis for choosing the suitable variables affecting the women life with PCOS. It contains information collected from individuals, and the variables are carefully selected based on their correlation associated with PCOS and it is shown in Figure 2 for visual appreciation. The variables selected have a dominant correlation with PCOS for the data length selected for the study. In the current study, the dataset is split into training and testing sets with 80% of the data being used for training the multi-linear regression model, and the remaining 20% for testing.

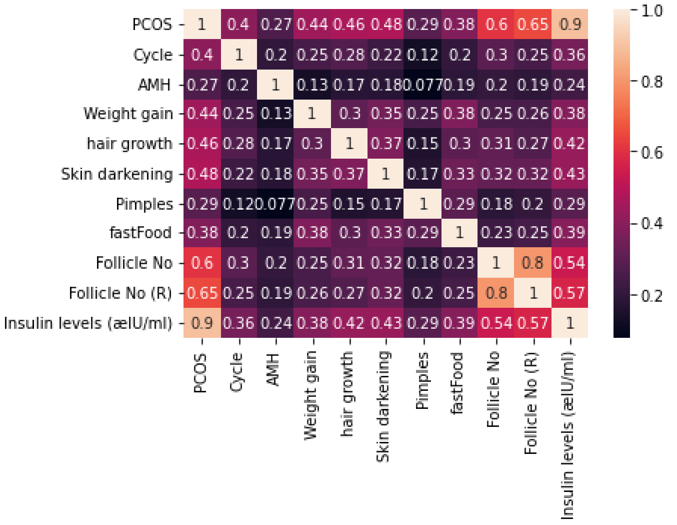


Figure 2: Correlation Matrix constructed between PCOS and the clinical measures of the body changes

EXPERIMENTAL METHODS

Regression Model

The multi-linear regression model has been used here with the underlying information of the features affecting PCOS and is shown in the correlation matrix in Figure 1. For the data sample available, the correlation coefficient value beyond 0.25 is considered as the dominant one using correlation test between the variables. The model given in equation (1) is trained for the data set which is used for the training the model.

The multiple linear regression equation for modelling PCOS in terms of the variables included is expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \varepsilon \text{ -----} \quad (1)$$

Here:

Y is the dependent variable (PCOS values in this study)

β_0 is the intercept

$\beta_1, \beta_2 \ldots\ldots\ldots \beta_{10}$ are the coefficients corresponding to the independent variables $X_1, X_2, \ldots\ldots X_{10}$ such as Cycle, AMH, Growth of hair, obesity, unwanted food, pimples, Skin darkening, Follicle No (Left), Follicle No (Right), Insulin levels respectively. ε is the error term.

The model is trained to estimate the parameters $(\beta_1, \beta_2 \ldots\ldots\ldots \beta_{10})$ from the observed (Y) and predicted $(\hat{Y})$ values. The parameters in the model are obtained by minimizing the squared errors between the data and model in (1) using least squares approach.

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10}$$

To find the efficiency of the model, two primary measures are employed such as the coefficient of association (R^2) and the mean square error (MSE) between the observed and the predicted data. The formula for finding MSE between the observed and the predicted values is shown in equation (2).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ldots\ldots\ldots\ldots\ldots \quad (2)$$

Where, n is the number of observations. y_i is the observed PCOS values and $\hat{y}_i$ is the predicted PCOS values. In this study the value of MSE is 0.04, it indicates that the model is efficient and further it can be used for testing purpose for which the data is available. Testing of the model provides an indication of the model efficiency better as the data is not biased for the parameters calculated in general.

The coefficient of association (R^2) given in equation (3) represents the proportion of the variance in the dependent variable that is explained from the model with the parameters estimated

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \ldots\ldots\ldots\ldots \quad (3)$$

Here $\bar{y}$ is the mean of the observed data. In this study the R^2 value is 0.88, it indicates 88% of the variance has been explained by the model developed in the present study. This percentage is a significant one as the number of parameters used in this model is 11 and is found to be less number in comparison with the model referred in literature review which uses more number of parameters.

Statistical Analysis

The descriptive statistical analysis provides a summary of the PCOS data for both the training and testing data sets. In the training data set with 432 samples, the average actual PCOS value is 0.34, with a moderate level of variability indicated by a standard deviation of around 0.47. Concurrently, the predicted PCOS values in the training set, derived from the model, exhibit a comparable mean of approximately 0.34, and a slightly lower standard deviation of about 0.44, suggesting the model's consistency in capturing the patterns observed in the observed PCOS values. The testing data set, consisting of 109 samples, demonstrates a similar trend with an average observed PCOS value of approximately 0.29 and a standard deviation of about 0.46, signifying a certain level of variability. The predicted PCOS values during testing data set have an average of around 0.28 and a standard deviation of about 0.42.

The statistics provided in the previous paragraph reveal the model's ability to provide predictions that align with the observed PCOS values in both training and testing data sets, with close mean values and reasonable consistency between observed and predicted PCOS data. The mean of observed and predicted PCOS values are close in both the training and testing sets, indicating that, on average, the model is providing better predictions. The standard deviations of the observed and predicted PCOS values show some difference which is acceptable as the model used here is not a perfect model. In both datasets, the standard deviation of predicted values is slightly lower than that of observed values. These statistics provide the central tendency and spread of PCOS values in the dataset and how well the model is capturing these patterns. Additionally, similar means between actual and predicted values suggest that the model is providing predictions in line with the observed data are depicted in Table 1.

Table 1: Basic Statistics of the model with the observed data of training data set and Testing data set

	Training		Testing	
	Observed Data	**Predicted Data**	**Observed Data**	**Predicted Data**
Sample length	432	432	109	109
Mean	0.34	0.34	0.29	0.28
Variance	0.22	0.19	0.21	0.18
Standard Deviation	0.47	0.44	0.46	0.42

The primary objectives of this research study are divided into three main categories. Firstly, the aim is to identify and extract pertinent features from a myriad of attributes, focusing on those that bear significant relevance in predicting Polycystic Ovary Syndrome (PCOS). This involves a comprehensive analysis to filter through the various attributes and focused the most influential factors in the context of PCOS prediction. Secondly, the research attempts to classify the identified parameters based on the Rotterdam criteria, a widely recognized diagnostic framework for PCOS. This classification is crucial for a complex understanding of the multifaceted nature of PCOS and aids in tailoring predictions to specific criteria. Lastly, the research aims to construct a robust multivariate regression model capable of predicting PCOS. By integrating various attributes and leveraging statistical techniques, the objective is to develop a predictive model that enhances our understanding of PCOS and contributes valuable insights for clinical applications. This comprehensive analysis involves delving into attributes such as hyperandrogenism indicators (hair growth, acne), Anti-Mullerian Hormone (AMH) levels, polycystic ovaries based on follicle count, and menstrual irregularities like oligomenorrhea, considering cycle length. Secondly, the research seeks to categorize these identified parameters according to the Rotterdam criteria, a pivotal diagnostic framework for PCOS. This classification facilitates a nuanced understanding of PCOS, aligning predictions with specific diagnostic criteria. Lastly, the study attempts to construct a robust multivariate regression model integrating these features, including lifestyle factors such as obesity and fast food consumption.

In summary, the multiple linear regression models are utilized to predict PCOS values based on various independent variables. The evaluation metrics and statistical tests offer valuable insights into the model's performance and highlight the significance of each variable in the prediction process. The combination of mathematical explanations and statistical tests ensures a comprehensive understanding of the model's behavior and its applicability.

The graph (Figure 3) displaying observed and Predicted PCOS values provide a visual representation of the model's performance. Each bar represents an individual sample index, with two bars side by side for each index - one for the observed PCOS value and one for the predicted PCOS value. The Mean Squared Error (MSE) and R-squared values provide quantitative measures of the model's performance depicted in Table 2. Correlation matrices for both the training and testing datasets underscore the strong positive correlations between actual and predicted PCOS values, indicating a close alignment between model predictions and observed data, which is depicted in Table 3 and Table 4. The data plots for both actual and predicted dataset for training and testing period is shown in figure 3 and figure 4 respectively.

Table 3: The performance of the model for Training and Testing PCOS data sets

PCOS	Training	Testing
Mean Squared Error (MSE)	0.03	0.04
R - squared	0.88	0.85

Table 4: Comparison of the model for Training data sets of PCOS

Correlation for training dataset			
	Observed data	**Predicted data**	**Error**
Observed data	1.00	0.93	-0.07
Predicted data	0.93	1.00	0.02
MSE	-0.07	0.02	1.00

Table 5: Comparison of the model for Testing data sets of PCOS

Correlation for testing dataset			
	Observed data	**Predicted data**	**Error**
Observed data	1.00	0.91	-0.08
Predicted data	0.91	1.00	0.04
MSE	-0.08	0.04	1.00

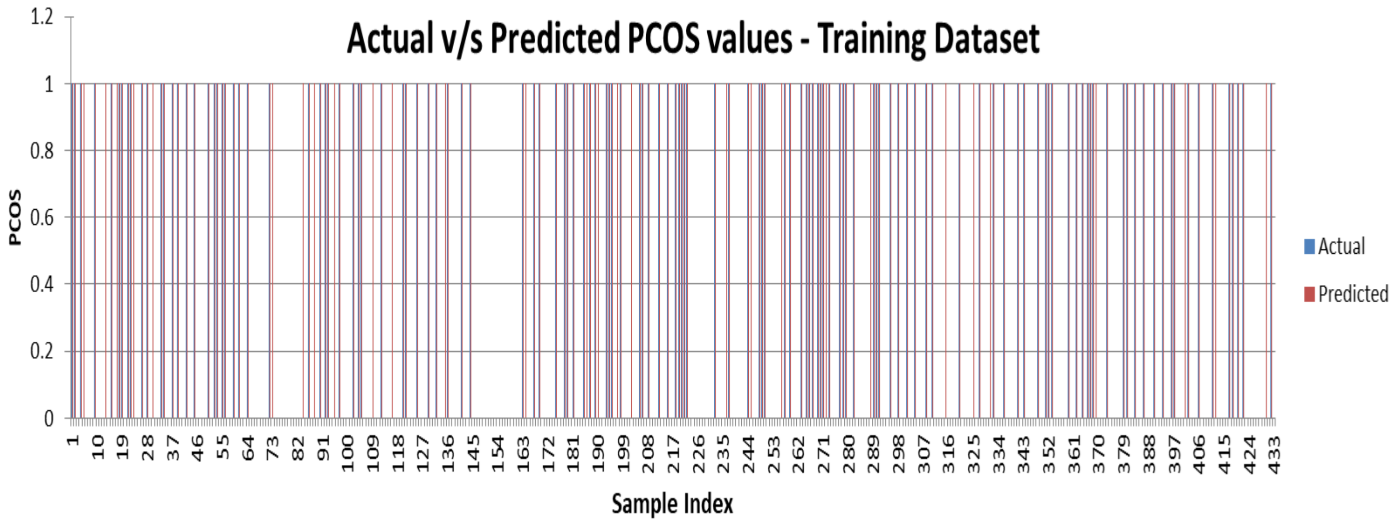


Figure 3: Visual Comparison of the observed and predicted PCOS values for the Training data sets

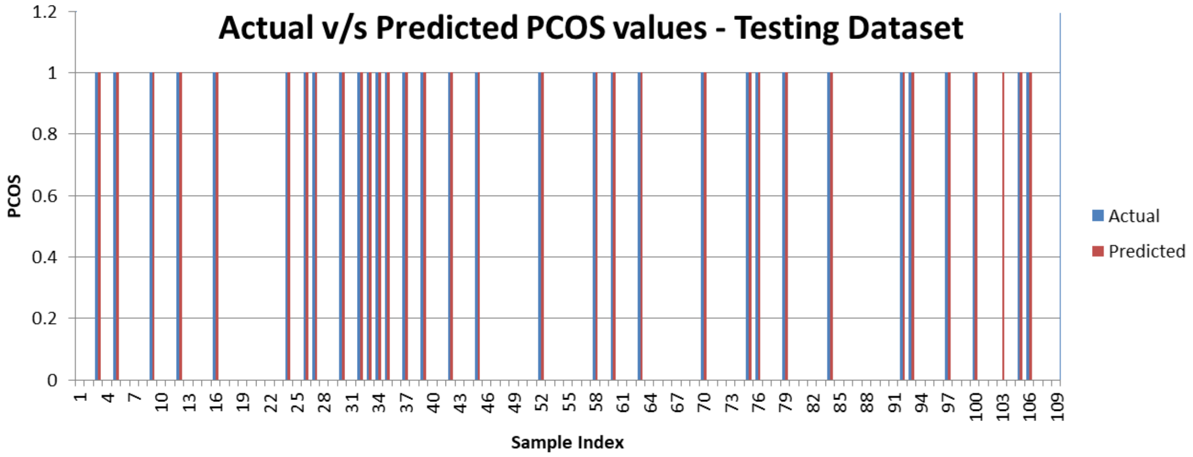


Figure 4: Visual comparison of observed and predicted PCOS values for the Testing data sets

The authors of the present paper focused on using a reduced set of 10 parameters including menstrual cycle length, AMH levels, hair growth, weight gain, fast food consumption, pimples, skin darkening, follicle count, and insulin levels. Despite using fewer parameters, the model achieved a respectable efficiency of 85-88%. This study demonstrates that with careful selection of key parameters, it is possible to achieve high model efficiency while reducing the complexity and dimensionality of the data.

The conclusion highlights the significance of the current study's approach in optimizing model efficiency with a reduced number of parameters, making it an effective strategy in scenarios where data is scarce or difficult to collect.

The conclusion highlights the significance of the current study's approach in optimizing model efficiency with a reduced number of parameters, making it an effective strategy for predicting PCOS using this model. This approach demonstrates that even with limited resources or data availability, high accuracy can be achieved by carefully selecting the most relevant parameters. The comparative study with different models have tabulated in Tale 6.

Table 5: The comparative study of PCOS

Study	Approach	Parameters Considered	Model Efficiency	Key Findings
Abrar Alamoudi et al.	Deep learning techniques	23 parameters (Age, BMI, Marital Status, Hormonal levels, Lipid tests, etc.)	82.46%	Demonstrated that using a moderate number of parameters with deep learning can be effective, though efficiency is lower compared to models with more parameters.
Vaidehi Thakre et al.	Machine learning classifiers used include Random Forest, Support Vector Machine (SVM), Logistic Regression, Gaussian Naive Bayes, and K-Nearest Neighbors (KNN).	30 parameters (hormonal levels, metabolic indicators, physical characteristics, lifestyle factors, etc.)	90.09%	Inclusion of a wider range of parameters and the use of diverse classifiers contributed to higher accuracy in predicting PCOS.
Wan Azani et al.	Machine learning algorithms	43 parameters (comprehensive health and lifestyle indicators)	90.70%	The comprehensive approach with a large number of parameters resulted in the highest prediction accuracy.
Current Study	Focused on reduced set of parameters	10 parameters (menstrual cycle length, AMH levels, hair growth, weight gain, fast food consumption, pimples, skin darkening, follicle count, insulin levels)	85-88%	Demonstrated that high model efficiency can be achieved with fewer parameters if they are carefully selected, reducing the complexity and dimensionality of the data.

DISCUSSION

This research aimed to explore the intricate relationship between various health indicators and the presence of Polycystic Ovary Syndrome (PCOS) in women. Leveraging a carefully curated dataset from www.kaggle.com, encompassing factors such as menstrual cycle length, Anti-Mullerian Hormone (AMH) levels, weight gain, and other pertinent variables, a multiple linear regression model was employed to predict PCOS values.

The choice of the linear regression model was rooted in its simplicity and interpretability, offering valuable insights into the significance of each predictor variable. The model was trained on 80% of the dataset and tested on the remaining 20%, with

Mean Squared Error (MSE) and R-squared serving as primary evaluation metrics. The resultant MSE of 0.04 indicated a relatively small error, indicative of a well-performing model.

The multiple linear regression equation, with coefficients determined through training, illustrated the relationship between the independent variables and the dependent variable (PCOS values). The discussion of the research findings encompasses an in-depth analysis of the statistical measures and model performance metrics pertaining to the multi-linear regression model utilized in predicting Polycystic Ovary Syndrome (PCOS) values based on various health indicators. The examination of basic statistics for both actual and predicted data during both the training and testing periods provides valuable insights into the model's performance and predictive accuracy.

The statistical analysis reveals several noteworthy observations regarding the PCOS prediction model. Firstly, the mean PCOS values for both actual and predicted data remain relatively consistent across both the training and testing periods. Furthermore, the standard deviation and variance statistics are found. While the standard deviation measures the spread of data within each dataset, the variance quantifies the variability of PCOS values. The comparable variances between actual and predicted PCOS values in both the training and testing periods indicate that the model accurately captures the variability observed in the dataset.

The assessment of model performance through Mean Squared Error (MSE) and R-squared metrics provides additional insights into the predictive accuracy and explanatory power of the regression model. The relatively low MSE values for both the training (0.03) and testing (0.04) periods indicate minimal errors in predicting PCOS values, suggesting a high level of accuracy in the model's predictions.

Moreover, the coefficient of determination (R-squared) metrics, which measure the proportion of variance in PCOS values explained by the independent variables, further validate the effectiveness of the regression model. With R-squared values of 0.88 for the training data set and 0.85 for the testing data set, the model demonstrates a strong fit to the data, indicating that approximately 88% and 85% of the variability in PCOS values, respectively, is accounted for by the selected health indicators.

The explanation of variance and MSE elucidates key aspects of model performance and predictive accuracy. The variance statistics highlight the spread of data points around the mean and provide insights into the variability of both observed and predicted PCOS values. Meanwhile, the MSE metrics quantify the average squared difference between predicted and actual PCOS values, serving as a measure of predictive accuracy. Correlation matrices for both the training and testing datasets have been studied. Additionally, the calculation of model efficiency further confirms the effectiveness of the regression model in capturing the variability of PCOS values, with efficiency percentages of 88% for the training data set and 85% for the testing data set.

The visual representation of the model's performance, presented in Figure 3 and Figure 4 as a bar graph comparing Actual and Predicted PCOS values, provided an intuitive and complementary assessment. This visual aid, coupled with the quantitative metrics, offered a comprehensive understanding of the model's strengths and generalization to unseen data.

In summary, the combination of rigorous statistical analyses, evaluation metrics, and visual representations in this research contributes to a robust understanding of the multiple linear regression model's behavior and its applicability in predicting PCOS values based on various health indicators. The findings not only enhance our knowledge of the intricate interplay between health factors and PCOS but also provide a foundation for further research and clinical applications in the field.

CONCLUSION

The study concludes that a number of health markers, such as the length of the menstrual cycle, hormone levels, and lifestyle choices, are important predictors of PCOS. These interactions are effectively captured by the multi-linear regression model, which also offers insightful information about the relative importance of each variable. The research employs a well-established linear regression model, allowing for straightforward interpretation of the relationships between variables. In conclusion, the discussion of statistical measures and model performance metrics provides a comprehensive evaluation of the multi-linear regression model's effectiveness in predicting PCOS values based on various health indicators. The findings highlight the model's accuracy, reliability, and ability to capture the complexity of PCOS, thus offering valuable insights for clinical applications.

Future research could benefit from incorporating more diverse datasets to enhance the external validity of the findings and exploring alternative machine learning algorithms may provide a more understanding of the relationships within the data. Overall, the research contributes valuable insights into the predictive factors of PCOS but should be considered as part of a broader exploration of the complex nature of this health condition.

References

1. *S. Singh et al., "Polycystic Ovary Syndrome: Etiology, Current Management, and Future Therapeutics," Journal of Clinical Medicine, vol. 12, no. 4. MDPI, Feb. 01, 2023. doi: 10.3390/jcm12041454.*
2. *K. Motlagh Asghari et al., "Burden of polycystic ovary syndrome in the Middle East and North Africa region, 1990–2019," Sci. Rep., vol. 12, no. 1, pp. 1–11, Dec. 2022, doi: 10.1038/s41598-022-11006-0.*
3. *"Polycystic ovary syndrome." https://www.who.int/news-room/fact-sheets/detail/polycystic-ovary-syndrome (accessed Apr. 15, 2024).*
4. *M. D. Bharali, R. Rajendran, J. Goswami, K. Singal, and V. Rajendran, "Prevalence of Polycystic Ovarian Syndrome in India: A Systematic Review and Meta-Analysis," Cureus, vol. 14, no. 12, Dec. 2022, doi: 10.7759/cureus.32351.*
5. *C. Kitzinger and J. Willmott, "'The thief of womanhood': Women's experience of polycystic ovarian syndrome," Soc. Sci. Med., vol. 54, no. 3, pp. 349–361, Feb. 2002, doi: 10.1016/S0277-9536(01)00034-X.*
6. *R. Pasquali et al., "PCOS Forum: research in polycystic ovary syndrome today and tomorrow," Clin. Endocrinol. (Oxf)., vol. 74, no. 4, pp. 424–433, Apr. 2011, doi: 10.1111/j.1365-2265.2010.03956.x.*
7. *K. W. Kim, "Unravelling polycystic ovary syndrome and its comorbidities," Journal of Obesity and Metabolic Syndrome, vol. 30, no. 3. Korean Society for the Study of Obesity, pp. 209–221, 2021. doi: 10.7570/JOMES21043.*
8. *H. Teede, A. Deeks, and L. Moran, "Polycystic ovary syndrome: A complex condition with psychological, reproductive and metabolic manifestations that impacts on health across the lifespan," BMC Medicine, vol. 8. BMC Med, Jun. 30, 2010. doi: 10.1186/1741-7015-8-41.*
9. *K. Williamson, A. J. Gunn, N. Johnson, and S. R. Milsom, "The impact of ethnicity on the presentation of polycystic*

ovarian syndrome," Aust. New Zeal. J. Obstet. Gynaecol., vol. 41, no. 2, pp. 202–206, Jan. 2001, doi: 10.1111/j.1479-828X.2001.tb01210.x.

10. *U. A. Ndefo, A. Eaton, and M. R. Green, "Polycystic ovary syndrome: A review of treatment options with a focus on pharmacological approaches," P T, vol. 38, no. 6, pp. 336–355, Jun. 2013.*
11. *K. M. Hoeger, A. Dokras, and T. Piltonen, "Update on PCOS: Consequences, Challenges, and Guiding Treatment," J. Clin. Endocrinol. Metab., vol. 106, no. 3, pp. e1071–e1083, Mar. 2021, doi: 10.1210/clinem/dgaa839.*
12. *H. Xu et al., "A Model for Predicting Polycystic Ovary Syndrome Using Serum AMH, Menstrual Cycle Length, Body Mass Index and Serum Androstenedione in Chinese Reproductive Aged Population: A Retrospective Cohort Study," Front. Endocrinol. (Lausanne)., vol. 13, Mar. 2022, doi: 10.3389/fendo.2022.821368.*
13. *Z. Zad et al., "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records.," Front. Endocrinol. (Lausanne)., vol. 15, p. 1298628, 2024, doi: 10.3389/fendo.2024.1298628.*
14. *C. Neto, M. Silva, M. Fernandes, D. Ferreira, and J. Machado, "Prediction Models for Polycystic Ovary Syndrome Using Data Mining," in Advances in Intelligent Systems and Computing, 2021, vol. 1352, pp. 210–221. doi: 10.1007/978-3-030-71782-7_19.*
15. *A. Alamoudi et al., "A Deep Learning Fusion Approach to Diagnosis the Polycystic Ovary Syndrome (PCOS)," Appl. Comput. Intell. Soft Comput., vol. 2023, 2023, doi: 10.1155/2023/9686697.*
16. *S. Sreejith, H. Khanna Nehemiah, and A. Kannan, "A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier," Healthc. Anal., vol. 2, p. 100102, Nov. 2022, doi: 10.1016/j.health.2022.100102.*
17. *A. Zigarelli, Z. Jia, and H. Lee, "Machine-Aided Self-diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study," JMIR Form. Res., vol. 6, no. 3, Mar. 2022, doi: 10.2196/29967.*
18. *H. Habehh and S. Gohel, "Machine Learning in Healthcare," Curr. Genomics, vol. 22, no. 4, pp. 291–300, Jul. 2021, doi: 10.2174/1389202922666210705124359.*
19. *"PCOS Dataset." https://www.kaggle.com/datasets/shreyasvedpathak/pcos-dataset (accessed Apr. 15, 2024).*
20. *Vedpathak, S., & Thakre, V. S. (2020, December). "PCOcare: PCOS detection and prediction using machine learning algorithms". Bioscience Biotechnology Research Communications, 13(14), 56.*
21. *Mustafa, W. A., Alkhayyat, A., Al-Azzawi, W., & Alquran, H. H. (2022, September). "Detection of Polycystic Ovary Syndrome (PCOS) using machine learning algorithms". Presented at the International Conference on Emerging Technologies and Intelligent Systems (ICETIS).*