International Journal of Electronics and Communication Engineering (IJECE)
ISSN(P): 2278-9901; ISSN(E): 2278-991X
Vol. 7, Issue 1, Dec–Jan 2018; 9–16

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

MACHINE LEARNING-BASED PREDICTION OF POLYCYSTIC OVARY SYNDROME

Pavithra M J[1*], Nirmala[2] & Rekha S Kambli[3]

*[*1, 2]Lecturer, Department of Electronics and Communication Engineering, Government Polytechnic K R Pet, INDIA*

[3]Lecturer, Department of Computer Science and Engineering, Government Polytechnic K R Pet, INDIA

ABSTRACT

Polycystic ovary syndrome (PCOS) is a common hormonal disorder in women of reproductive e age. It is very difficult to fix the exact cause for PCOS. Timely diagnosis and treatment may reduce the risk of long-term complications of PCOS. A prediction model for the diagnosis of PCOS using a Category based feature selection method is proposed, which assist doctors in the diagnosis of PCOS with few available features. The data set used for the study was taken from KAGGLE and the major prediction method used are Logistic regression, Neural Networks and Random forest, the highest accuracy of 91% is obtained using logistic regression with ultra sound and symptom based attributes.

KEYWORDS: *Polycystic Ovary Syndrome .Machine Learning, Logistic Regression, Normalization, Neural Networks, Random Forest*

Article History

Received: 16 Jan 2018 | Revised: 24 Jan 2018 | Accepted: 31 Jan 2018

INTRODUCTION

It is now well accepted that diet and nutrition have a significant impact on an individual's health during the past century. Because of the numerous changes a woman's body experiences throughout pregnancy, diet and nutrition become crucial. If the woman suffers from polycystic ovarian syndrome, the problem becomes even more complex. The primary cause of PCOS is the overproduction of hormones linked to male sex, such as testosterone and androgens. PCOS upsets the equilibrium of luteinizing hormone (LH) and follicle-stimulating hormone (FSH) [1].

Pregnant women with PCOS and their babies are exposed to several factors during the time between ovulation and live birth. Metabolic, inflammatory, and hormonal changes and obesity have unfavourable effects on ovulatory function, oocyte quality and endometrial receptivity, thus complicating the initiation and progress of pregnancy. Recent studies demonstrate an association between PCOS and increased rates of unfavourable pregnancy and birth outcomes [2, 3]

So, the precise diagnosis of PCOS is very essential. These days the diagnosis made by doctors is to manually count the number of follicular cysts in the ovary, which is used to arbiter whether the PCOS exists or not. The most important work of the PCOS diagnosis is to accurately count the number of follicular cysts. However, the manual way may lead to problems of variability, reproducibility, and low efficiency. Automated detection of PCOS will overcome these problems.

This paper aims to provide an overview of research on the use of machine learning techniques in PCOS and the relationship between PCOS and gestational diabetes mellitus. Additionally, we suggested a prediction model in the research for PCOS prediction.

LITERATURE SURVEY

Meena and colleagues have examined and provided a survey report regarding PCOS and data mining methodologies. They have offered the notion that a hybrid method may be used to establish an accurate classification for the positive aspects driving polycystic ovarian syndrome. Additionally, they have said that data mining approaches address the actual causes of the PCOS issue [1]. In addition to decision tree classification strategies like ID3 and J48, they have also suggested a novel feature selection methodology called information gain. The neural fuzzy rough subset evaluation method combined with attribute selection yields superior results [2].

The most effective method is to use data mining techniques to detect the classification accuracy and predict PCOS in its early phases. This leads to the traits being filtered, which ultimately produces the best outcomes. When these strategies are applied together instead of one at a time, this method works [3].

M. Alotaibi,A. Alsinan,[4]did a case study on PCOS in Gulf Countries. They presented the system using mobile health care technologies. Analysis has been done using statistical software SPSS.Sandy Rihana et al,[5] proposed an automated cysts detection and classification in the ovary using multiscale morphological and SVM.

In order to categorize patients into normal and PCOS-affected groups, Sharvari S. et al. ,[6] Suggested an automated method of PCOS identification that counts the number of follicles in an ultrasonography of the ovary and incorporates clinical and biochemical imaging markers. The multiscale morphological technique is used to extract features, and the support vector (SVM) algorithm is used to classify all of the data.Yinhui Deng et al [7], introduced filtering method named adaptive morphological filter and watershed algorithm to extract outlines of targets, a clustering method is applied to identify expected follicular cysts.Palak Mehrotra et al,[8] written a paper in Automated Ovarian Follicle Recognition for Polycystic Ovary Syndrome used the techniques like multiscale morphological approach for contrast Enhancement and scanline thresholding is used to extract the contours of the follicles.

S. Sheela et al[8], mentioned about theoretical investigations on PCOS. They used Wiener filters for the removal of speckle noise, extraction of region of interest using segmentation, classify images in maximum accuracy to detect ovarian cyst in short time. Bedy Purnama et al [9],did a classification of polycystic Ovary Syndrome based on follicle detection of ultrasound images by using the methodology feature extraction using Gabor wavelet. The images of follicle are categorized into two groups of texture features as dataset A and dataset B[9].

A research on PCOS prediction using menstrual type was proposed by S. Rethinavalli et al[10] ,the original dataset had 32 attributes; feature selection was used in the pre-processing step to limit the attributes to 7. The ANN algorithm and NFRS (Neural Fuzzy Rough Set) are used to compare PCO syndrome and non-PCO syndrome. The relationship between menstrual types and PCOS condition was demonstrated by the chi-square test.

METHODOLOGY

Data Set Used

The study's dataset, sourced from the KAGGLE repository platform, formed the backbone of its analysis, comprising 541 instances and 42 features or attributes, inclusive of a pivotal class attribute. Within this dataset, a demographic of females was represented, totaling 364 instances categorized as non-PCOS cases and 177 instances identified as PCOS cases.

Central to the predictive analysis were three overarching characteristics: physical, clinical, and hormonal, each providing unique insights into the dataset. Comprehensively, the dataset encompassed 32 continuous attributes, allowing for nuanced examination, alongside 9 categorical attributes, enriching the breadth of analysis with distinct categorizations.

To enhance clarity and facilitate meaningful analysis, the attributes within the dataset were systematically organized into groups, as delineated in Table 1 of the study. This systematic arrangement allowed for a structured understanding of the dataset, grouping attributes based on the methodology employed to ascertain their values. Such meticulous organization was fundamental in streamlining the analytical process, enabling researchers to navigate the dataset effectively and derive actionable insights.

Table 1: Attributes Groups

Hormonal:	Follicle Stimulating Hormone Luteinizing Hormone Thyroid Stimulating Hormone Anti-mullerian hormone Progesterone Prolactin
Physical/Metabolic	Age (yrs) Weight (Kg) Height(Cm) BMI Blood Group Fast food Reg.Exercise Marriage Status No. of aborptions Hip(inch) Waist(inch) Waist:Hip Ratio Cycle length(days
Ultrasound	Follicle No. (L) Follicle No. (R) Avg. Follicle size (L) (mm) Avg. Follicle size (R) (mm) Endometrium (mm)
Lab Tests Related	BP _Systolic (mmHg) BP _Diastolic (mmHg) Pulse rate(bpm) RR (breaths/min) Haemoglobin I beta-HCG(mIU/mL) II beta-HCG(mIU/mL) Vit D3 (ng/mL)
Signs/Symptoms Related	Weight gain hair growth Skin darkening Hair loss Pimples

Feature Selection

In navigating the expansive landscape of dimensional datasets, selecting the most pertinent features for classification emerges as a pivotal undertaking in predictive analytics. In this study, a meticulously crafted feature selection methodology was proposed, tailored to distill the essence of the dataset and optimize classification accuracy.

The cornerstone of this approach lies in the adoption of a category-based feature selection method, which offers a systematic framework for discerning relevant features. Within the dataset sourced from KAGGLE, comprising a comprehensive array of 42 features, a strategic categorization scheme was devised to delineate features into six distinct groups based on the manner in which their values were derived.

- **Category-I:** Ultrasound Related Attributes: Encompassing features solely derived from ultrasound data, this category provides insights into the anatomical and morphological aspects relevant to the diagnosis of PCOS.
- **Category-II:** Symptom-Based Attributes: Focused exclusively on features derived from symptomatic manifestations, this category sheds light on the clinical presentation and subjective experiences associated with PCOS.
- **Category-III:** Ultrasound and Symptom-Based Attributes: Combining ultrasound and symptom-based features, this category offers a comprehensive perspective by integrating both structural and clinical insights.
- **Category-IV:** Hormonal Attributes: Centered on features pertaining to hormonal profiles, this category delves into the endocrine dysregulations characteristic of PCOS, offering valuable biomarkers for diagnosis and prognosis.
- **Category-V:** Hormonal and Lab Test Related Attributes: Expanding beyond hormonal markers, this category incorporates additional laboratory-derived parameters, enriching the diagnostic landscape with quantitative metrics.
- **Category-VI:** Hormonal, Physical/Metabolic Attributes, and Lab Test Related Attributes: The most comprehensive category, amalgamating hormonal, physical, metabolic attributes, and laboratory parameters, this grouping provides a holistic panorama of PCOS, encompassing multifaceted dimensions spanning endocrinology, physiology, and pathology.

By delineating features into these distinct categories, the feature selection process gains clarity and precision, facilitating the identification of salient predictors for PCOS classification. This methodological framework not only streamlines the feature selection process but also enhances the interpretability and generalizability of the predictive model, fostering robust insights into the complex interplay of factors underlying PCOS.

Normalization

Normalization, a fundamental data preprocessing technique, aims to standardize the scales of attributes within the dataset, rendering them comparable and alleviating the disproportionate influence of attributes with larger scales on distance-based algorithms. By rescaling attribute values to a common scale, typically within the range of 0 to 1, normalization fosters uniformity in the treatment of attributes, enhancing the stability and efficacy of subsequent predictive modeling techniques.

Classification

Following normalization, a suite of popular prediction techniques is employed to prognosticate PCOS, encompassing a diverse array of methodologies tailored to discern patterns and relationships within the dataset:

- **Random Forest:** Leveraging the power of ensemble learning, Random Forest constructs a multitude of decision trees and amalgamates their predictions to yield robust and accurate classifications. Renowned for its resilience to overfitting and versatility across diverse datasets, Random Forest excels in capturing intricate relationships within the data, making it a formidable contender for PCOS prediction.
- **Naive Bayes:** Grounded in the principles of Bayesian probability, Naive Bayes offers a probabilistic framework for classification, predicated on the assumption of feature independence. Despite its simplistic assumptions, Naive Bayes often yields competitive performance, particularly in scenarios characterized by high-dimensional datasets and categorical attributes.
- **Logistic Regression:** A stalwart in the realm of statistical modeling, Logistic Regression furnishes a linear model for binary classification, characterizing the relationship between input attributes and the probability of PCOS occurrence. Renowned for its interpretability and computational efficiency, Logistic Regression remains a cornerstone of predictive modeling endeavors.
- **k-Nearest Neighbor (k-NN):** Embracing the ethos of instance-based learning, k-NN discerns the class label of a given instance by aggregating the labels of its nearest neighbors in the feature space. Esteemed for its simplicity and intuitive appeal, k-NN offers a non-parametric approach to classification, adept at capturing local patterns and adapting to the inherent intricacies of the dataset.
- **Neural Network:** Unleashing the power of deep learning, Neural Networks orchestrate intricate architectures of interconnected nodes to discern complex patterns and relationships within the data. Endowed with the capacity to learn hierarchical representations, Neural Networks excel in capturing nonlinear dependencies and extracting latent features, rendering them indispensable tools for predictive modeling endeavors.

By harnessing the collective prowess of these diverse prediction techniques, the study endeavors to unravel the intricate tapestry of factors underpinning PCOS, empowering clinicians and researchers with actionable insights to enhance diagnostic accuracy and inform therapeutic interventions.

RESULT AND DISCUSSION

In the experimental section of the study, a rigorous evaluation of classification models is conducted on both raw and normalized data, illuminating the impact of data preprocessing on predictive performance. Initially, the classifiers are deployed on the raw data, thereby providing a baseline assessment of their efficacy in discerning patterns within the dataset. Subsequently, the same classifiers are applied to the normalized data, thereby elucidating the extent to which normalization mitigates the influence of disparate attribute scales and enhances predictive accuracy. The comparative performances of the classifiers on raw and normalized data are systematically tabulated in Table 2 and Table 3, respectively, furnishing a comprehensive overview of the efficacy of each classification model across different data preprocessing paradigms. Building upon this foundational analysis, the study delves deeper into the realm of feature selection, recognizing the pivotal role of feature subsets in elucidating salient predictors and enhancing predictive accuracy. Employing a repertoire of feature selection techniques, the dataset is scrutinized through diverse lenses, each tailored to accentuate distinct subsets of features deemed most informative for PCOS classification.

Crucially, the study endeavors to delineate the impact of feature selection across different categories of attributes, leveraging insights garnered from the category-based feature grouping outlined earlier. Through meticulous experimentation, classification models are applied to subsets of features derived from distinct categories, thereby illuminating the differential impact of feature subsets on predictive accuracy across various dimensions of the dataset. The culmination of these endeavors is encapsulated in figure 1, where the accuracy scores obtained for different sets of features, stratified by category, are meticulously cataloged. This tabulation affords a granular understanding of the predictive efficacy conferred by feature subsets derived from different categories, elucidating the relative importance of various dimensions of the dataset in informing PCOS classification.

By systematically unraveling the interplay between data preprocessing, feature selection, and classification performance, the study engenders a nuanced comprehension of the underlying factors governing PCOS classification. Armed with these insights, clinicians and researchers are empowered to navigate the intricacies of PCOS diagnosis with precision and efficacy, thereby advancing the frontiers of medical knowledge and enhancing patient care.

Table 2: Performances of the Classifiers

Model	AUC	CA	F1	Precision	Recall
kNN	0.659	0.682	0.604	0.640	0.682
Random Forest	0.927	0.879	0.876	0.879	0.879
Neural Network	0.935	0.866	0.865	0.865	0.866
Naive Bayes	0.939	0.870	0.870	0.871	0.870
Logistic Regression	0.934	0.879	0.879	0.878	0.879

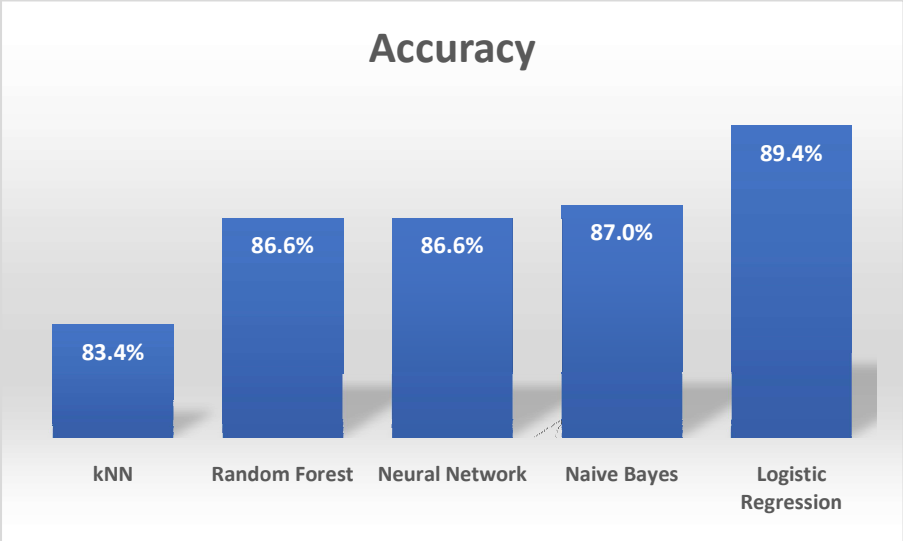

Figure 1: Accuracy of the Classifiers for the Normalized Data.

Table 3: Accuracy of the Classifiers for Different Category Based Features

Model	Category-I	Category-II	Category-III	Category-IV	Category-V	Category-VI
kNN	0.84	0.84	0.85	0.67	0.70	0.72
Random Forest	0.83	0.84	0.88	0.68	0.68	0.71
Neural Network	0.85	0.83	0.89	0.70	0.68	0.71
Naive Bayes	0.84	0.83	0.87	0.67	0.66	0.74
Logistic Regression	0.85	0.83	**0.91**	0.69	0.69	0.74

It is challenging to diagnose PCOS using a particular test or set of procedures; in addition to the patient's medical history, a number of other criteria must be taken into account. Because PCOS is difficult to diagnose, the suggested prediction model uses the limited features that are available to help doctors make the right decision. It is evident from the result that greater than 90% accuracy can be achieved when employing ultrasound results or symptoms/signs related qualities.

CONCLUSION

Our study introduces a predictive model tailored for PCOS diagnosis, leveraging diverse feature selection techniques. Through the application of wrapper and correlation-based methods, we explore the efficacy of various classifiers, with logistic regression emerging as the top performer, achieving a notable accuracy rate of 90%.

Furthermore, our investigation reveals that focusing solely on ultrasound findings or symptom-related attributes yields promising results, with an accuracy of approximately 80%. This underscores the potential for targeted diagnostic approaches based on specific diagnostic indicators.

Moreover, our research delves into the relationship between gestational diabetes mellitus (GDM) and PCOS. Extensive literature supports this association, with multiple studies elucidating the interplay between these conditions. By examining this link, we contribute to the growing body of knowledge surrounding PCOS and its comorbidities, shedding light on potential implications for clinical practice and further research endeavors.

REFERENCES

1. *K. Meena, M. Manimekalai, and S. Rethinavalli, "A Literature Review on Polycystic Ovarian Syndrome and Data Mining Techniques," Int. Jouranl Adv. Res. Comput. Sci. Softw. Eng, vol. 4, pp. 780-785, 2014.*
2. *D. K. Meena, D. M. Manimekalai, and S. Rethinavalli, "A Novel Framework for Filtering the PCOS Attributes using Data Mining Techniques," International Journal of Engineering Research & Technology (IJERT), vol. 4, pp. 702-706, 2015.*
3. *K. Meena, M. Manimekalai, and S. Rethinavalli, "Correlation of Artificial Neural Network Classification and NFRS Attribute Filtering Algorithm for PCOS Data," Int. J. Res. Eng. Technol, vol. 4, pp. 519-524, 2015.*
4. *M. Alotaibi and A. Alsinan, "A mobile Polycystic ovarian syndrome management and awareness system for Gulf countries: System architecture," in 2016 SAI Computing Conference (SAI), 2016, pp. 1164-1167.*
5. *S. Rihana, H. Moussallem, C. Skaf, and C. Yaacoub, "Automated algorithm for ovarian cysts detection in ultrasonogram," in 2013 2nd International Conference on Advances in Biomedical Engineering, 2013, pp. 219-222.*
6. *S. S. Deshpande and A. Wakankar, "Automated detection of polycystic ovarian syndrome using follicle recognition," in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014, pp. 1341-1346.*
7. *Y. Deng, Y. Wang, and P. Chen, "Automated detection of polycystic ovary syndrome from ultrasound images," in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, pp. 4772-4775.*

8. *P. Mehrotra, C. Chakraborty, B. Ghoshdastidar, S. Ghoshdastidar, and K. Ghoshdastidar, "Automated ovarian follicle recognition for polycystic ovary syndrome," in 2011 International Conference on Image Information Processing, 2011, pp. 1-4.*

9. *B. Purnama, U. N. Wisesti, F. Nhita, A. Gayatri, and T. Mutiah, "A classification of polycystic Ovary Syndrome based on follicle detection of ultrasound images," in 2015 3rd International Conference on Information and Communication Technology (ICoICT), 2015, pp. 396-401.*

10. *S. Rethinavalli and M. Manimekalai, "A Hypothesis Analysis on the Proposed Methodology for Prediction of Polycystic Ovarian Syndrome," International Journal of Science, Engineering and Computer Technology, vol. 6, p. 396, 2016.*