An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOS

Pijush Dutta (✉ pijushdutta009@gmail.com)  
Global Institute of Management Technology  https://orcid.org/0000-0001-5144-315X

Shobhandeb Paul  
Gurunanak Institute of Technology

Madhurima Majumder  
Mirmadan Mohanlal Government Polytechnic

Research Article

Keywords: PCOS, Machine learning, Classification, Logistics regression, Decision Tree, Random Forest, Support Vector Machine & K Nearest neighbour

Posted Date: November 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1043852/v1

License: ☺️ 📧 This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Polycystic ovary syndrome (PCOS) is a well known catabolism based and reproductive endocrinopathies disease which affects up to 18% of women in reproductive age. PCOS results irregularity in menstrual periods which might directed to numerous entanglements such as embodiment of cysts (follicles) in either or both ovaries as well as impotence for giving birth, turmoil in androgen levels, increased BMI and some other deviancy in hormone level of LH, FSH, DHEAS, Fasting insulin and Fasting blood sugar. Detection & prevention of this disease as early as possible it has attracted research attention. Overall research in this therapeutic unhealthiness categorisation using machine learning technique is considered less to some extent because of data imbalance. In this research we can make a lay on the line for a novel forecast model using Synthetic Minority Oversampling Technique (SMOTE) with five machine learning algorithms like Logistic Regression, Random Forest, Decision Tree Support vector machine and K-Nearest Neighbor (KNN) model which automates the PCOS detection in an early stage. The framework of the recommended classification of PCOS is comprised of three different layers. The first flapping which is responsible for handling the missing values & oversampling the splinter class in datasets. In second layer, the most eloquent attributes are tabbed using Principal component analysis (PCA). In third layer, the projected model is skilled & its efficiency is reckoned in a fourth layer in terms of classification accuracy, Training time (TT), F1 score, Recall (Sensitivity), Precision & Area under the ROC (AUROC). The propounded algorithm clarify the outperformed of its corresponded & achieves the preeminent accuracy, precision, F1 score, Sensitivity, training time & AUROC. The best model achieve accuracy, Training time (TT), F1 score, Recall (Sensitivity), Precision & Area under the ROC (AUROC) is 97.11%, 0.010 sec, 98%, 98%, 98% & 95.6% respectively.

1. Introduction

Polycystic Ovary Disease (PCOD) or Polycystic Ovary Syndrome (PCOS) is a hormonal lopsidedness issue found in women of reproductive age. In today's generation, there has been a constant rise in the occurrence of this disease due to the hindered lifestyle routines and adulterated food. There are three types of ovaries and is classified as normal ovary, cystic ovary and polycystic ovary.(Child et al., 2001) Poly Cystic Ovarian Syndrome or PCOS is a complex hormonal plights distressing up to around 1 in every 10 women at their concepitive age. PCOS manifests during adolescence and is formed as a result of hormonal disturbances (Witchel et al., 2015). Peripherally inside the ovary, fluid-filled sacs are present which are called follicles or cysts. A polycystic ovary (PCO) can be characterized by twelve or more follicles with a diameter of 2-9 mm(Jarrett et al., 2020). PCOS affects both health and the quality of women's life. The symptoms include cardiovascular diseases, failure to ovulate and infertility, late menopause, type 2 diabetes, acne, baldness, hair loss, hirsutism, obesity, anxiety, depression, and stress(Khomami et al., 2015).

The early diagnosis and treatment can be used to control based on the symptoms and by the prevention of long-term problems. PCOS can be detected through ultrasonography by a doctor by reckoning the number and size of follicles situated in the ovaries. However, this process takes a protracted interval, need
good image quality and high accuracy to detect the presence of PCOS (Balen et al., 2003). Another approach for PCOS detection is through biochemical parameters such as hormone levels examination. Since hormone examination is very expensive, other clinical parameters such as body mass index (BMI), menstrual cycle length, etc. are taken into consideration for the detection of PCOS (Ranjzad et al., 2011).

Among the plentiful problems that subsisted around us, the problems that are pertinent to the conceptive health of women was chosen as an area of our relevance, due to its materiality in this contemporary society. A exhaustive survey of studies on PCOS and systems for its supportive diagnosis was carried out. about 5-10% of Indian women in reproductive age are attained by the multifaceted endocrine disorder called Polycystic Ovary Syndrome (PCOS)(Denny et al., 2019). It is a supreme instigation of an ovulatory infertility and increases the risk for insulin resistance, obesity, cardiovascular disease and psychosocial irregularities (Pauli et al., 2011). The symptoms for PCOS might be divaricate from patient to patient. Some of them are inconsistency in menstrual periods, acne, overweight, increased tendency for infertility, intense hair fall, balding of front head, increased growth of facial hair(Ding et al., 2017). Traditionally the PCOS can be questionable when number of follicles in an ovary is more than 12 per unit area and visible in radiological scan (Kenigsberg et al., 2015). Few recent studies are preempted in elemental research direction, scrutinising the affiliated aspects such as obesity(Kalantar-Zadeh et al., 2005) and genetic factors(McCrory et al., 2010). In recent years, machine learning (ML) classification and feature selection algorithms have been used by researchers and clinicians for the prediction of diseases as a non-invasive method (Dutta et al., 2021). PCOS datasets which consist of heterogeneous attributes related to biochemical, clinical, medical history, symptoms of the patients and ultrasound images are used to build predictive models (Vagios et al., 2021). Various machine learning model used to detection & prediction of PCOS (Boomidevi & Usha, 2021). SPSS V 22.0 software used to identify the potential attributes out of 23. The best model Random Forest achieves maximum accuracy of 89.02%. An effective classification performance for PCOS, datasets are resample using a combination of SMOTE(Synthetic Minority Oversampling Techniques) & ENN (Edited Nearest Neighbor) (Khan Inan et al., 2021).From the experimental results it was seen that for a 10 fold cross validation Extreme Gradient Boosting classifier outperformed than all other classifiers. A system was proposed to classify PCOS on ultrasound images using the feature extraction (Gabor Wavelet method) and Competitive Neural Network (CNN) (Dewi et al., 2018). Highest accuracy & testing time of the proposed model were 80.84% and 60.64 seconds respectively. Diagnose of PCOS data five different machine learning approaches were applied (Hassan & Mirza, 2020). From the Result analysis it was seen that maximum accuracy obtain by random Forest is 96%. A novel hybrid structure to determine the chances of PCOS that coalesce navies Bayes and artificial neural network algorithm to produce the best result (Thomas & Kavitha, 2020). A comparative study was performed on two different tool Rapid Miner & Python, from the result analysis it was seen that Random Forest outperformed than the other classification when it was utilized in Rapid Miner (Satish et al., 2020).

Four classification algorithms and five feature selection methods for PCOS dataset was performed to predict the disease(Boomidevi & Usha, 2021; Sumathi et al., 2021; Wagha et al., 2020).
In this paper, our objective is to know the better performance machine learning classification algorithm after applying the Principle Component Analysis (PCA) & Synthetic Minority Oversampling Technique (SMOTE). The left over paper has been collaborated in the following way Section 2 emphasises the proposed framework which is being propounded for the coordination of the given dataset. Section 3 bouncing off and sorts out classification results that are handed for getting a result of the application of the proposed methodology. Section 4 puts the lid on for the findings of the proposed work in an trenchant way.

2. Proposed Framework

The data are pre-processed for mislaid values and then regimented using the standard scalar technique. Ensuingly, most significant attributes are selected using Principal component analysis (PCA) enduring within the dataset. Additionally, various machine learning algorithms are attained to the given dataset, and their results are estimated. The main aspire was to button down the algorithm which can distinguish the given dataset in the best manner. The following section wholly exchange the view of the proposed framework and its components in detail.

2.1 Datasets

The dataset which has been used is the UCI PCOS without fertility dataset. The UCI dataset reposed of 541 records having 41 attributes. Most published research works have used subset of 12 attributes. The main discourse for excerpting above mentioned attributes is that these attributes were mulled over the most decisive while vaticinating PCOS for a particular patient. These attributes are distinctively recognised by UCI repository while promulgating dataset for public. Sample datasets for infertility PCOS shown in Figure 1. Attributes characteristics are shown in Whisker plot (Figure 2). These attributes along with their interrelated values show how a itemised attributes can be associated to PCOS. Among these 12 attributes, 11 attributes are used in prediction of PCOS disease. One attribute “Target” dishes up as output variable whose value resolve the subsistence or truancy of PCOS. Data pre-processing is brought about on given dataset before the pertinence of classification models. The main steps are attained for data pre-processing were as follows: the data were normalized between 0 and 1 to escalation the performance of a model. Left behind values were redeemed with mean values for corresponding columns. The output class “target” was shift gear from multiclass into binary class where 1 exemplify presence of PCOS and 0 represents absence of PCOS.

2.2 Imbalance datasets

The given dataset contains 178 instances of the positive class (1) and 363 instances of the negative class (0). There is an mismatched distribution of classes within the dataset. This uneven distribution is one of the considerable causes of curtailed accuracy of classification models. The main reason is that most machine learning models cannot learn patterns for both positive and negative classes persuasively because of their imbalance numbers in a dataset. Moreover, as minority class, i.e., positive class is less in number, so results procreated by this class often become abortive because of its less number. Most
literature studies do not keep document minority class contribution toward producing overall classification results. One of the key contributions of the proposed work is the unbalanced nature of the given dataset is managed adequately via SMOTE technique. Likewise, results for the majority and minority classes are chronicled separately in order to excavate the performance of each class’s contribution in producing overall prediction results.

2.3 Synthetic Minority over Sampling Technique (SMOTE)

SMOTE is a famous approach (Bhagat & Patil, 2015) used for the elevation of a classifier for the imbalance dataset. An imbalance dataset subsists of disparate distribution of nitty-gritty output classes. SMOTE is heavily used in classification problems of imbalance dataset (Douzas et al., 2018). SMOTE preprocessing technique is premediated as one of the most trustworthy techniques when dealing with the imbalanced dataset. SMOTE is considered to be one of the most powerful preprocessing techniques in the machine learning and information mining domain (Shuja et al., 2020). The aim of SMOTE is to endure interpolation within data of minority class samples so that their numbers can be marked up. This makes remedy in negotiating generalization in classification. In SMOTE, minority class is generally oversampled by generating artificial examples (Elreedy & Atiya, 2019). SMOTE is a very efficacious technique while tackling with imbalanced datasets. It tends to commeasure the number of majority and minority class instants in training examples.

2.4 Proposed flow Diagram

Figure 3 shows the recommended flow of experimentation where at first dataset is processed to shed out null values & button down the potential attributes by PCA. Afterward, SMOTE technique is applied to the given dataset so that an equal quantity of positive and negative examples can be secured. Moreover, at the final stage, various machine learning algorithms are enforced on the given dataset and corresponding results are obtained. Stratified K-Fold validation is applied to the given dataset to make certain of the fidelity of the results.

2.5 Tools & Techniques:

This algorithm works on Python version 3.7.6 with Jupyter notebook 6.0.3 & Operating system Ubuntu. As the dataset size was comparatively small, the required machine learning algorithms are run on the machine which is geared by Intel Core i3 6th Generation processor. It has 8 GB RAM and has 500 GB HDD storage.

2.6 Parameters tuning:

One of the crucial aspects of classification problems in machine learning is about how to tune the hyperparameters for the given classification model. These hyperparameters have a high effect on the sweeping verity and loyalty of classification results (Sun et al., 2021). Although various methods are defined for hyperparameters tuning (Hutter et al., 2015), yet so far, there is no state-of-the-art algorithm
present for hyperparameters tuning. In the proposed research work, grid search technique (Veloso et al., 2021) is fanned out while tuning the hyperparameters for different algorithms. In the grid search technique, some random values are given as hyperparameters values. Subsequently, the algorithm will find the best value among all possible values for which the best classification results are originated.

| Classification model | Parameters Tuning |
|----------------------|-------------------|
| Random Forest        | n_estimators = 100, max_depth = 16, bootstrap = true, min_samples_leaf = 0.1, min_samples_split = 2, min_leaf_nodes = 10 |
| KNN                  | n_neighbors = 10, weights = “uniform,” algorithm = “auto”, leaf_size = 30 |
| Logistic regression  | Stopping criteria-1 e – 9, bias = true, maximum iteration for convergence = 100 |
| Decision Tree        | DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=None, splitter='best') |

3. Result Analysis

The experiment is accomplished on the given dataset, and related results are secured. Stratified K-fold validation is depicted for each experiment so that results are free from any biasness. The main aspiration was to avert any biasness in the results as feature engineering often results in inattentiveness of some features which may have an impression on overall prediction results. Furthermore, the feature engineering process often ascertained to be very expensive. Raw data after some preprocessing are fed into machine learning algorithms. Afterward, results are attained and segregated with the existing state of the art systems.

There were two datasets, out of which one was PCOS data with infertility and the other was PCOS data with fertility. The dataset was analyzed and accordingly the algorithms were applied and the model was trained to get the better accuracy.

3.1 Performance Measure

One of the extensive misinterpretations regarding the assessment of the machine learning model is that every dataset can be computed with the same evaluation matrices disregarding of its nature. Most machine learning models tend to be evaluated in terms of exactness. This approach often proves to be deceitful, when we are dishing out with the cumbersome dataset. For that reason, different standard evaluation matrices are utilised along with exactness. Precision, recall, $F_1$ measure, and ROC curve (Dutta
et al., 2021) have been applied for the evaluation of the proposed work. Accuracy is the ratio of the number of accurate predictions divided by the total number of inputs. The confusion matrices are brought about by calculating true positive (TP), true negative (TN), false-positive (FP), and false negative (FN). Sensitivity and specificity are two measures that are calculated as TP/(FN + TP) and FP/(FP + TN), respectively. The receiver operation curve (ROC) is another metric that is extensively used to evaluate the classification accuracy of a given model.

**Table 2.** Average Performance measure after experimentation

| Algorithm          | Avg. Accuracy | Avg. Precision | Avg. Recall | Avg. F1 Score | Computation Time | AUC       |
|--------------------|---------------|----------------|-------------|---------------|------------------|-----------|
| SMOTE based LR     | 95.5          | 96             | 96          | 96            | 0.023            | 94.50     |
| SMOTE based DT     | 88.99         | 89             | 89          | 89            | 0.031            | 81.86     |
| SMOTE Based RF     | 88.99         | 90             | 89          | 89            | 0.010            | 93.40     |
| SMOTE Based KNN    | 89.90         | 90             | 90          | 90            | 0.015            | 95.60     |
| SMOTE based SVM    | 91.74         | 92             | 92          | 92            | 0.012            | 89.60     |

**Table 3.** Average Performance measure for Positive & negative class after experimentation

| Algorithm          | Precision Positive Class | Precision Negative Class | Recall Positive Class | Recall Negative Class | F1 Score Positive Class | F1 Score Negative Class |
|--------------------|---------------------------|--------------------------|-----------------------|-----------------------|-------------------------|-------------------------|
| SMOTE based LR     | 94.9                      | 97.1                     | 92                    | 98                    | 92                      | 98                      |
| SMOTE Based DT     | 86.8                      | 91.2                     | 86                    | 92                    | 86                      | 92                      |
| SMOTE Based RF     | 86.8                      | 91.2                     | 86                    | 92                    | 86                      | 92                      |
| SMOTE Based KNN    | 87.7                      | 92.3                     | 87                    | 93                    | 87                      | 93                      |
| SMOTE Based SVM    | 87.8                      | 94.2                     | 89                    | 95                    | 89                      | 95.5                    |

Table 2 shows the tabular form result analysis of Average Accuracy, Precision, F1 score, Precision Area under ROC & Computational Time for SMOTE based Logistic Regression, Random Forest, Decision tree, KNN & SVM models when trained and tested on the Tabular data consisting of actual 540 records. Results are obtained after principle component analysis & SMOTE. Average parameters are calculated for both negative & positive classes’ cases. We found that SMOTE based Logistic Regression performed best
among the entire SMOTE based algorithm with the Average Accuracy, precision, Recall & F1 Score values of 95.5%, 96%, 96% & 96% respectively. However the SMOTE based Random Forest and SMOTE based KNN & SVM have least computational time & maximum area under ROC with the values of 0.10 sec & 95.6% respectively.

In Table 3 we found that for negative classes SMOTE based Logistic regression performed best among the entire SMOTE based algorithm as well as positive classes case with precision, Recall & F1 Score value of 97.1%, 98% & 98% respectively.

Figure 4 shows graphical comparison for the entire SMOTE based machine learning algorithm on the basis of computational time. From the plot we found that SMOTE based Random Forest have least computational time of 0.010 sec while SMOTE based Decision Tree have a maximum time of 0.031 sec. Figure 5, we found that both SMOTE based SVM & KNN have maximum AUC with values of 95.6% while SMOTE based Decision Tree have least AUC with the value of 81.86%.

3.2 Segregation with existing system

The results of the propounded work are being make segregated with results of other state-of-the-art extant systems so that the authenticity of the proposed work can be confirmable.

| Source                      | Algorithm/ Technique                          | Maximum Accuracy |
|-----------------------------|-----------------------------------------------|------------------|
| (Denny et al., 2019)        | RF, DT, KNN, SVM, LR & Naïve Bayes.           | 89.02%           |
|                             | Out of 23 potential features, classification model is designed with the help of 8 features |                  |
| (Satish et al., 2020)       | RF, KNN & SVM                                  | 93.12%           |
|                             | Out of 24 potential features, classification model is designed with the help of 10 features |                  |
| Proposed work               | SMOTE based LR, DT, RF, KNN & SVM             | 97.11%           |

The proposed SMOTE-based five different machine learning model without feature engineering is compared with two different systems are delectably shown in recent years for the given UCI dataset. Table 4 shows the benchmark achievement of the proposed model with two extant systems. The proposed model has surpassed all existing systems with the highest accuracy of 97.11% obtain by SMOTE based Logistic regression for negative class. Furthermore, as irradiated earlier, most existing
literature only keyed on evaluating results in terms of “accuracy” which may become a disingenuous metric when compromising with the shortcoming dataset.

4. Conclusion

Polycystic Ovary Syndrome (PCOS) is one of the most familiar types of endocrine infirmity in reproductive age women. This may payoff infertility and an ovulation. The diagnostic criterion includes the clinical and metabolic parameters which are biomarker for the disease. Here to perform the research we haven 541 number of non fertility POCS datasets from UCI repository. The overall datasets contains 178 instances of the positive class (1) and 363 instances of the negative class (0). Disparated dissemination of classes within the dataset is one of the major causes of waning exactness of classification models. Hence in this research we developed a novel prediction model using Synthetic Minority Oversampling Technique (SMOTE) with five machine learning algorithms like Logistic Regression, Random Forest, Decision Tree Support vector machine and K-Nearest Neighbor (KNN) model which automates the PCOS detection in an early stage with higher degree of efficiency.

Among Five SMOTE based algorithm it is seen that proposed SMOTE based logistic Regression outperformed than other algorithms empower to classify dataset fastidiously, without any discernible data pre-processing. Furthermore, in term of execution time, SMOTE based Random Forest has taken remarkably less time with 0.10 seconds and both SMOTE based SVM & KNN has taken maximum area under ROC.

In future practice similar type disease prediction systems can be evolved for other diseases like Heart, diabetes or cancer. Moreover, IOT technology may be embedded with the proposed model so that patient’s health parameters can be incidentally call the shots for developing an compelling healthcare system.

Declarations

Compliance with Ethical Standards:

Funding: For this research authors does not get any fund.

Conflict of Interest: All the author declares that there has no conflict of interest

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Balen, A. H., Laven, J. S., Tan, S.-L., & Dewailly, D. (2003). Ultrasound assessment of the polycystic ovary: International consensus definitions. Human Reproduction Update, 9(6), 505–514.
2. Bhagat, R. C., & Patil, S. S. (2015). Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest. *2015 IEEE International Advance Computing Conference (IACC)*, 403–408. https://doi.org/10.1109/IADCC.2015.7154739

3. Boomidevi, R., & Usha, S. (2021). Performance Analysis of Polycystic Ovary Syndrome (PCOS) Detection System Using Neural Network Approach. In *Data Engineering and Communication Technology* (pp. 449–459). Springer.

4. Child, T. J., Abdul-Jalil, A. K., Gulekli, B., & Tan, S. L. (2001). In vitro maturation and fertilization of oocytes from unstimulated normal ovaries, polycystic ovaries, and women with polycystic ovary syndrome. *Fertility and Sterility*, 76(5), 936–942.

5. Denny, A., Raj, A., Ashok, A., Ram, C. M., & George, R. (2019). i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques. *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 673–678.

6. Dewi, R. M., Adiwijaya, Wisesty, U. N., & Jondri. (2018). Classification of polycystic ovary based on ultrasound images using competitive neural network. *Journal of Physics: Conference Series, 971*, 012005. https://doi.org/10.1088/1742-6596/971/1/012005

7. Ding, T., Hardiman, P. J., Petersen, I., Wang, F.-F., Qu, F., & Baio, G. (2017). The prevalence of polycystic ovary syndrome in reproductive-aged women of different ethnicity: A systematic review and meta-analysis. *Oncotarget*, 8(56), 96351.

8. Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences, 465*, 1–20. https://doi.org/10.1016/j.ins.2018.06.056

9. Dutta, P., Paul, S., & Kumar, A. (2021). Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19. In *Electronic Devices, Circuits, and Systems for Biomedical Applications* (pp. 521–540). Elsevier.

10. Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32–64.

11. Hassan, M. M., & Mirza, T. (2020). Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. *International Journal of Computer Applications, 975*, 8887.

12. Hutter, F., Lücke, J., & Schmidt-Thieme, L. (2015). Beyond manual tuning of hyperparameters. *Künstliche Intelligenz, 29*(4), 329–337.

13. Jarrett, B. Y., Vanden Brink, H., Oldfield, A. L., & Lujan, M. E. (2020). Ultrasound characterization of disordered antral follicle development in women with polycystic ovary syndrome. *The Journal of Clinical Endocrinology & Metabolism, 105*(11), e3847–e3861.

14. Kalantar-Zadeh, K., Abbott, K. C., Salahudeen, A. K., Kilpatrick, R. D., & Horwich, T. B. (2005). Survival advantages of obesity in dialysis patients. *The American Journal of Clinical Nutrition, 81*(3), 543–554.
15. Kenigsberg, L. E., Agarwal, C., Sin, S., Shifteh, K., Isasi, C. R., Crespi, R., Ivanova, J., Coupey, S. M., Heptulla, R. A., & Arens, R. (2015). Clinical utility of magnetic resonance imaging and ultrasonography for diagnosis of polycystic ovary syndrome in adolescent girls. *Fertility and Sterility, 104*(5), 1302–1309.

16. Khan Inan, M. S., Ulfath, R. E., Alam, F. I., Bappee, F. K., & Hasan, R. (2021). Improved Sampling and Feature Selection to Support Extreme Gradient Boosting For PCOS Diagnosis. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 1046–1050. https://doi.org/10.1109/CCWC51732.2021.9375994

17. Khomami, M. B., Tehrani, F. R., Hashemi, S., Farahmand, M., & Azizi, F. (2015). Of PCOS symptoms, hirsutism has the most significant impact on the quality of life of Iranian women. *PLoS One, 10*(4), e0123608.

18. McCrory, E., De Brito, S. A., & Viding, E. (2010). Research review: The neurobiology and genetics of maltreatment and adversity. *Journal of Child Psychology and Psychiatry, 51*(10), 1079–1095.

19. Pauli, J. M., Raja-Khan, N., Wu, X., & Legro, R. S. (2011). Current perspectives of insulin resistance and polycystic ovary syndrome. *Diabetic Medicine, 28*(12), 1445–1454.

20. Ranjzad, F., Mahban, A., Shemirani, A. I., Mahmoudi, T., Vahedi, M., Nikzamir, A., & Zali, M. R. (2011). Influence of gene variants related to calcium homeostasis on biochemical parameters of women with polycystic ovary syndrome. *Journal of Assisted Reproduction and Genetics, 28*(3), 225–232.

21. Satish, C. N., Chew, X., & Khaw, K. W. (2020). *Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques*.

22. Shuja, M., Mittal, S., & Zaman, M. (2020). Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. In H. Sharma, K. Govindan, R. C. Poonia, S. Kumar, & W. M. El-Medany (Eds.), *Advances in Computing and Intelligent Systems* (pp. 195–211). Springer. https://doi.org/10.1007/978-981-15-0222-4_17

23. Sumathi, M., Chitra, P., Prabha, R. S., & Srilatha, K. (2021). Study and detection of PCOS related diseases using CNN. *IOP Conference Series: Materials Science and Engineering, 1070*(1), 012062.

24. Sun, D., Xu, J., Wen, H., & Wang, D. (2021). Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest. *Engineering Geology, 281*, 105972.

25. Thomas, N., & Kavitha, A. (2020). PREDICTION OF POLYCYSTIC OVARIAN SYNDROME WITH CLINICAL DATASET USING A NOVEL HYBRID DATA MINING CLASSIFICATION TECHNIQUE. *Technology, 11*(11), 1872–1881.

26. Vagios, S., James, K. E., Sacha, C. R., Hsu, J. Y., Dimitriadis, I., Bormann, C. L., & Souter, I. (2021). A patient-specific model combining antimüllerian hormone and body mass index as a predictor of polycystic ovary syndrome and other oligo-anovulation disorders. *Fertility and Sterility, 115*(1), 229–237.

27. Veloso, B., Gama, J., Malheiro, B., & Vinagre, J. (2021). Hyperparameter self-tuning for data streams. *Information Fusion, 76*, 75–86.
28. Wagha, P., Panjwanib, M., & Amrutha, S. (2020). Early detection of PCOD using machine learning techniques. *Artificial Intelligence and Speech Technology, 9*.

29. Witchel, S. F., Oberfield, S., Rosenfield, R. L., Codner, E., Bonny, A., Ibáñez, L., Pena, A., Horikawa, R., Gomez-Lobo, V., & Joel, D. (2015). The diagnosis of polycystic ovary syndrome during adolescence. *Hormone Research in Paediatrics, 83*(6), 376–389.

### Figures

| PCOS (Y/N) | Follicle No. (R) | Follicle No. (L) | Skin darkening (Y/N) | hair growth(Y/N) | Weight gain(Y/N) | Cycle(R/I) |
|------------|------------------|------------------|----------------------|------------------|-----------------|------------|
| 0          | 0                | 3                | 0                    | 0                | 0               | 2          |
| 1          | 0                | 5                | 3                    | 0                | 0               | 2          |
| 2          | 1                | 15               | 13                   | 0                | 0               | 2          |
| 3          | 0                | 2                | 2                    | 0                | 0               | 2          |
| 4          | 0                | 4                | 3                    | 0                | 0               | 2          |

**Figure 1**

Sample datasets for PCOS without infertility

**Figure 2**

Box plot of some potential important parameters
Proposed workflow for PCOS

![Proposed workflow for PCOS](image)

**Figure 3**

Proposed workflow for PCOS

**Figure 4**

Computational time for various SMOTE based Algorithm

![Computational time for various SMOTE based Algorithm](image)
Figure 5

ROC curve for various SMOTE based Algorithm