**Detection of Endocrine Disorder Based on Machine Learning using Physical and Clinical Database**

Prafull Pandey

Email: - [prafullpandey@united.ac.in](mailto:prafullpandey@united.ac.in)

United Institute of Technology

Amit Kumar Tiwari

Email: - [kumartiwariamit@gmail.com](mailto:kumartiwariamit@gmail.com)

United Institute of Technology

Utkarshani Tiwari

[utkarshani15@gmail.com](mailto:utkarshani15@gmail.com)

United Institute of Technology

Ashutosh Kumar Singh

[ashutoshsingh4415@gmail.com](mailto:ashutoshsingh4415@gmail.com)

United Institute of Technology

Ashish Singh

[singhashishmi7800@gmail.com](mailto:singhashishmi7800@gmail.com)

United Institute of Technology

Muskan Singh

[singhmuskan1509@gmail.com](mailto:singhmuskan1509@gmail.com)

United Institute of Technology

**Abstract**

These days PCOS is not very unusual among girls of reproductive age. It’s miles a hormonal disease that causes irregular period, cyst formation and infertility amongst women. Now it emerges as a worldwide issue and analysis of PCOS is hard because everyone with polycystic ovaries no longer has PCOS subsequently whole diagnosis of that is required like pelvic ultrasound, adjustments in frame and others. Consequently, we’ve efficiently evolved a version of the use of Jupyter notebook and data samples from website Kaggle, and the data samples used for this version contain several records of patients, we've got used diverse machines getting to know set of rules like Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) and different to assess accuracy, sensitivity, precision. Jupyter notebook has proven a terrific model fitting embedded tactics for scoring high accuracy. The result has confirmed that Random Forest algorithm and AdaBoost set of rules accomplished the highest accuracy of 91.3%.

**Keywords: -** Machine Learning, Hormonal Imbalance Disorder, Random Forest, AdaBoost.

**1. Introduction**

Polycystic ovary syndrome commonly Known as PCOS; Hormonal Imbalance Disorder is the maximum not unusual hormonal sickness occur in women of reproductive age. PCOS influence the fitness of women through causing hormonal imbalance and metabolism problems, generally the ovaries produce hormones known as estrogen (lady hormones) and androgen (male hormones) for normal fitness however girls tormented by PCOS hormones are imbalanced with better androgen and less estrogen than ordinary. This reason lumps and those lumps gradually enlarge and affect the ovulation process. [[9]](#ref21) PCOS has many symptoms like irregular menstrual cycles, zits, extra hair increase, weight benefit, oily pores and skin and trouble in getting pregnant. The length of PCOS remedy is rely upon the signs and symptoms of the patient, however those symptoms vary from individual to individual each ladies having cyst aren't recollected to have PCOS. PCOS growth the threat of diabetes, high blood pressure (blood strain), depression and different health problems. The female reproductive system includes fallopian tubes, uterus and vagina. The ovaries store a limited number of eggs in a woman's life, each attached to a follicle. In ordinary instances the egg matures, and estrogen launch from follicles those results in ovulation but in the case of PCOS these manners get disrupt and menstrual cycle get affected. Spotting and treating PCOS is not smooth. Even though breeds and genetics were determined to be two principal elements at the back of PCOS in ladies, the precise cause of PCOS remains unknown. A gradual, inactive lifestyle and beside the point nutrients can also help girls get PCOS. [[12]](#ref22) Diagnosing PCOS isn't clean. Due to this fact no longer, all people have a polycystic ovary with PCOS, and not everybody has a PCOS with ovarian cysts. PCOS is straightforward to diagnose.

**2. Literature Review**

Correctly as it is hard to assemble the version using algorithms for hormonal issues, medical exams, symptoms, and machine getting to know. related running this detail, we’ve mentioned the existing literature alongside the used strategies and algorithms. This section offers a table summarizing current studies, with their major findings and an overview of the principal results and conclusions made from the research carried out in the area. Over the past few years, numerous researchers have introduced AI models for analyzing PCOS, using data like clinical parameters and key symptoms. [[5]](#ref1) Here are some of these studies: Sumathi et al. [[4]](#ref12) presented image processing CNN-based for feature extraction for detection of Ovulatory Dysfunction. The model is fundamentally based on image processing to recognize the cyst in ovaries. For usage of parameter measurement, the algorithm Watershed is used for feature extraction and OpenCV. However, there is limitation in detecting cysts and it is not clear if an individual has PCOS or not. The accuracy achieved by this model is 86%. Silva et al. In this research, Boruta and SHAP methods were used to identify the most impressive properties before training a random forest model. The dataset consisted of both PCOS, and both individuals diagnosed unconditionally. [[2]](#ref15) A total of 58 features were evaluated based on their relevance and significance. As a result, the model received 86% accuracy. Denny et al. devised an approach to identify PCOS at an early stage with a reduced number of medical and metabolic features using Principal Component Analysis (PCA). Experimented with some of the popular machine learning algorithms, such as logistic regression, K-nearest neighbors, Random Forest, Naïve Bayes, and Support Vector Machine. The best results were obtained using the random forest model, achieving an accuracy of 89.02%. Zigarelli et al. [[3]](#ref18) employed the Kaggle dataset with 541 patients from 10 hospitals and 44 features and built a diagnostic model. They obtained 82.5% accuracy for aggressive and 90.1% for non-invasive methods using the CatBoost algorithm in conjunction with K-fold cross-validation and thus indicated that non-invasive methods can provide more diagnostic accuracy.

Concerning the research in two of the fine classifiers have been KNN along with logistic regression resulting in accuracy of 85.58% and 87.20% respectively. characteristic choice and facts filtering had been carried out by way of some researchers that assure good prediction outcomes. A few of them choose the capabilities counting on their medical dependability, resulting in achieving high accuracy and precision rate.

Silva et al. [[1]](#ref19) used Boruta Sharp method and they developed Random Forest model for further analysis in their study. In their data sample both healthy people and people diagnosed with PCOS are considered. According to their significance and relevance, a total of 58 features were taken. Overall, their accuracy is 86%.

**3. Methodology**

On this, we provide an explanation step by step what is hard to carry out to make a machine gaining knowledge of undertaking. For using the data sample, we are using Kaggle, and we should prepare the data sample, we must easy and preprocess the records to make them understood for machine training. Even as we're operating in healthcare, we can't afford the error in facts therefore having correct data of patient is important. Several parameters we have used for detecting physical parameters that we've got used are weight, age, height and many others. For the detection we've used several attributes that will help in increasing the accuracy, the following attributes are shown in (Table 1). In this undertaking, we are the usage of exclusive machine learning algorithms like Random Forest, Logistic Regression, Support Vector Machine, and AdaBoost increase for the prediction of PCOS. The first step we use to put together the data sample is the preprocessing of facts which consists of deletion of unnecessary columns, doing away with all duplicate records to keep away from repetition, dealing with missing values, and changing all text-based data into numerical information. This step ensures that your data sample is smooth, prepared and equipped for devices gaining knowledge of features. [[10]](#ref22) The following step in making an undertaking is characteristic choice, it's far used to locate the most applicable attributes, which assist to decrease the computational complexity and in enhancing the version overall performance. The next step is splitting of dataset which is used to assess the performance of the special system-studying features. We train and evaluate all the features with each different, and the features that are covered are Random Forest, Logistic Regression, AdaBoost, and Support Vector Machine. Based on key performance like accuracy and precision decide the only set of rules for the detection of PCOS. [[11]](#ref3) In parent 1 we will see how our internet site works and tell every step virtually and concisely. The following device studying features which might be used to determine the PCOS are described under evaluation of features here, we examine all the gadgets getting to know models based on the accuracy we get. After the implementation of the device, gaining knowledge of algorithms is the subsequent accuracy table we have obtained (Table 2). These were generated in the Jupyter lab. This shows that Random Forest is seen as performing better compared to others. Random Forest algorithm is used for our final hypothesis which will result in the use of check statistics. The accuracy we get the use of the Random Forest algorithm is 91.3%.

Random Forest: - Random Forest is described as mastering algorithms used for classification and regression. It’s a set of Decision Tree that combines to make predictions. It uses plenty of Decision Tree to provide the prediction. It is also an ensemble getting to know approach. It is the combination of two ideas randomness and forests. The first element refers to randomly selected records samples and the 2nd part refers to selecting trees. [[8]](#ref19) It is a relatively versatile algorithm this is utilized in diverse programs like disorder prediction, junk mail detection and prediction of inventory prices. Random Forest Algorithm handles the missing information, with larger and greater complex datasets it presents correct predictions, and we integrate more than one choice of tree which also reduces the hazard of overfitting.

Ensemble Learning: Random Forest is based totally on ensemble learning. Ensemble getting to know is defined as a mixture of predictions of a couple of trees to make steady and reliable predictions. The goal is to reduce errors and improve performance.

Support Vector Machine: - Support Vector Machine is a system mastering algorithm that is used for both classification and regression. It is to categories the statistics effectively. In classification terms, it is a discriminative classifier to discover the optimal hyperplane to categorize records into specific instructions. SVM aimed to find the best plan in an N-dimensional area to separate records points into different training. Support Vector Machine (SVM) algorithm is designed to classify the data by determining an optimal hyperplane that separates two different categories. The purpose is to maximize the margin, which is the minimum distance between the hyperplane and the nearest data points from each class. These closest data points called support vectors play an important role in defining the border. When the classification is strictly used without permission, this approach is known as a hard margin SVM. To make the facts linearly separable, SVM uses the approach which is kernels. Kernels map the statistics into a better dimensional area where it turns into separable.

Logistic Regression: - Logistic Regression is defined as the statistical method which is employed to predict the outcome based on input variables. It’s a form of supervised machine learning. Logistic regression is used for type tasks. It is a method used for binary classification, it takes the entering and produces the output in 0,1. [[7]](#ref3) There are 3 types of Logistic regression -Binomial, Multinomial, Ordinal. In logistic regression every remark is impartial to other remarks. The core of logistic regression is shaped by Sigmoid feature; it's miles a function that is used to map the expected values to possibilities.

A diagram of a patient's condition

AI-generated content may be incorrect.Adaboost: - Adaboost, known as Adaptive Boosting, is used to create a strong classifier by combining more than one vulnerable classifier. For correcting the errors made through the preceding features it adds the classifiers itself it is a learning set of rules that improves the accuracy of classification by combining a couple of classifiers. Adaboost can be sensitive to outliers or noisy data. There are three sorts of boosting set of rules -Gradient, XGBoost, CatBoost. The intention of this algorithm is to reduce bias, on these features are weighed on the idea of their overall performance.

Fig 1: 1- Flowchart of Website

|  |  |  |  |
| --- | --- | --- | --- |
| **N0** | **Parameters** | **N0** | **Parameters** |
| 1 | Effects of PCOS | 23 | Cycle Duration (days) |
| 2 | Age (years) | 24 | Duration of marriage (years) |
| 3 | Body Weight (kg) | 25 | Pregnancy status (Yes/No) |
| 4 | Body Height (cm) | 26 | Count of Abortions |
| 5 | Body Mass Index | 27 | FSH Level |
| 6 | Blood Group | 28 | LH Level (mIU/mL) |
| 7 | Pulse rate(bpm) | 29 | I β - HCG (mIU/mL) |
| 8 | Respiratory Rate (Breaths/min) | 30 | II β -HCG (mIU/mL) |
| 9 | Hemoglobin level(g/dl) | 31 | FSH / LH Ratio |
| 10 | Menstrual Cycle Regularity (R/I) | 32 | Hip Measurement (inch) |
| 11 | Menstrual Cycle Length (days) | 33 | Waist Measurement (inch) |
| 12 | Waist and Hip Ratio | 34 | Pimples (Yes/No) |
| 13 | Thyroid Stimulating Hormone Level | 35 | Regular Exercise (Yes/No) |
| 14 | AMH Level (ng/mL) | 36 | Systolic Blood Pressure (mmHg) |
| 15 | Prolactin Concentration (ng mL) | 37 | Diastolic Pressure (mmHg) |
| 16 | Vitamin D3 (ng/mL) | 38 | Follicles count (Left) |
| 17 | Progesterone Level (ng/mL) | 39 | Follicles count (Right) |
| 18 | Random Blood Glucose  (mg/dl) | 40 | Average Follicles size  (left) (mm) |
| 19 | Weight Change (Yes/No) | 41 | Average Follicles size  (Right) (mm) |
| 20 | Hair Development (Yes/No) | 42 | Endometrial  Measurement (mm) |
| 21 | Hyperpigmentation (Yes/No) |  |  |
| 22 | Hair Fall (Yes/No) |  |  |

Table-1 – Attribute Table

**4. Results and Discussion**

The model achieved an impressive 91.3% accuracy in predicting PCOS cases. We got the data that we used for training and testing from Kaggle. The effectiveness of the machine learning model is measured by employing performance measures like accuracy and precision. Random Forest is recognized as a valuable tool in Ovulatory Dysfunction detection, which offers reliability, efficiency and initial identity. The dataset was analyzed to measure performance indicators such as accuracy, uniqueness, sensitivity and general efficiency. Many machines learning algorithms, including Random Forest, Logistic regression, Adaboost classifier and Support vector machine, were used to evaluate and compare their efficiency. Random forest obtained the highest accuracy of 91.3 %.

|  |  |  |
| --- | --- | --- |
| **SR. No** | **Models** | **Accuracy** |
| 0 | LR | 82.56% |
| 1 | SVM | 70.64% |
| 2 | RF | 91.32% |
| 3 | GB | 82.56% |

Table-2 – Accuracy of given algorithms

A graph showing different models

AI-generated content may be incorrect.

Fig 2: Comparision char of different models.

**5. Conclusion**

In this study, a high-performance ovulatory dysfunction detection model was distributed to detect the presence or absence of PCOS. There is a medical condition that causes female endocrine system disorders. PCOS is an ovarian disorder affecting women in their bleeding wounds. Hormonal imbalance leads to delay in the menstrual cycle. Women with PCOS suffer from facial hair growth, obesity, blackness in the skin, acne, pelvic pain and difficulty getting pregnant. If PCOS is uncontrolled in its early stages, it can contribute to infertility and reduce the possibility of natural conceving. Requiring the condition and requiring proper medical intervention can help handle the effect of reproductive health. Due to the limits of the initial detection, the record PCOS patient is higher than the previous year. In response, the proposed research has developed an online ovulatory dysfunction detection system using machine learning. Original identity can authorize patients to take the necessary steps such as exercise, diet and proper medicare, as recommended by their doctor, which can lead to a well-balanced life. The aim of our research is to use the ML model to understand the pattern of this reproductive system disorder.

**6. References**

1: - M. Sumathi ‘‘Study and detection of PCOS related diseases using CNN,’ IOP Conf. Ser., Mater. Sci. Eng., vol. 1070, Nov. 2021, Art. no. 012062.

2: - Silva I, et al. Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models. J Endocrinol Invest 2022:1–9.

3: - R.P.R. Murugan "Application of Machine Learning in PCOS Diagnosis: A Survey" published in 2021.

4: - P.K. Jain “Predicting Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms" published in 2019.

5: - K.M.S.R. Reddy "Machine Learning Models for PCOS Diagnosis Using Clinical Data" published 2021.

6: - K.R. Anusha "A Review on PCOS Diagnosis with Machine Learning Models" published in 2020

7: - Araviiskaia E. The role of topical dermocosmetics in acne vulgaris. J. Eur. Acad. Dermatol. Venereol 30, 926-935 (2016)

8: -Gold, M.H, Goldberg Current treatments of acne: Medications, lights, laser and a novel 650mus 1064nm Nd: YAG laser. J. Cosmet-Dermatol.16,303-318 (2017).

9: - Rohit Mishra et al,” Comparative Analysis of Conventional IP Network and MPLS Network over VoIP Application” / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 4496-4499

10: - Rohit Mishra et. al, “A Predicate Based Fault Localization Technique Based on Test Case Reduction” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 10, OCTOBER 2015.

11: - Rohit Mishra et. al, “Protein Structure Prediction Using Needle Man Wunsch Algorithm” Journal of Harbin Engineering University ISSN: 1006-70.

12: - Rohit mishra et. al, “Augmented Reality in Supply Chains of Indian Micro and Small Enterprises” Rivista Italiana di Filosofia Analitica Junior ISSN: 2037-4445

13: - Rohit Mishra et. al, “ENHANCED IMAGE ANALYSIS FOR GASTRIC CANCER USING CNN” DOI:10.48047/ecb/2023.12.9.51

14: - Rohit Mishra et. al, “Real-time Vehicle Tracking System Using Geofencing” https://doi.org/10.1007/978-981-97-6318-4\_20