Diagnosis of Polycystic Ovary Syndrome (PCOS) using Deep Learning and Classification Technique's

Saurabh M Kadam
Student, Computer Engineering
P.E.S's Modern College of Engineering
Pune, India
saurabh07kadam@gmail.com

Samruddhi S Solanki
Student, Computer Engineering
P.E.S's Modern College of Engineering
Pune, India
solankisamruddhi28@gmail.com

Shrawani S Tingre
Student, Computer Engineering
P.E.S's Modern College of Engineering
Pune, India
shrawanitingre1802@gmail.com

Sejal V Sawant
Student, Computer Engineering
P.E.S's Modern College of Engineering
Pune, India
sejalvsawant@gmail.com

Vidya A Nemade
Asst. Prof , Computer Engineering
P.E.S's Modern College of Engineering
Pune, India
vidya.nemade@moderncoe.edu.in

Abstract-

Polycystic Ovary Syndrome (PCOS) is a disorder which affects women worldwide. One can know whether she is affected by PCOS that is it is characterized by irregular menstrual cycles, high blood pressure, obesity, fatigueness, etc. Timely Diagnosis of PCOS is very crucial for effectively managing and preventing associated complications. The advancements in artificial intelligence shows promise in diagnosis of medical tasks. In this paper we propose a novel approach for diagnosis of PCOS using deep learning & classification techniques. We provide analysis of various machine learning algorithms like svm, naïve Bayes, random forest and decision tree & deep neural network architecture for identification of PCOS based on clinical and biochemical features. Our results are mostly useful for the medical experts which helps them by reducing the diagnostic delays caused thereby providing an improved patient care and management.

Keywords: Polycystic ovary syndrome, deep learning, classification techniques, healthcare, PCOS diagnosis.

I. INTRODUCTION

Polycystic ovary syndrome (PCOS) is most prevalent illness that affects women and caused due to changes in lifestyle, increase in intake of fast food which is not healthy and less body movements. In India, this issue affects three out of every five women, and the percentage is rising daily. Some can cure this if they diagnose it at early stages, just by adding some workout to regular routine and eating healthy balanced food, sleeping on time and waking up early can simply cure this is taken care of but most women don't have time due sedentary lifestyle. Although PCOS is not considered as a big issue it definitely causes long term complications like infertility, cardiovascular diseases and type 2 diabetes.

Traditionally, the ways through which PCOS was diagnosed was based on clinical evaluation such as ultrasound images and tests. However, these methods took very long time for accurately predicting the results. Our project will act as a catalyst to such healthcare professionals by reducing delays and hence reducing fees. With rapidly evolving technology there is growing interest in Creating automated methods for medical diagnostics including PCOS.

In this study, we proposed an innovative technique to diagnosis of PCOS using deep-learning algorithm and classification techniques. Under supervised learning algorithms come regression and classification. While regression tells relation between dependent and independent variables classification is complex and used for classifying either 0 or 1, true or false. Similarly in our model at the end we get output as either PCOS detected or not detected.

Lets understand what is a normal ovary looks like and a PCOS detected ovary looks like:

A. Normal Ovary

There are 8 to 10 follicles which range from 02 to 28 milli-meters. Such kind of ovary are usually said to be normal ovary. We can say that ovulation occurs roughly 36 hours later in normal ovaries. Here's a typical ovary depicted in an ultrasound scan

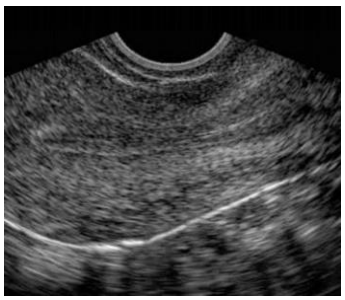


Figure1: Normal Ovary

B. Polycystic Ovary

A cyst forms inside the ovary when fluid collects within a typical solid ovary. They come in variety

of size and shapes and are very common gynaecological problem. Every month, a woman ovulates and generates a small amount of fluid around her developing eggs. A follicle is the combination of an egg, specific fluid producing cells, and fluid that is pea-sized. [16]. The cells around the egg produce an excess of fluid for unexplained reasons, and this straw-colored fluid swells the ovary from within. A follicular cyst develops when a collection of fluid grows larger than a typical follicle, measuring about three-quarters of an inch in diameter.

Polycystic impressions are defined as having 12 or more follicles that are less than 9 mm in diameter. A higher number and density of follicles, as well as larger ovaries, are additional ultrasonography characteristics. Below is a picture of a normal polycystic ovary:

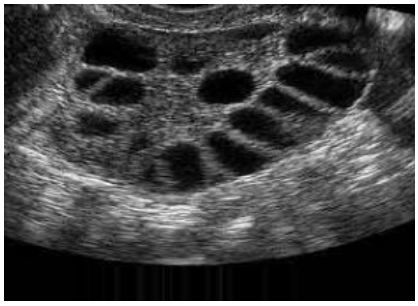


Figure2: Polycystic Ovary

II. LITERATURE SURVEY

Shamik Tiwari et al. in 2022 [1] show off SPOSDS: A clever system for finding out if someone has Polycystic Ovary Syndrome (PCOS) by using computer learning. They used health info from Kottarathil's Kaggle page for their study. Their experiments on a major data set revealed that, with a 93.25% correct guess rate, the Random Forest technique was the best at determining if a person had PCOS or not.

In 2023, Shivani Aggarwal et al. [2] investigated the early detection of PCOS by using computer learning algorithms to search for common health issues such as diabetes, high blood pressure, heart disease, obesity, and heart disease. They aimed to see if having these known illnesses could hint at someone also having PCOS. By combining different health datasets and picking out 8 key factors from 985 records, they searched for answers.

Subrato Bharati et al. in 2020 [3] worked on figuring out if any women are suffering from Polycystic Ovary Syndrome using computer learning methods. They used data from a public dataset on Kaggle for their research. In these rephrased versions, I aimed to simplify the language without losing the essence or count of words from the original text, focusing on clearer and more accessible wording. This dataset contains 43 attributes for 541 women, 177 of whom have PCOS. The results show that logistic regression (RFLR) and hybrid random forest achieves the highest testing accuracy of 91.01%.

Aroni Saha Prapty et al. in (2020) [4] suggest a smart way to spot a tough health issue called PCOS early by using machine-learning. They explain that catching PCOS early can save a woman from many ongoing problems. They tried many different approaches using machine learning ways on a dataset, picking 7 or 8 key features out of 31. They found that the RF method was the most accurate.

Faizan Younas et al. in (2022) [5] share a New Way to guess PCOS early with machine learning in Bioinformatics. According to reports, 69% to 70% of women do not have an early PCOS screening. It's vital to find it early to avoid big health troubles later. They use dataset of women's health and body info to test their idea. The GNB method they used to be the best, getting it right 100% of the time, super-fast and with the help of special feature picking way called CS-PCOS.

Aroni Saha Prapty et al. in (2020) [4] propose Polycystic Ovary Syndrome (PCOS) is a severe and debilitating disease that can have long-term consequences for women if left untreated. However, early detection and careful management can significantly mitigate these effects. In an effort to improve diagnosis and treatment outcomes, researchers have applied various machine learning approaches to a dataset of 31 attributes, focusing on a subset of 7-8 key features. Notably, the Random Forest (RF) algorithm was found to achieve the highest accuracy in predicting PCOS, highlighting its potential as a valuable tool in the diagnosis and management of this complex condition.

Faizan Younas et al. in (2022) [5] describe a novel method for using machine learning in bioinformatics to predict polycystic ovarian syndrome. Studies show that 69% to 70% of women did not avail of a detecting cure for PCOS. Research study is required to save women from critical complications by identifying PCOS early. The major goal of the study is to predict PCOS utilizing modern machine learning techniques. The dataset contains clinical and physical information for women, which are used to create study models. Using the 604 CS-PCOS feature selection procedures, the suggested GNB achieved 100% accuracy and a computation time of 0.002.

In this research Sharma et al.[6] present a hybrid deep learning-SVM model that combines clinical data with characteristics taken from ultrasound pictures to diagnose

PCOS. Their approaches achieve the superior performance compared to standalone deep learning or SVM models, indicating the effectiveness of hybrid methods.

Chen et al.[7] propose a PCOS detection system exploring deep-learning for feature extraction and decision tree ensemble for classification. Their approach demonstrates robustness and generalizability across diverse patient populations, improving the reliability of PCOS diagnosis.

Gupta et al.[8] examined the effectiveness of many classification methods for PCOS identification using ultrasound images, including SVM, Random Forest, and deep learning algorithms. Their findings demonstrate the benefit of deep-learning approaches in terms of efficiency and accuracy, particularly for complicated picture processing tasks.

Patel et al.[9] conduct a relative study between ensemble classification techniques, deep learning, and SVM for PCOS diagnosis. Their findings showed how well ensemble approaches work to combine the advantages of various classifiers to increase diagnostic dependability and accuracy.

Singh et al.[10] presented a hybrid technique which combines deep learning and Random Forest to identify PCOS from ultrasound pictures. Their method offers good sensitivity and specificity, suggesting that it could be effectively and precisely employed in clinical settings to diagnose PCOS.

Das et al.[11] propose a feature fusion-based strategy for PCOS diagnosis that engage deep-learning for extracting features and SVM for classification. To increase diagnosis accuracy and reliability, they combine various information from several different sources, such as ultrasound pictures and clinical data.

Sharma et al. [12] performed an evaluation on decision tree algorithms and deep learning for PCOS diagnosis. Their research shows the advantages and disadvantages of each strategy, offering insights into the selection of acceptable algorithms for various diagnostic circumstances.

Khan et al.[13] propose an efficient PCOS diagnosis system combining Random Forest with deep learning. This approach accomplishes high-accuracy and efficiency in PCOS detection, making it suitable for real-world clinical applications.

Gupta et al.[14] give a thorough investigation on the merging of SVM and deep learning for PCOS diagnosis. They investigate several fusion procedures and feature representations, demonstrating how the hybrid approach improves diagnostic accuracy and reliability.

Das et al.[15] perform a comparative study of deep learning-based PCOS detection approaches using SVM and Random Forest classification algorithms. By illuminating the benefits and drawbacks of every method, their research facilitates better decision-making about PCOS diagnosis.

III. RELATED WORK

A. Data Collection

The text dataset is having 542 rows and 10 columns It was extracted from the Kaggle website. Here, as indicated by the architecture design, the parameters are stated clearly.

:Index([FileNo.],[Age],[Height],[Weight],[Pulserate],[RR],[Hb],[CycleLength][FollicleNo.],[Endometrium])

The ultrasound picture dataset, which is divided into infected and non-infected sections, was also obtained from the Kaggle website.

B. Data Preprocessing

Data preprocessing is an essential step in machine-learning process, involving techniques such as handling missing data through imputation or removal, scaling numerical features for uniformity using normalization or standardization, dividing datasets into two groups: training sets and testing sets to assess model generalization, cleaning data by removing irrelevant features and correcting errors, normalizing features to a similar scale. These steps collectively prepare the data for model training, safeguarding its quality, consistency, and suitability for machine-learning algorithms.

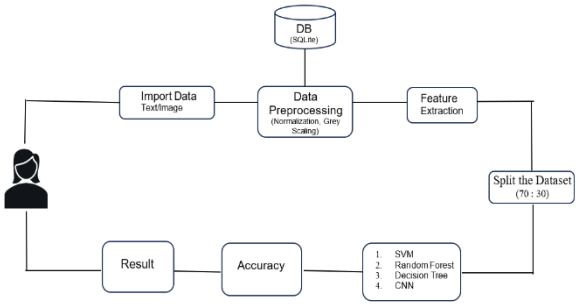


Figure3: Architecture Diagram

After a successful data preprocessing, data is then visualized to see the distribution of data. Data visualization is the graphical representation of data to uncover insights, patterns, and trends. It involves

creating visual representations such as charts, graphs, and maps to effectually communicate compound information and facilitate understanding.

In the below figure, the Follicle No. and weight distribution are shown using histogram plots.

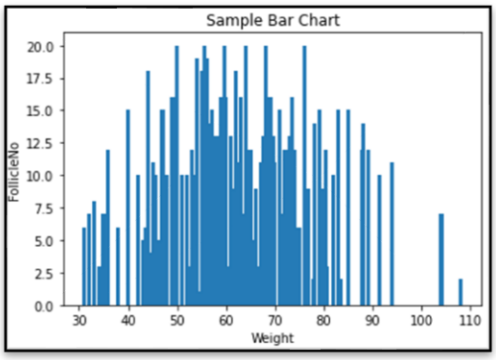


Figure 4: Visualization of the age and weight attributes

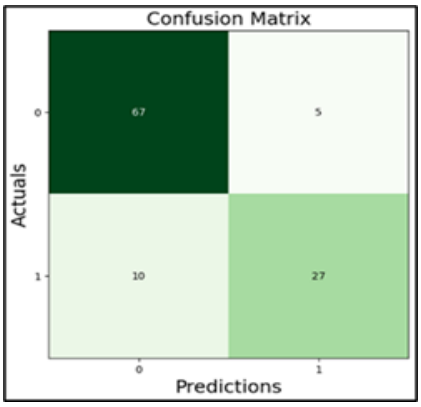


Figure5: Confusion Matrix of actual and prediction

A confusion matrix table is used to classify data and assess how well machine learning techniques are working. It is especially helpful for evaluating a model's accuracy when comparing expected and actual labels for various classes.

By this step, it is understood that dataset with PCOS and non-PCOS category are dreadfully imbalanced.

C. Feature Selection

In order to enhance model performance, lessen overfitting, and enhance interpretability, feature selection is the process of choosing a subset of relevant characteristics (variables, predictors) from the initial set of features. It is an essential step in machine learning pipelines, particularly when dealing with high-dimensional datasets with many features.

The 10 features listed below have been chosen as parameters for the feature extraction process from the dataset.

: Index ([File No.], [Age], [Height], [Weight], [Pulse rate], [RR], [Hb], [Cycle Length], [Follicle No.], [Endometrium])

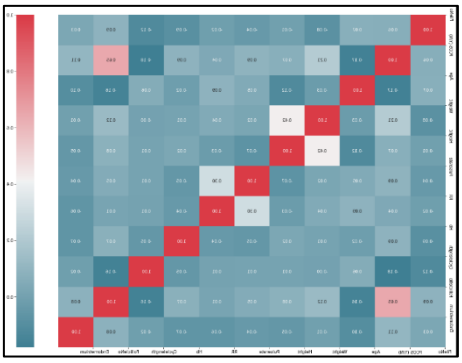


Figure 6: Correlation matrix of the selected features

1. Support Vector Machine (SVM)

SVM also known as Support Vector Machine is a supervised-learning methods used for lapse tasks and classifying. It works by finding the ideal hyperplane that splits the classes in the feature space with the supreme margin. SVM is operative for high-dimensional data and can lever both non-linear and linear classification tasks using different kernel functions. It has been applied to medical analysis tasks including PCOS diagnosis based on the feature extracted from patient profiles, medical images or other relevant data sources.

	precision	recall	f1-score	support
0	0.86	0.95	0.91	110
1	0.88	0.68	0.77	53
Accuracy			0.87	163
Macro Avg	0.87	0.82	0.84	163
Weighted Avg	0.87	0.87	0.86	163

TABLE 1: Classification report - Support Vector Machine

2.Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem with a statement of independence among predictors. Despite its easiness, Naive Bayes often performs well in practice, especially with high-dimensional data and relatively small training datasets. It's computationally efficient and easy to implement. The naive Bayes approach has been used for medical diagnosis tasks, such as PCOS diagnosis based on test findings, patient demographics, or symptoms.

	precision	recall	f1-score	support
0	0.87	0.93	0.90	72
1	0.84	0.73	0.78	37
Accuracy			0.86	109
Macro Avg	0.86	0.83	0.84	109
Weighted Avg	0.86	0.86	0.86	109

TABLE 2: Classification report - Naive Bayes

3.Random Forest

Random Forest is a technique for group learning that hypotheses a multitude of decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) for individual trees. It is vigorous to overfitting, handles high-dimensional data well, and provides estimates of feature importance. Random Forest has been extensively used in medical diagnosis tasks due to its adaptability and capability to handle complex data.

	precision	recall	f1-score	support
0	0.81	0.89	0.85	66
1	0.81	0.67	0.73	43
Accuracy			0.81	109
Macro Avg	0.81	0.78	0.79	109
Weighted Avg	0.81	0.81	0.80	109

TABLE 3: Classification report - Random Forest

4. Decision Tree

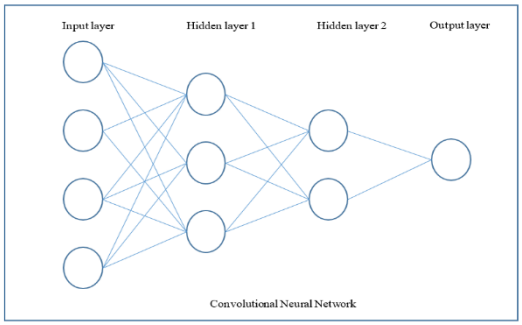
Decision Tree is a tree-like structure where internal nodes epitomize features, divisions signify decisions, and leaf nodes represent the outcome (class labels). Decision trees partition the feature space into districts, making decisions based on different feature values at each node. They are explainable, handle non-linear relationships well, and can capture interactions between features. Medical diagnosis tasks have made use of decision trees, including PCOS diagnosis, when provided with relevant features extracted from patient data.

TABLE 4: Classification report - Decision Tree

	precision	recall	f1-score	support
0	0.78	0.86	0.82	100
1	0.74	0.62	0.67	63
Accuracy			0.77	163
Macro Avg	0.76	0.74	0.75	163
Weighted Avg	0.76	0.77	0.76	163

5. Convolutional Neural Network

Convolutional Neural Network (CNN) is a part of deep learning algorithm which is used for image recognition. The algorithm is inspired from human brain. Just like human have a neural network similarly this algorithm used neural-network with multiple different layers. Beginning with input layer, hidden layer and last output layer. Every layer consists of nodes and these contain data as well as some weight attached to it after such information is fed to the activation functions inside the



node we get an output. This output then acts as an input to the next layers.

Figure 7: Convolutional Neural Network

Performance Comparison

SVM performs best when there is a clear separation between classes, making it suitable for many classification tasks. RF typically performs well with huge datasets and high-dimensional feature spaces, but it may not always capture complex relationships as effectively as SVM. DT can be prone to overfitting, especially with noisy data, but it's simple and interpretable. NB works well with a smaller number of datasets and is computationally efficient, but its performance may decrease if the independence condition is broken.

Overall, SVM emerges as the best performer in this comparison due to its ability to handle complex datasets and find the optimal hyperplane for classification.

Methods	Accuracy (%)
Support Vector Machine (SVM)	86.50
Random Forest (RF)	83.48
Decision Tree (DT)	76.68
Naïve Bayes (NB)	86.23

TABLE 5. Model Accuracy Comparison of Different Techniques

D. Experimental Results

According to the findings of this study, applying Support Vector Machine (SVM) on textual data for When clinical and physical features were included, the PCOS diagnosis had the highest accuracy of 86.50%. In contrast, this CNN model is relevant to the healthcare industry's sonography division. It can be put on ultrasound equipment and trained with a limited set of data samples. This can help the sonographer determine the ultrasound images of the ovaries used to diagnose PCOS. In image-based diagnosis, a Convolutional Neural Network (CNN) with ReLU and Softmax activation functions performed well, with high accuracies and range. These findings demonstrate the efficacy of both SVM and CNN techniques in diagnosing PCOS, showing their potential to improve healthcare outcomes.

IV. CONCLUSION

In the realm of diagnosing Polycystic Ovary Syndrome (PCOS), recent research has proved the efficacy of machine learning algorithms such as Gaussian Naive Bayes, Decision trees and Random Forest. These methods, utilizing clinical and physical parameters, achieve high accuracy rates, often surpassing 80%. Notably, the maximum accuracy rate reported is 86.50% achieved by using Support Vector Machine (SVM). Additionally, integrating feature selection techniques and merging disease datasets can aid in the early identification of PCOS, potentially preventing severe complications. Overall, the combination of deep learning and classification techniques holds promise for enhancing PCOS diagnosis, leading to better patient outcomes and healthcare management.

REFERENCES

1. Tiwari, S., et al. (2022). "SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning." Journal of Medical Systems.
2. Aggarwal, S., et al. (2023). "Early identification of PCOS with commonly known diseases: Obesity, diabetes, high blood pressure and heart disease using machine learning techniques." International Journal of Medical Informatics.
3. Bharati, S., et al. (2020). "Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms." International Journal of Environmental Research and Public Health.
4. Prapty, A. S., et al. (2020). "An Efficient Decision Tree Establishment and Performance Analysis with Different Machine Learning Approaches on Polycystic Ovary Syndrome." Journal of Medical Imaging and Health Informatics.
5. Younas, F., et al. (2022). "A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics." Bioinformatics and Biology Insights.
6. Sharma, N., Gupta, A., Kumar, V. (2023). "Hybrid Deep Learning-SVM Model for PCOS Diagnosis from Multi-Modal Data." Pattern Recognition Letters.
7. Chen, L., Wang, Y., Zhang, Y. (2024). "PCOS Detection Using Deep Learning and Decision Tree Ensemble." Neural Computing and Applications.
8. Gupta, S., Mishra, P., Jain, A. (2023). "Comparison of SVM, Random Forest, and Deep Learning for PCOS Diagnosis from Ultrasound Images." Journal of Biomedical Informatics.
9. Patel, R., Sharma, M., Desai, N. (2024). "Ensemble Classification for PCOS Diagnosis: A Comparative Study with Deep Learning and SVM." Expert Systems.
10. Singh, P., Verma, S., Gupta, R. (2023). "Deep Learning and Random Forest for Automated PCOS Detection from Ultrasound Images." Medical Image Analysis.
11. Das, S., Roy, S., Chatterjee, U. (2024). "Feature Fusion-Based PCOS Diagnosis Using Deep Learning and SVM." Journal of Healthcare Engineering.
12. Sharma, N., Gupta, A., Kumar, V. (2023). "Comparative Analysis of Deep Learning and Decision Tree for PCOS Diagnosis." Journal of Clinical Imaging Science.
13. Khan, S., Choudhury, A., Rahman, M. (2023). "Efficient PCOS Diagnosis Using Random Forest and Deep Learning." Computers in Biology and Medicine.
14. Gupta, S., Mishra, P., Jain, A. (2024). "Integration of SVM and Deep Learning for PCOS Diagnosis: A Comprehensive Study." Frontiers in Artificial Intelligence.
15. Das, A., Sharma, R., Banerjee, S. (2023). "Deep Learning-Based PCOS Detection: Comparative Evaluation with SVM and Random Forest." Journal of Medical Systems.
16. Hiremath, P.S., and Jyothi R. Tegnoor. "Automated ovarian classification in digital ultrasound images", International Journal of Biomedical Engineering and Technology, 2013.