ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

Progress in Healthcare: Utilizing Machine Learning for the Diagnosis of Polycystic Ovary Syndrome

[1] Chetna Sharma, [2] Renu Vadhera, [3] Dr. Sarika Chaudhary

[1] M.Tech Student (CSE), DPGITM Engineering College
[2] Assistant Professor (CSE-AI/DS), DPGITM Engineering College
[3] Associate Professor (CSE), DPGITM Engineering College
Corresponding Author Email: [1] chetnaa.chetu@gmail.com, [2] vadhera.renu@gmail.com, [3] sarikacse23@gmail.com

Abstract— A frequent endocrine disorder affecting fertile women is PCOS. It is characterized by hormonal abnormalities that cause ovarian cysts, irregular menstruation, and infertility, among other symptoms. For PCOS to be effectively managed and related problems to be avoided, early and correct diagnosis is essential. An advanced healthcare system using machine learning (ML) approaches to diagnose PCOS is presented in this abstract. Comprehensive health data, such as medical history, hormone profiles, ultrasound imaging, and clinical symptoms, are integrated with machine learning algorithms in the suggested system. The system improves diagnostic accuracy by detecting important biomarkers and patterns suggestive of PCOS through feature engineering and selection. In addition, it makes use of advanced machine learning methods, such as support vector machines, neural networks, and random forests, to evaluate multidimensional data and offer individualized diagnostic information. This cutting-edge healthcare system has several benefits, such as decreased rates of misdiagnosis, increased diagnostic accuracy, and prompt PCOS patient action. Additionally, it makes proactive healthcare easier by estimating the likelihood that at-risk persons may acquire PCOS based on lifestyle, genetic, and demographic factors. Furthermore, the technology provides healthcare providers with an intuitive interface that expedites decision-making and improves patient care. To sum up, the integration of machine learning into healthcare systems signifies a noteworthy progression in the identification and treatment of polycystic ovarian syndrome. Applying data-driven methods, this innovative strategy has the potential to enhance clinical outcomes, patient empowerment, and the field of tailored medicine for reproductive health.

Index Terms— ADA Boost, PCOS, Random Forest.

I. INTRODUCTION

PCOS, also known as polycystic ovarian syndrome, is a complicated endocrine condition that affects a large percentage of women globally. Hormonal irregularities are its defining characteristic, which leads to infertility, ovarian cysts, and irregular menstruation. It is critical to diagnose PCOS as soon as possible and accurately in order to effectively manage the condition and avoid any related consequences. Recent developments in medical technology, especially regarding machine learning (ML), have demonstrated potential to improve PCOS diagnosis. The development and application of a state-of-the-art healthcare system that diagnoses PCOS using machine learning techniques is examined in this introduction. The first section provides a thorough analysis of the research and literature on the subject of machine learning-based PCOS identification. It draws attention to the difficulties in using conventional diagnostic techniques and investigates the utilization of ML techniques to analyze a range of data sources, including clinical symptoms, hormone profiles, medical histories, and ultrasound images, in order to improve diagnostic precision. The suggested healthcare system is developed on the basis of earlier studies conducted in this field. The variety of technologies used in the model's development is described in the second part. This covers, among other things, feature engineering, data preparation, and model optimization tactics, as well as machine learning technologies such as support vector machines, random forests, and neural networks. Utilization of cloud computing infrastructure, medical imaging software, and electronic health records (EHR) may also be covered, highlighting the interdisciplinary aspect of the suggested healthcare system. An explanation of the research methodologies applied in the development and evaluation of the healthcare system is given in the third section. This includes methods for gathering data, standards for recruiting participants, and the definition of the experimental design. It also goes into detail about how the machine learning models are trained and validated, including how datasets are divided, how cross-validation works, and what performance measures are used to gauge how effective the models are. The study's results are shown in the fourth part, which also highlights how well the developed healthcare system diagnoses PCOS. It talks about important metrics including sensitivity, specificity, and accuracy as well as how well the model can find important biomarkers and patterns that point to PCOS. Additionally, based on empirical findings, the section may provide comparison analyses with other diagnostic techniques and emphasize the system's advantages and disadvantages. The significance of the constructed healthcare system for PCOS diagnosis is summarized in the final portion of the study, which serves as the conclusion. The main conclusions, new developments in technology, and practical ramifications for clinical practice are outlined. It also addresses future research directions and the implementation of the suggested system in actual healthcare environments. In the end, the conclusion restates how machine learning is altering the field of reproductive health by enhancing patient outcomes and diagnostic processes.

II. BACKGROUND STUDY

Few paper highlight PCOS issue and where they notice the impact of PCOS among women around 5-10% of women in age group, with varying prevalence among different ethnicities. PCOS is associated with a number of health dangers including obesity, hypertension, type 2 diabetes, and cardiovascular disease. It also heightens the chances of gynecological cancers and first-trimester miscarriages. Symptoms include irregular menstrual cycles, excessive male hormone production leading to issues like acne and hirsutism (excessive hair growth), and problems with follicle development in the ovaries, often resulting in infertility. Early detection is crucial, and while automated screening systems based on ultrasound images exist, there's a gap in utilizing machine learning techniques for PCOS diagnosis based on clinical parameters. study aims to fill that gap by statistically analyzing metabolic and clinical features and employing logistic regression and Bayesian classifiers to predict PCOS, offering a more accurate and efficient means of detection [1]. Some paper used machine learning models, specifically K-NN, are employed for PCOS detection, with Logistic Regression proving more accurate (92% accuracy compared to K-NN's 90.74%). The study's structure includes methodology, results, conclusions, and references, aiming to enhance PCOS diagnosis through advanced computational techniques [2]. Hormonal abnormalities are a hallmark of PCOS, sometimes known as Stein-Leventhal syndrome, particularly elevated androgen levels, and metabolic issues, potentially leading to irregular periods, enlarged ovaries with microcysts, and infertility. Diagnosis typically relies on clinical symptoms, though ultrasound evidence of ovarian cysts can aid identification. Common symptoms include acne, hirsutism, baldness, skin pigmentation, obesity, and irregular periods. PCOS is believed to have genetic roots, with insulin resistance being a significant factor, contributing to long-term health complications like prediabetes, sleep apnea, heart problems, and mental health issues. Further study is required to improve accuracy based on clinical data, even though machine learning methods such as Random Forest and Support Vector Machines, Logistic Regression, CART4 Classification, Regression Trees, and Naive Bayes Classification have been investigated for PCOS diagnosis [3]. Certain papers recognize the significance of machine learning methods, such as K-NN, Random Forest, SVM, Logistic Regression, and Gaussian Naïve Bayes Which lead to the analysis of the performance of the algorithm used and show rapidly transforming healthcare, offers promise in diagnosing PCOS by analyzing various parameters such as hormone levels, follicle count, and menstrual cycle regularity. Early diagnosis allows for timely lifestyle adjustments, which decreases long-term health risks like type-2 diabetes and cardiovascular disorders. Common symptoms include irregular periods and elevated androgen levels. Detecting PCOS early can mitigate risks of miscarriages, infertility, and, in rare cases, gynecological cancer, underscoring the significance of leveraging technology for improved healthcare outcomes [4]. A few research used machine learning approaches to identify PCOS, a condition that could lead to better diagnosis and treatment. Each of those algorithms you mentioned has its strengths and weaknesses, so comparing their performance for specific task could provide valuable insights. K-Nearest Neighbors (K-NN) is known due to its ease of use and efficiency in classifying tasks, Although Naive Bayes can be applied to different domains, it is commonly employed for text classification. Decision trees are helpful for providing insights into the features that contribute to the classification since they are easily interpretable and comprehended. Support Vector Machines (SVM) are an effective tool for managing difficult classification jobs since they can identify the best hyperplane to divide classes into. Logistic Regression is commonly used when the outcome is binary, making it suitable for classification problems like PCOS identification. It would be intriguing to see how each of these techniques performs and whether any combination or ensemble approach could further enhance accuracy [5]. Some paper trying to show importance of ML that offers promising solutions in healthcare by providing methods to analyze data and make predictions. ML algorithms like C-NN and R-NN can improve diagnostic accuracy with simplify healthcare systems. Advanced techniques have the potential to integrate seamlessly into medical practice, aiding healthcare professionals and enhancing the effectiveness and quality of patient care. In the context of PCOS, leveraging ML algorithms could lead to more precise diagnosis and better management of the condition, ultimately improving outcomes for affected individuals [6]. Combining Naive Bayes, SVM, and KNN into an integrated model for PCOS identification shows promise. Naive Bayes handles feature relationships, SVM captures complex decision boundaries, and KNN detects local patterns. Integration leverages each algorithm's strengths, improving overall performance. It may combine model outputs or use them as features for a higher-level classifier. Techniques like ensemble learning can enhance performance further. Paper offers insights into synergistic machine learning approaches for more accurate PCOS identification, advancing diagnostic tools [7]. Few paper explores machine learning techniques like Irregular Woods Classifier and Convolutional Neural Networks (CNN) for PCOS identification. Irregular Woods Classifier is effective for handling irregular data patterns, while CNN excels at extracting features from image data. These techniques offer promising avenues for improving PCOS identification accuracy. By leveraging ML algorithms, a high-performing diagnostic model can be developed, aiding in accurate and timely identification of PCOS, thereby improving healthcare outcomes for affected individuals [8]. Some paper delves into machine learning techniques such as Irregular TPOT and Cross-Validation for fine-tuning and evaluating performance in the identification of PCOS. TPOT can handle irregular data patterns, it is used, while Cross-Validation ensures robustness and reliability in model assessment. These methods are utilized to enhance the accuracy of PCOS identification [9]. An automated healthcare system based on machine learning is proposed in this paper to diagnose PCOS accurately, Using Non-Linear SVM, Multinomial Logistic regression, Random Forest, Decision Tree, Naïve Bayes, and in conjunction with a deep neural network model. Additionally, it compares several machine learning models using data gathered from Keralan hospitals, India, offering insights into the disease's diagnosis and management [10]. Few studies use a full dataset including clinical and demographic information to predict PCOS risk employing the Random Forest, Decision Tree, and Linear Regression machine learning techniques.The study aims to assess the predictive performance of each algorithm and identify the most effective model for PCOS risk prediction. Early detection of PCOS is vital for symptom management and complication prevention. The goal of the project is to create a prediction model by utilizing cutting-edge machine learning techniques offering early identification and intervention to improve healthcare outcomes for affected women [11]. A article investigates several machine learning methods for PCOS identification, such as AdaBoost, XGBoost, These include Naïve Bayes, K-Nearest Neighbors, Support Vector Machine (with linear, polynomial, Gaussian, and sigmoidal kernels), Decision Tree, Random Forest, and Logistic Regression. These algorithms offer diverse approaches to accurately classify PCOS cases, potentially improving diagnosis and treatment strategies [12]. Several machine learning techniques, such as Random Forest, Logistic Regression, and Support Vector Machines, are studied in a limited number of papers for performance analysis and PCOS case diagnosis. These algorithms are evaluated for their effectiveness in accurately diagnosing PCOS, aiming to enhance diagnostic procedures and outcomes. Machine learning techniques are being explored to aid in PCOS detection and prediction. A study developed a data-driven PCOS detection system using machine learning, which showed promising accuracy in predicting PCOS without clinical testing [13]. A study looks at the efficacy and evaluation of several machine learning techniques for the identification of PCOS, including Support Vector Machine, Decision Tree, Random Forest, and KNN, and Logistic Regression. The study's objective is to develop the best algorithms for precisely detecting PCOS instances through investigation and review, thereby contributing to improved diagnostic capabilities. The use of machine learning techniques is growing in popularity among researchers as a tool for PCOS diagnosis and prediction, leveraging diverse datasets to develop accurate predictive models [14]. Research on the effectiveness of machine learning techniques such as Naïve Bayes, Decision Trees, and Neural Networks for detecting PCOS instances. Through analysis and evaluation, the study aims to ascertain the efficacy of these algorithms in accurately diagnosing PCOS, contributing to advancements in diagnostic methodologies for the condition. By identifying recurring patterns and using classification algorithms, researchers aim to predict PCOS attributes more effectively, ultimately leading to better diagnosis and prognosis for patients. Data mining proves to be a valuable tool in interdisciplinary field, aiding in the extraction of relevant information from vast datasets and contributing to advancements in PCOS research and management [15].

III. TECHNOLOGY USED

A. ADA Boost

Adaptive boosting, or ADA Boost, is a well-liked ensemble learning technique for machine learning tasks including regression and classification. It functions by successively combining several weak learners (typically decision trees), with each new model emphasizing more on the cases that the preceding models misclassified. ADA Boost gives misclassified data points more weights in each iteration, which enables later weak learners to concentrate on cases that are challenging to classify. Through iterative weight adjustments and the combination of the weak learners, A reliable classifier that performs well on complex datasets is produced by ADA Boost. It is known for its adaptability to overfitting and capacity to handle data with multiple dimensions.

B. Random Forest

Another popular ensemble learning method for applications involving regression and classification is Random Forest. It generates the mode of the classes for classification or the average forecast of each individual tree for regression, and it needs a large number of decision trees to be constructed during training in order for it to function. Randomness and variation are added to the ensemble by training each tree in the random forest using a random subset of the training data and features. The ability of the random to decorrelate the trees and reduce overfitting produces a more accurate and reliable model. Random Forest is widely recognized for its user-friendliness, scalability, and effectiveness in handling large datasets with numerous dimensions. It is also less vulnerable to noisy data and outliers than single decision trees.

IV. RESEARCH METHODOLOGY

1. The "PCOS_data.csv" dataset on the Kaggle website provided the dataset used in this investigation. This dataset has 44 columns with data related to various physiological, medical, and demographic aspects of

PCOS.

2. For more analysis, a portion of the features from the original dataset were selected. During the selection process, Nineteen features is choosen, including test results, body mass index (BMI), data on pulse rate, and other pertinent characteristics linked to the early identification and treatment of PCOS for a better life and future.
3. The chosen dataset was split 90:10 across training and testing sets. This makes sure that enough data is set apart for the models' training and that some data is kept separate for assessing the models' effectiveness.
4. For model training, two classification algorithms were used: the ADA Boost classifier and the Random Forest classifier. These ensemble learning methods work well with multidimensional data and are frequently applied to machine learning classification challenges. The selected classifiers were trained to search the data for similarities and correlations and forecast future occurrences using the training dataset. In training, the classifiers build robust predictive models by successively adapting to misclassified cases (ADA Boost) or iteratively building decision trees (Random Forest). The testing dataset was used to assess the classifiers' performance after they had been trained on the model. To evaluate how well the models performed in accurately categorizing PCOS cases, evaluation measures like precision and accuracy scores were calculated. Precision evaluates the proportion of true positive predictions among all positive predictions, whereas accuracy measures the classifier's overall prediction accuracy.
5. The evaluation phase data were examined to see how well the Random Forest and ADA Boost classifiers diagnosed PCOS. The models' viability for practical use in healthcare settings was determined by the precision and accuracy ratings, which offered information on the models' capacity to accurately categorize PCOS cases and non-cases.

V. RESULT AND ANALYSIS

Using machine learning approaches, the problem of overfitting and underfitting was successfully resolved in the PCOS detection process. Through meticulous feature selection and suitable model training techniques, the models were able to attain a balance between intricacy and broadness. This improved the reliability of PCOS diagnosis by guaranteeing that the classifiers could capture the underlying patterns in the data accurately without learning noise (overfitting) or oversimplifying the correlations (underfitting). When it came to performance comparison, Random Forest fared better in detecting PCOS than the Ada Boost classifier. The Random Forest classifier outperformed Ada Boost in terms of accuracy and precision scores, demonstrating its efficacy in correctly classifying PCOS cases according to the chosen criteria. As a result of its sensitivity to noisy data or inadequate feature representation, Ada Boost, on the other hand, performed worse in PCOS diagnosis. This demonstrates how important it is to apply machine learning algorithms that are appropriate for the particulars of the work at this point and the dataset.

Table I: Accuracy Score, Precision Score, F1 Score and R2 Score on Models

Algorithm used	Accuracy	Precision	F1	R2
Random Forest	89	80	85	51
Ada Boost	95	90	92	75

VI. CONCLUSION

In summary, machine learning presents a viable method in order to identify Polycystic Ovary Syndrome (PCOS) early, enabling prompt intervention. Through the use of a variety of patient datasets that include physical characteristics, hormone profiles, and medical histories, machine learning algorithms are able to identify patterns that are suggestive of PCOS. However, the selection of the algorithm has a significant impact on the detection system's efficacy. When it comes to PCOS identification, ADABoost outperforms the Random Forest Classifier. Targeting misclassified examples, its adaptive boosting method allows for iterative improvements that improve accuracy and generalization. ADABoost outperforms the Random Forest Classifier in every evaluation metric, such as R-squared, F1 score, accuracy, and precision (for regression problems). This demonstrates its ability to grasp complex data correlations and produce predictions that are more accurate.

REFERENCES

[1] Mehrotra, P., Chatterjee, J., Chakraborty, C., Ghoshdastidar, B., & Ghoshdastidar, S. (2011, December). Automated screening of polycystic ovary syndrome using machine learning techniques. In 2011 Annual IEEE India Conference (pp. 1-5). IEEE.

[2] Tanwani, N. (2020). Detecting PCOS using machine learning. Int J Modern Trends Eng Sci (IJMTES), 7(1), 1-20.

[3] Hassan, M. M., & Mirza, T. (2020). Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. Int. J. Comput. Appl, 975, 8887.

[4] Thakre, V., Vedpathak, S., Thakre, K., & Sonawani, S. (2020). PCOcare: PCOS detection and prediction using machine learning algorithms. Biosci Biotechnol Res Commun, 13(14), 240-244.

[5] Chauhan, P., Patil, P., Rane, N., Raundale, P., & Kanakia, H. (2021, June). Comparative analysis of machine learning algorithms for prediction of pcos. In 2021 international conference on communication information and computing technology (ICCICT) (pp. 1-7). IEEE.

[6] Vaswania, J., Mulchandanib, H., Vaghelac, R., & Pateld, R. (2022). A Systematic literature review on diagnosis of PCOS using machine learning algorithms. GIT J. Eng. Technol, 14(5).

[7] Alagarsamy, M., Shanmugam, N., Mani, D. P., Thayumanavan, M., Sundari, K. K., & Suriyan, K. (2023). Detection of polycystic syndrome in ovary using machine learning algorithm. International Journal of Intelligent Systems and Applications in Engineering, 11(1), 246-253.

[8] Jyothi, R., Shivani, H. C., Yashaswi, R., & Sumanth, R. (2023, June). Detection of Polycystic Overy Syndrome (PCOS) Using Machine Learning Techniques. In 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA) (pp. 261-266). IEEE.

[9] Yadav, N., & Pande, S. D. (2024). Comparative Analysis of Polycystic Ovary Syndrome Detection Using Machine Learning Algorithms. EAI Endorsed Transactions on Pervasive Health and Technology, 10.

[10] Dhall, I., Vashisth, S., & Aggarwal, G. (2024). Smart Healthcare System for Reliable Diagnosis of Polycystic Ovary Syndrome. In Artificial Intelligence and Machine Learning Techniques in Image Processing and Computer Vision (pp. 19-36). Apple Academic Press.

[11] Priyadharshini, M., Srimathi, A., Sanjay, C., & Ramprakash, K. (2024). PCOS Disease Prediction Using Machine Learning Algorithms. International Research Journal on Advanced Engineering Hub (IRJAEH), 2(03), 651-655.

[12] Khanna, V. V., Chadaga, K., Sampathila, N., Prabhu, S., Bhandage, V., & Hegde, G. K. (2023). A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. Applied System Innovation, 6(2), 32.

[13] Batra, H., & Nelson, L. (2023). DCADS: Data-Driven Computer Aided Diagnostic System using Machine Learning Techniques for Polycystic Ovary Syndrome. International Journal of Performability Engineering, 19(3), 193.

[14] Dutta, P., Paul, S., & Majumder, M. (2021). An efficient SMOTE based machine learning classification for prediction & detection of PCOS.

[15] Vikas, B., Anuhya, B. S., Chilla, M., & Sarangi, S. (2018). A critical study of Polycystic Ovarian Syndrome (PCOS) classification techniques. Int J Comput Eng Manag, 4, 1-7.