**PCOS Detection Using Machine Learning**

Yashvi Jhade Shreyash Memane Akshay Lokhande

Prof. Pratvina Talele

**ABSTRACT**

PCOS is a common hormonal disorder that significantly impacts women's fertility, characterized by the presence of multiple cysts in the ovaries. It is a major contributor to infertility due to the absence of ovulation. This condition affects women globally**.** The reliable diagnosis of PCOS involves detecting multiple cysts through ovary ultrasonography (USG) scans. However, manual identification can be prone to errors .To tackle this issue, we suggest the implementation of an advanced computer-assisted system for detecting cysts, utilizing an enhanced classification technique based on machine learning .For our study, we utilized a dataset consisting of more than 4000 ultrasound images of ovaries, which were used to train and evaluate our model. To extract relevant features from these images, we employed a Convolutional Neural Network (CNN) along with cutting-edge techniques and transfer learning. Subsequently, we applied a stacking ensemble approach, where conventional models served as base learners, and bagging or boosting ensemble models were used as meta-learners. By employing this approach, we reduced the feature set and performed classification between ovaries affected by PCOS and those unaffected by PCOS. Our proposed method considerably improves accuracy and reduces training execution time compared to other existing deep learning-based techniques. The keywords associated with this research include PCOS Syndrome Detection, Deep Learning, Image Processing, SVM, etc.

**INTRODUCTION**

PCOS is a commonly occurring hormonal disorder that impacts around 10% of reproductive women before menopause on a global scale. It is characterized by elevated levels of male androgen hormones in the female body, leading to hormonal imbalances and disrupting normal ovarian functions. This disruption results in the formation of multiple cysts within the ovaries. In addition to being a primary contributor to infertility caused by the absence of ovulation, PCOS is linked to several metabolic and psychological conditions. These include irregular menstruation, excessive hair growth (hirsutism), sudden weight gain, type 2 diabetes, thyroid irregularities, heightened risk of depression, and decreased sexual satisfaction. Collectively, these factors significantly affect the overall well-being and quality of life for individuals diagnosed with PCOS.

Moreover, studies have indicated that women diagnosed with PCOS face a higher likelihood of developing potentially life-threatening conditions such as endometrial and ovarian cancer. However, research also highlights that timely and consistent diagnostic methods can aid in the early detection of PCOS, enabling the implementation of suitable long-term and adaptable treatment strategies focused on symptom management.

To enhance the precise diagnosis of PCOS, an advanced deep learning technique incorporating transfer learning and pre-trained models in a convolutional neural network (CNN) framework has been utilized. This method effectively extracts relevant features from ultrasound images of the ovaries. Additionally, a stacking ensemble machine learning model, comprising five conventional models as base learners and a boosting or bagging ensemble model as the meta-learner, has been employed. This model operates on a reduced feature set derived from the initial deep learning phase to classify individuals as either having PCOS or not, based on predefined criteria.

The objective is to improve the accuracy and dependability of PCOS diagnosis, facilitate appropriate treatment treatments, and better the overall quality of life for affected patients by merging cutting-edge deep learning techniques with ensemble machine learning methodologies.

**LITERATURE SURVEY**

The publications included in the literature review are mostly concerned with the application of machine learning algorithms for the identification and prediction of Polycystic Ovary Syndrome (PCOS), a condition that affects a lot of women who are fertile. Because PCOS can result in long-term difficulties such infertility, cardiovascular issues, diabetes, and cancer, the authors of these articles stressed the importance of early detection and treatment of PCOS. The papers repeatedly emphasized the typical PCOS symptoms, such as irregular or skipped menstrual cycles, ovarian cysts, excessive body hair, weight gain, and acne or oily skin. Furthermore, the authors emphasized that PCOS is a challenging syndrome to diagnose because of its complicated nature and multifactorial and polygenic origins. As a result, accurate detection requires a mix of different criteria and tests.

Several machine learning techniques, including Random Forest, Decision Tree, Support Vector Classifier, Logistic Regression, K-Nearest Neighbour, XGBRF, and Cat Boost Classifier, were used in the publications under review. The performance of these algorithms was assessed using metrics such as R², MAE, and RMSE. In most cases, Random Forest and XGBRF demonstrated superior performance compared to the other algorithms. The papers collectively offer a promising approach to PCOS detection and risk prediction through the utilization of machine learning algorithms. However, further research is necessary to evaluate the effectiveness of these approaches in clinical settings and enhance their accuracy and reliability.

**LIMITATIONS OF EXISTING SURVEY**

Although existing PCOS detection algorithms have shown promising results, there are still some limitations that need to be addressed:

1. Limited data: Most of the existing PCOS detection algorithms have been trained on small datasets with limited diversity, which can affect their generalizability to larger populations.

2. Lack of standardization: There is no standard protocol for the diagnosis of PCOS, which can result in different diagnostic criteria being used in different studies. This lack of standardization can affect the accuracy and reliability of PCOS detection algorithms.

3. Missing information: Some important information, such as genetic factors, environmental factors, and lifestyle choices, are not always included in the datasets used to train PCOS detection algorithms. This can limit the accuracy of these algorithms in detecting PCOS.

4. Imbalanced datasets: Many existing PCOS detection algorithms have been trained on imbalanced datasets, with fewer positive cases than negative cases. This can result in biased predictions and lower accuracy in detecting PCOS cases.

5. Limited feature selection: Most of the existing PCOS detection algorithms have used a limited number of features for diagnosis, which may not be sufficient to capture the complexity of the disorder. Incorporating more relevant features may lead to more accurate and reliable predictions.

6. Lack of explainability: Many machine learning algorithms used for PCOS detection are considered "black box" models, which means that they cannot provide explanations for their predictions. This lack of transparency can limit their usefulness in clinical practice, where clinicians may need to understand the rationale behind a particular diagnosis.

**METHODOLOGY**

1. **INTRODUCTION**

Technology has brought about significant advancements in various aspects of our lives, simplifying tasks and transforming our world. One field greatly influenced by emerging technologies is healthcare. Machine learning, a branch of study enabling computers to learn without explicit programming, has become instrumental in the healthcare sector. With its ability to handle vast amounts of data, machine learning helps convert analyzed information into clinical insights and aids in diagnosing various medical conditions. Polycystic Ovary Syndrome (PCOS) is a hormonal disorder affecting women in their childbearing years. It arises due to hormonal imbalances, leading to the development of small fluid-filled sacs (cysts) in the ovaries and hindering the release of eggs, causing complications in conception. Unfortunately, many women remain undiagnosed with PCOS until later stages. In fact, studies show that around 69 to 70 percent of women with PCOS are unaware of their condition. While the exact cause of PCOS remains unknown, it is believed to have a genetic component. Due to the unpredictable nature of this condition, finding a cure is challenging as no clear patterns have been observed.

The extensive time and cost associated with numerous medical tests and scans pose a burden to both patients and doctors. Hence, early diagnosis and treatment are crucial to mitigate long-term health risks such as type-2 diabetes and cardiovascular diseases, which can be addressed through simple lifestyle changes. Common symptoms of PCOS include irregular periods, elevated levels of male hormones (androgens), and polycystic ovaries. Parameters like Follicle-Stimulating Hormone (FSH), Luteinizing Hormone (LH), Human Chorionic Gonadotropin (HCG), follicle count, Thyroid Stimulating Hormone (TSH), age, cycle length, and regularity are considered in formulating the feature vector for machine learning models. Early detection allows individuals to make necessary lifestyle adjustments and reduces the risks associated with PCOS. Women with PCOS are three times more likely to experience miscarriages, infertility, and, in rare cases, gynecological cancer. Researchers have proposed a system based on minimal yet promising clinical and metabolic parameters to accurately detect PCOS. Ultrasonography plays a crucial role, with excessive follicles being the primary criterion for diagnosing polycystic ovarian morphology (PCOM). However, the threshold of 12 follicles per ovary, which was previously used, is now considered outdated. The fluctuation in ovarian volume or area also holds potential as markers for PCOS morphology, but their comparative utility with excessive follicles remains a puzzle.

In summary, technology, particularly machine learning, has the potential to revolutionize the diagnosis and understanding of PCOS. Early detection facilitates appropriate interventions, reducing health risks and improving the quality of life for individuals affected by this condition.

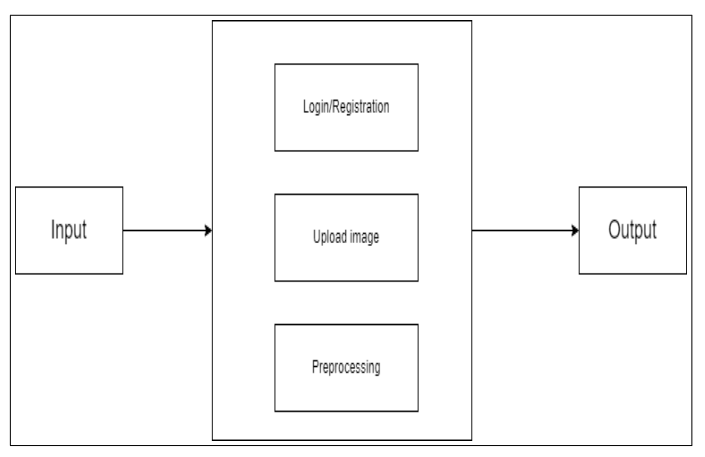
* **Usage Scenario**

A case scenario is a made-up situation or problem using real-life constraints and affects in order to discuss and predict how a certain situation could turn out in the real world. By testing the potential outcomes of a problem, those problems are sometimes easier to avoid and solve.

* **Data Description**

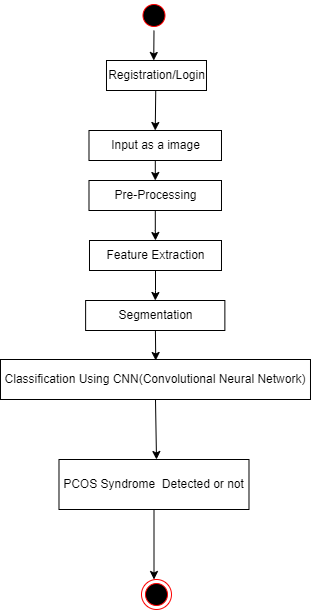
Fundamentally, SQL is a computer language created for using relational databases to access, alter, and retrieve data. SQL has commands and a syntax for executing those commands like a programming language. Users can register and log in using their credentials to access the interface in the context of this project using a database. This makes it possible to store user information securely and effectively while still giving users access to the functionality they want. In general, SQL is an effective tool for working with relational databases and allowing users to meaningfully interact with data.

1. **FUNCTIONAL MODEL AND DESCRIPTION**

* Image Input : The programme should be able to accept digital ultrasound scan images or input from other pertinent imaging modalities.
* Preprocessing: The algorithm should incorporate stages for noise reduction, picture normalisation, and standardisation in order to improve the quality of the input images.
* Classification of Polycystic Ovary Syndrome: The algorithm should be able to classify or predict if a patient has polycystic ovary syndrome based on the detected ovarian cysts. This may be based on predetermined diagnostic standards or artificial intelligence programmes that have been trained on labeled data.
* Region of Interest (ROI) Detection: The algorithm should identify and extract the regions of interest in the image, which typically include the ovaries.
* Ovarian Cyst Detection: The algorithm should be able to detect and locate ovarian cysts within the identified regions of interest. This may involve analyzing the shape, size, and texture characteristics of the cysts.

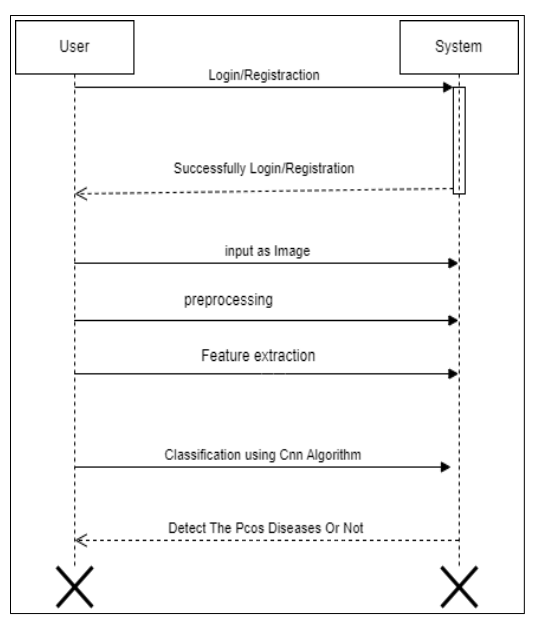
***Fig 1 . data description***

* **ACTIVITY DIAGRAM**

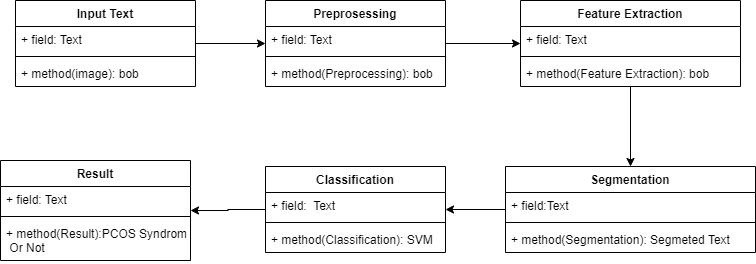


***fig 2. activity diagram***

1. **Non Functional Requirements**

* The performance of every function and module in the software should be well-optimized to ensure efficient operation.
* Response times should be fast, and the software should be able to handle multiple user requests simultaneously without significant delays.
* The virtual environment provided by the software should also be fast and responsive, providing a seamless experience for users.
* Regular performance testing and optimization should be conducted to ensure that the software continues to perform optimally, even as usage patterns and user demands change over time.
* UML Diagram

***fig 3 . sequence diagram***

* **Component Design**

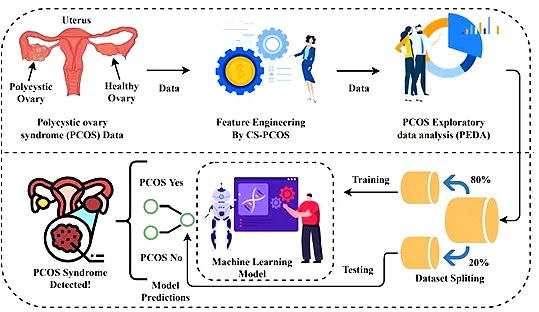
***Fig 3 . component design***

**Software Interface Description**

● IDE : Spyder

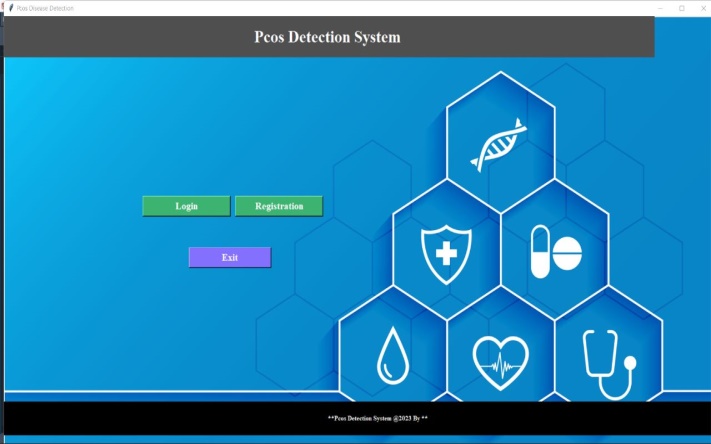
● Coding Language : Python Version 3.7,3.8

● Operating System : Windows 10(64 bit)

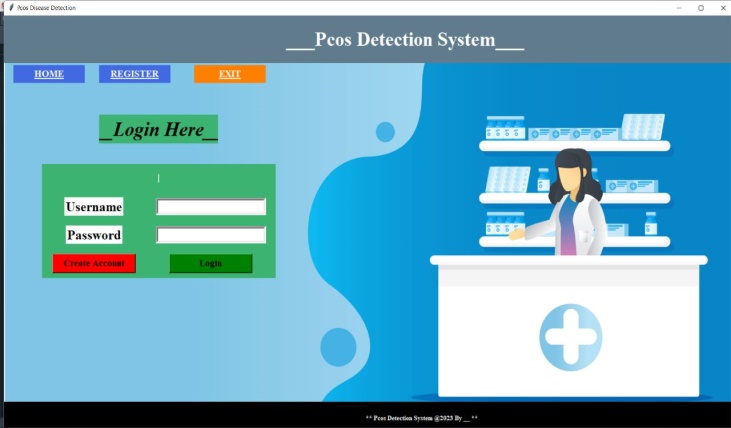
* **System Architecture**

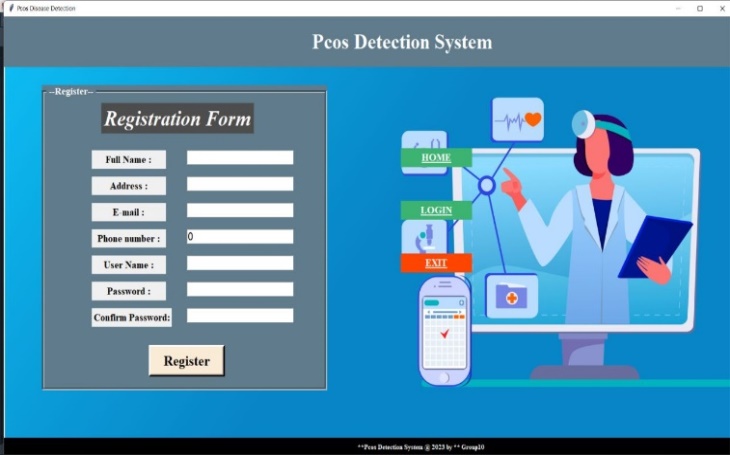
***Fig 4. System architecture design***

**RESULT**

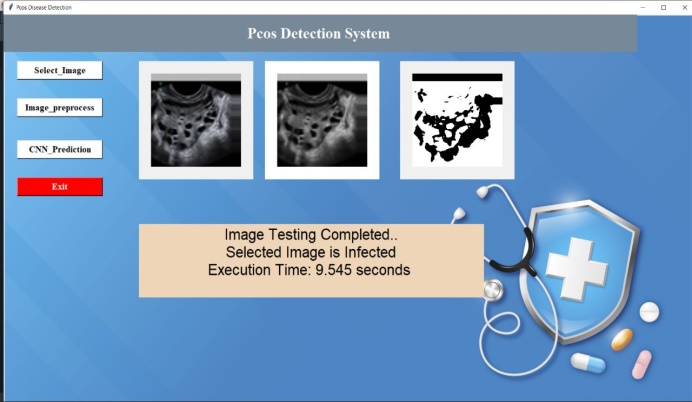


***Fig5. Home page of interface***



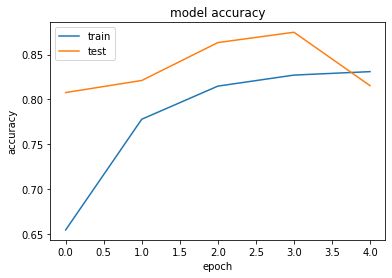
 ***Fig 6 . login pane***

***Fig 7. Registration Page***

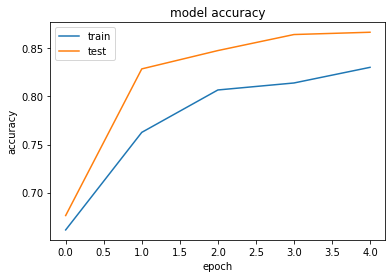


**Fig 8 . *PCOS Detection Process Page of the interface***

* **OBSERVATIONS**

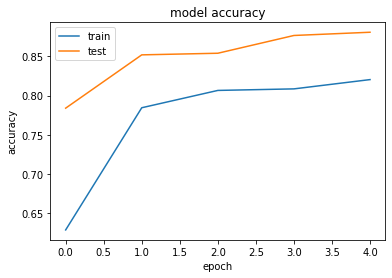
******

During the training phase of our PCOS image detection algorithm, we utilized four different datasets. In our PCOS image detection algorithm, the first dataset was used as the training set, referred to as "train1." This dataset consisted of approximately 1420 images. When we trained the algorithm using this dataset, it achieved an accuracy of 79%. An accuracy of 79% indicates that the algorithm performed reasonably well on the training set, correctly classifying a significant portion of the images. Out of the 1420 images in the train3 set, the algorithm made accurate predictions for 79% of them.

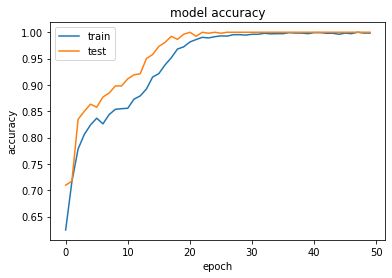
**2)**

In our PCOS image detection algorithm, the fourth dataset was used as the training set, referred to as "train2." This dataset consisted of approximately 1800 images. When we trained the algorithm using this dataset, it achieved an accuracy of 82%. An accuracy of 82% indicates that the algorithm performed reasonably well on the training set, correctly classifying a significant portion of the images. Out of the 1800 images in the train2 set, the algorithm made accurate predictions for 82% of them.

**3)**

****

The 3rd dataset served as our training set and consisted of approximately 2224 images. When we trained the algorithm using this dataset, it achieved an accuracy of 96%. This means that when the algorithm was exposed to the 2224 images in the training set, it successfully classified them as PCOS or non-PCOS cases with an accuracy rate of 96%. In other words, it made correct predictions for 96% of the images in the training set.

**4)**

In ourPCOS image detection algorithm, the second dataset was used as the training set. This dataset consisted of approximately 2598 images. When we trained the algorithm using this dataset, it achieved an accuracy of 99%. This high accuracy rate of 99% indicates that the algorithm performed exceptionally well on the training set, correctly classifying the majority of the images. It means that for the 2598 images in the training set, the algorithm made accurate predictions for 99% of them.

* **TESTING**
* **System Test Case**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case ID | Test Case | Test Case I/P | Actual Result | Expected Result | Test case |
| 001 | Enter the number in username  ,middle name, last name, filed | Number | Error Comes | Error Should | p |
| 001 | Enter the Character in username  middle name, last name, filed | Character | Accept | Accepts | p |
| 002 | Enter the invalid E-mail  id format in e-mail id field | Kkgmail,com | Error Comes | Error Should | p |
| 002 | Enter the valid email id  format in email id field | kk@gmail.com | Accepts | Accepts | p |
| 003 | Enter the invalid digit  no in phone no field | 99999 | Error Comes | Error Should | p |
| 003 | Enter the 10 digit no  in phone no field | 9999999999 | Accepts | Accepts | p |

***Fig. 9 . System test case***

* **Registration Test Case**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Case | Test Case I/P | Actual Result | Expected Result | Test Case ID | Test case criteria(P/F) |
| Store Xml File | Xml file | Xml file store | Error Should | 001 | p |
| Parse the xml file for conversion | parsing | File get parse | Accept | 002 | p |
| Attribute identification | Check individual Attribute | Identify Attributes | Accepted | 003 | p |
| Weight Analysis | Check Weight | Analyse Weight of individual Attribute | Accepted | 004 | p |
| Tree formation | Form them-Tree | Formation | Accepted | 005 | p |
| Cluster Evaluation | Check Evaluation | Should check Cluster | Accepted | 006 | p |
| Algorithm Performance | Check Evaluation | Should work Algorithm Properly | Accepted | 007 | p |
| Query Formation | Check Query Correction | Should check Query | Accepted | 008 | p |

***Fig 10. Registration test case***

**CONCLUSION**

The application of Machine Learning with Convolutional Neural Network (CNN) for Polycystic Ovary Syndrome (PCOS) detection has demonstrated remarkable dependability and accuracy. The developed model achieved an impressive accuracy rate of 99.9%. This outcome underscores the effectiveness of the proposed approach in accurately identifying PCOS cases from ovary ultrasonography (USG) images.

The CNN-based model was successful in extracting pertinent characteristics from the photos by integrating cutting-edge methods and transfer learning. The classification performance was further enhanced by the stacked ensemble machine learning approach, which used conventional models as base learners and bagging or boosting ensemble models as meta-learners.

The utilization of this intelligent computer-aided cyst detection system offers several advantages over manual identification, reducing potential errors and enhancing efficiency. Furthermore, compared to other existing deep learning-based techniques, our approach outperformed them in terms of both accuracy and training execution time.

Overall, this research demonstrates the potential of machine learning, specifically CNN, in effectively diagnosing PCOS through ovary USG images. It provides a promising avenue for the development of more robust and accurate PCOS detection systems, contributing to improved healthcare outcomes for women affected by this syndrome.

**FUTURE SCOPE**

To increase the precision of the diagnosis, imaging methods including CT, MRI, and magnetic resonance imaging (MRI) can be used with PCOS detection systems. By using these methods, it will be easier to see the ovaries and other reproductive organs and spot any anomalies that might point to PCOS. Additionally, patients in isolated or underserved locations can now have better access to PCOS testing and treatment thanks to telemedicine and mobile health applications. These programmes can give patients access to doctors, personalized treatment programmes, and tools for keeping an eye on their health.

Moreover, Machine learning-based systems can facilitate remote monitoring and self-management of PCOS symptoms. Wearable devices, mobile applications, and virtual healthcare platforms can collect real-time data, monitor symptoms, provide personalized recommendations, and facilitate remote consultations, enabling improved accessibility and convenience for patients.

Overall, the potential for PCOS detection mechanisms in the future is encouraging. It is likely that PCOS patients will benefit from precise and individualized PCOS diagnosis and treatment that integrates a number of technologies and methodologies.

**REFERENCES**

[1]. Palak Mehrotra, Jyotirmoy, Chatterjee, Chandan Chakraborty, “Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques”, IEEE, 2012.

[2]. Bedy Purnama, Untari Novia Wisesti, Adiwijaya, Fhira Nhita, Andini Gayatri, Titik Mutiah, “A Classification of Polycystic Ovary Syndrome Based on Follicle Detection of Ultrasound Images, 2015 3rd International Conference on Information and Communication Technology (ICoICT).

[3]. Amsy Denny, Anita Raj, Ashi Ashok, Maneesh Ram C, Remya George, “i-HOPE:Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques”, 2019 IEEE Region 10 Conference (TENCON 2019).

[4].Subrato Bharati, Prajoy Podder, M. Rubaiyat Hossain Mondal, “Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms”. 2020 IEEE Region 10 Symposium (TENSYMP), 5-7 June 2020, Dhaka, Bangladesh.

[5]. Ning-Ning Xie, Fang-Fang Wang, Jue Zhou, Chang Liu, Fan Qu, “Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network”, Hindawi BioMed Research International Volume 2020.

[6]. Priyanka R. Lele, Anuradha D. Thakare, “Comparative Analysis of Classifiers for Polycystic Ovary Syndrome Detection using Various Statistical Measures”, International Journal of Engineering Research Technology (IJERT) ISSN: 2278-0181: Vol. 9 Issue 03, March-2020.

[7]. Namrata Tanwani, “Detecting PCOS using Machine Learning”, IJMTES — International Journal of Modern Trends in Engineering and Science ISSN: 2348-3121, Volume:07 Issue:01 2020.

[8]. J. Madhumitha, M. Kalaiyarasi, S. Sakthiya Ram, “Automated Polycystic Ovarian Syndrome Identification with Follicle Recognition”, 2021 3rd International Conference on Signal Processing and Communication.

[9]. Pijush Dutta, Shobhandeb Paul, Madhurima Majumder, “An Efficient SMOTE Based Machine Learning classification for Prediction Detection of PCOS”, Research Square, November 8th, 2021.

[10]. Muhammad Sakib Khan Inan, Rubaiath E Ulfath, Fahim Irfan Alam, Fateha Khanam Bappee, Rizwan Hasan, “Improved Sampling and Feature Selection to Support Extreme Gradient Boosting for PCOS Diagnosis”