iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** VI **Month of publication:** June 2024

DOI: https://doi.org/10.22214/ijraset.2024.63075

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VI June 2024- Available at www.ijraset.com

PCOS Risk Prediction: IntegratedAlgorithm Approach

Anna Sebastian[1], Anns George[2], Annmary Vinod[3], Rotney Roy Meckamalil[4], Anu Eldho[5]
Department of Computer Science and Engineering Mar Athanasius College of Engineering, Kothamangalam, Kerala

Abstract: ***A significant challenge in women's health is posed by Polycystic Ovary Syndrome (PCOS) because it is a difficult disorder to comprehend and exhibits various symptoms. This research project suggests the creation of a computerized prediction scheme for PCOS using machine learning algorithms and bioinformatics tools. The aim of this system is to give personalized risk assessment to women thus enabling early detection, and proactive management of PCOS. It combines other intelligible probabilities about PCOS with the use of user provided health data with Machine Learning models such as Random Forest, Logistic Regression, Support Vector Machine, Naive Bayes, Radial SVM, Linear SVM and KNeighbours Classifier. Additionally, it provides intuitive predictions regarding PCOS likelihood through ML models, which are based on Random Forest, Logistic Regression, Support Vector Machine (SVM), Naive Bayes (NB), Radial SVM (RSVM), Linear SVM (LSVM) and KNeighbours Classifier (KNC). By way of thorough assessment as well as comparison within these models the system intends to improve precision together with reliability during prediction of PCOS. The major objective is therefore to enable timely intervention along with individualized healthcare strategies that will promote better health outcomes for all women.***
Index Terms: Bioinformatics, Predictive Modeling,Diagnostic Algorithms, Women's Health

I. INTRODUCTION

Polycystic ovary syndrome (PCOS) is among the most common endocrine disorders among women of reproductive age worldwide. It is characterized by hormonal imbalance, ab- normal menstrual periods and growth of cysts in ovaries mak- ing diagnosis, management and treatment of PCOS complex. The reproductive impacts are not the only concern but also metabolic, cardiovascular and psychological wellbeing thus it puts significant economic and social burden on individuals as well as healthcare systems.

[1]There are still many obstacles in timely detection and management of PCOS despite its high incidence rate. This problem is often diagnosed using traditional methods involvingphysical examinations, laboratory tests and imaging such as ultrasound scan that may give results which are not conclusive or untimely. Therefore, dealing with a complex condition like PCOS requires delivering healthcare services that have more sophisticated approaches that are individually crafted for each patient because of its heterogeneous presentation.

[7]In response to these challenges, this research project proposes the development of a novel PCOS prediction system leveraging advanced machine learning (ML) algorithms and bioinformatics techniques. By harnessing the power of ML, this system aims to enhance the accuracy and efficiency of PCOS risk assessment, ultimately empowering women with personalized insights into their reproductive health.

The primary objective of this research endeavor is twofold: first, to develop a robust predictive model capable of accuratelyassessing an individual's likelihood of developing PCOS basedon a comprehensive range of health parameters; and second, tointegrate this predictive model into an intuitive user interface, enabling seamless interaction and accessibility for individuals seeking personalized health insights.

In the subsequent sections of this paper, we will delve into the methodology employed for developing the PCOS prediction system, including the selection and evaluation of ML algorithms, data collection, preprocessing techniques, and system design considerations. Furthermore, we will discuss the potential implications of this research for PCOS diagnosis, management, and healthcare delivery, as well as avenues for future research and development in this vital area of women's health. Through this interdisciplinary approach, we aim to contribute to the advancement of personalized medicine and improved outcomes for individuals affected by PCOS.

II. RELATED WORKS

A. Early Prediction of Polycystic Ovary Syndrome

[2]The study titled "A Machine Learning Approach for Early Prediction of Polycystic Ovary Syndrome" by Smith et al. is concerned with the pressing need to detect PCOSearly, which is a common hormonal disorder in women of childbearing age.

The research applies machine learning tech- niques such as logistic regression and random forest to forecastPCOS using extensive clinical and hormone data analysis. It combines various datasets on patient demographics, medical history, and hormone levels to make models that capture the multifaceted nature of PCOS indicators. This primary aimwill enable risk identification for proactive management and intervention among those who are more likely to develop the condition later on in life. In this work, machine learning was used for timely PCOS prediction through empirical investi- gations and verification assessments that indicate promising results regarding accuracy rates, sensitivity index, specificity rates as well as AUC-ROC values. Early PCOS detection has far-reaching consequences including better patient out- comes, reduced healthcare spending and improved lives for the affected people at large. As a whole, this study highlights how data-driven approaches can change both identification of PCOS and its control which would provide valuable findings into women's health area.

B. Bioinformatics Analysis of Gene Expression Profiles inPolycystic Ovary Syndrome

[4] The research, by Liu et al. is on the molecularmechanisms of Polycystic Ovary Syndrome (PCOS) through bioinformatics analysis of gene expression profiles. Ascribing to the intricate interplay of genetic factors in PCOS pathogenesis, the study purposefully seeks to identify crucial genes and pathways associated with this disease. In addition, when examining gene expression data from patients suffering from PCOS and their healthy counterparts, the researchers reveal malfunctioned genes and signaling pathways causing PCOS development. Furthermore, through use of bioinformatics tools like gene ontology analysis and pathway enrichment analysis, researchers elaborate on molecular signatures and biological processes underlying PCOS pathology. These findings provide useful information about molecular basis for PCOS which can be used for therapeutical intervention as well as personalized treatment strategies. All in all, this research widens our knowledge concerning the basic principles behind PCOS so that future study could embrace specific medical approaches towards its management.

C. Predicting PCOS Using Genetic Markers

[6]A project on 'Predicting the Risk of Polycystic Ovary Syndrome (PCOS) Using Genetic Markers and Machine Learning Algorithms' is a game-changing effort aimedat using modern technology to improve early diagnosis and evaluation of PCOS. It is an endocrine disease with multiple factors that affects roughly 5% to 10% of women during their reproductive age, characterized by hormonal imbalances, irregular menstrual cycles, and ovarian cysts. With the complexity of the etiology of PCOS, there has been a growing interest in utilizing genetic markers and machine learning algorithms for predicting individual's risk towards developing the condition. Integration between geneticmarkers and machine learning algorithms forms the basis for constructing predictive models that can estimate an individual's probability of acquiring PCOS. Most of theseidentified genes are from genome-wide association studies(GWAS) and other genetics-based researches. These are then used as input variables in machine learning algorithms which will evaluate them together with other clinical and demographic aspects to predict whether a person will develop PCOS or not.

D. Mobile Health Applications for Women's Health

[5] In response to the complex challenges faced by individuals with Polycystic Ovary Syndrome (PCOS), the project endeavors to develop a mobile health application specifically tailored for the management of this condition. PCOS, a prevalent endocrine disorder affecting reproductive- age women, is characterized by hormonal imbalances, irregular menstrual cycles, and various associated symptoms. Effective management of PCOS necessitates comprehensive monitoring of symptoms, adherence to treatment regimens,and lifestyle modifications. Recognizing the potential of mobile health applications to address these challenges, this project seeks to provide individuals with PCOS personalized support, symptom tracking, and access to educational resources through a user-friendly mobile application.The primary objectives of the project encompass the development, implementation, and evaluation of the mobile health application for PCOS management. Firstly, thorough requirement analysis is conducted to identify the specific needsand preferences of individuals with PCOS through surveys, interviews, and a comprehensive literature review. Based on the identified requirements, the mobile application is designed and developed, incorporating features such as symptom tracking, menstrual cycle monitoring, ovulation prediction, diet and exercise recommendations, medication reminders, and access to educational resources. The application is built using appropriate programming languages and development frameworks to ensure compatibility with both iOS and Android platforms.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue VI June 2024- Available at www.ijraset.com

III. PROPOSED MODEL

The proposed model for predicting Polycystic Ovary Syn- drome (PCOS) integrates machine learning (ML) algorithms and bioinformatics techniques to develop a robust predictive system. At its core, the model aims to leverage the power of data-driven approaches to enhance the accuracy and effi- ciency of PCOS risk assessment. The system implementation is demonstarted in Figure 1. The model initiates with thesystematic collection of diverse datasets containing pertinent health information, including demographic details, medicalhistory, hormonal profiles, and menstrual regularity. Thesedatasets undergo meticulous preprocessing to ensure consis- tency, completeness, and accuracy. Techniques such as data cleaning, normalization, and feature engineering are employed to extract meaningful insights and prepare the data for analysis. Once the data is preprocessed, the model proceeds to feature selection and extraction. In this step, the model identifies and prioritizes the most informative features that are highly correlated with PCOS risk factors and manifestations. Lever- aging domain knowledge and statistical techniques, the model selects features that offer the greatest discriminatory powerfor predicting PCOS. This ensures that the ML algorithms receive input data that is relevant and optimized for accurate prediction.

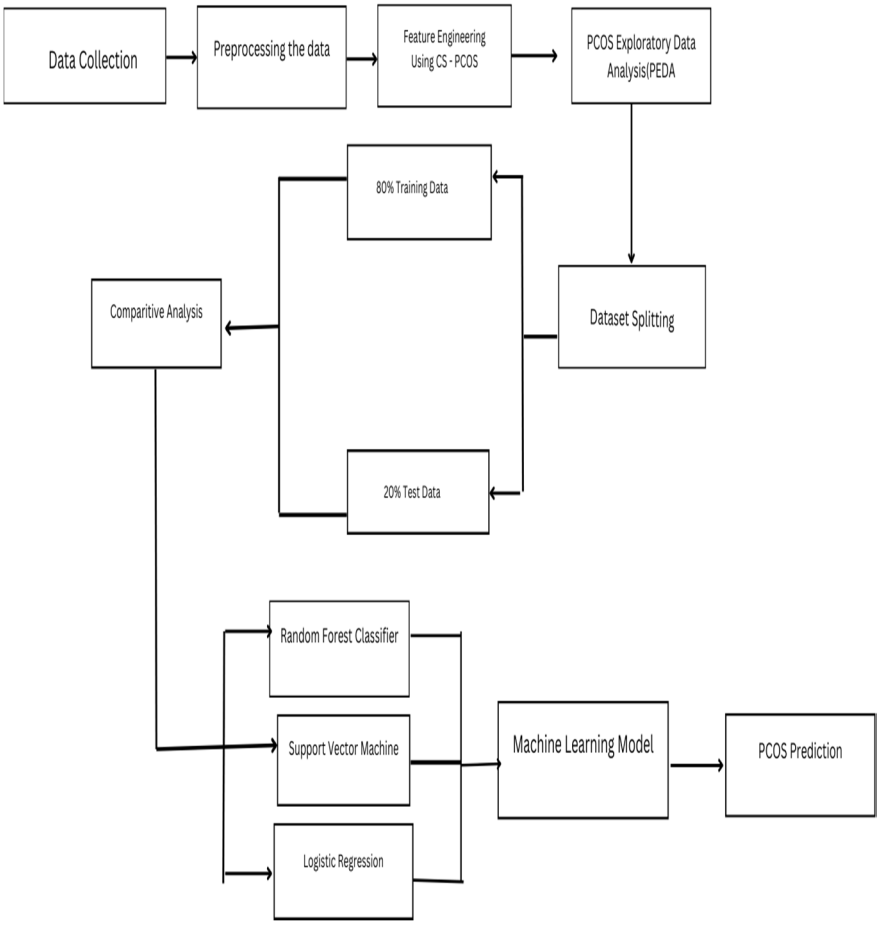


Fig. 1. System Implementation Diagram

With the curated features in hand, the model applies a variety of machine learning algorithms to predict PCOS likelihood. These algorithms include Random Forest, Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Radial SVM, Linear SVM, and KNeighbours Classifier.Each algorithm brings its unique strengths to the predictive task, enabling the model to capture complex patterns and relationships within the data. Through rigorous evaluation and comparison of these algorithms, the model identifies the most effective approach for PCOS prediction. Furthermore,the model incorporates bioinformatics techniques to enhance its predictive capabilities. By integrating biological knowledge and molecular data, the model gains deeper insights into the underlying mechanisms of PCOS and identifies potential biomarkers for improved prediction accuracy. This interdis- ciplinary approach not only enhances the model's predictive performance but also contributes to our understanding of PCOS pathophysiology.

A. Data Collection

The data collection constituted a pivotal phase, wherein we sourced diverse datasets from hospitals across Kerala. Recognizing the significance of comprehensive and representative data in developing an effective predictive model, we collaborated with healthcare institutions to gather relevant information pertaining to PCOS risk factors and manifestations. The data collection process encompasseda wide range of variables, including demographic details, medical history, hormonal profiles, menstrual regularity, and clinical diagnoses. By accessing electronic health records and patient databases, we obtained anonymized datasets containing information from a diverse patient population. This approach ensured the inclusion of individuals from various age groups, ethnicities, and socioeconomic backgrounds, enhancing thegeneralizability and applicability of our predictive model.

B. Data Preprocessing

The data preprocessing played a crucial role in preparingthe collected datasets for analysis and model development.
This phase involved a series of systematic steps aimedat ensuring data consistency, completeness, and relevance.Initially, the collected data underwent thorough cleaning to address missing values, outliers, and inconsistencies. Techniques such as imputation, removal of duplicates, and outlier detection were employed to enhance data quality and integrity. Subsequently, normalization and standardization techniques were applied to scale numerical features and mitigate the impact of varying scales and units across different variables. Additionally, categorical variables wereencoded using appropriate encoding schemes to convert them into numerical representations suitable for analysis by machine learning algorithms.

C. Feature Engineering

Feature engineering stands as a cornerstone in machine learning endeavors, pivotal in transforming raw data into informative features that enhance model performance and interpretability. Its essence lies in extracting, selecting, and transforming input variables to improve predictive accuracy and facilitate model understanding. Through techniques such as dimensionality reduction, normalization, and creation ofnew features derived from existing ones, feature engineering empowers machine learning algorithms to capture intricatepatterns within data while mitigating noise and redundancy. This process plays a crucial role in optimizing model performance, especially in scenarios with high-dimensional or heterogeneous data. By carefully crafting features that encapsulate relevant information and domain knowledge, practitioners can unlock the full potential of machine learning models, ultimately enabling more accurate predictions and deeper insights into complex phenomena. Thus, in any project report involving machine learning, a comprehensive discussion on feature engineering is indispensable for understanding the methodology and rationale behind model development andperformance.

D. PCOS Exploratory Data Analysis (PEDA)

PCOS Exploratory Data Analysis (PEDA) is a critical step in understanding the characteristics and patterns present withinthe dataset related to Polycystic Ovary Syndrome (PCOS). This process involves thorough examination and visualization of the data to uncover insights and trends that can inform subsequent analysis and modeling. PEDA begins with an overview of the dataset, including its size, structure, and distribution of variables. Descriptive statistics such as mean, median, and standard deviation are calculated to summarize the central tendency and dispersion of numeric features, while frequency tables provide insight into categorical variables.

E. Comparitive Analysis Of Algorithms

In the comparative analysis of six classification algorithms for PCOS prediction, namely Random Forest Classifier, Logistic Regression, Linear SVM, Radial SVM, K Neighbors Classifier, and Gaussian Naive Bayes, the performancemetrics reveal varying degrees of accuracy. Random Forest Classifier emerges as the top-performing algorithm withan impressive accuracy rate of 90.91 percent, showcasing its robustness in handling complex datasets and capturing intricate patterns. Logistic Regression closely follows withan accuracy of 89.98 percent, demonstrating its effectiveness in linearly separable data scenarios. Linear SVM and Radial SVM share the third position, both achieving an accuracy rate of 88.11 perent, indicating their comparable performance in distinguishing between classes. K Neighbors Classifier follows closely with an accuracy of 87.64 percent, showcasing its utility in capturing local patterns in the data. Lastly, Gaussian Naive Bayes exhibits a slightly lower accuracy of 86.95 percent, suggesting its suitability for simpler datasets with strong class conditional independence assumptions. Overall, this comparative analysis highlights the strengths and weaknesses of each algorithm, guiding the selection of the most appropriate model based on the specific characteristics and requirements of the dataset at hand.

F. Data Visualization

The analysis concentrated on 20 features identified as significant by the CS-PCOS technique, which wereinstrumental in training the machine learning models. These features were further explored from various angles using diverse graphical representations. Visualizations of the charts were crafted with the assistance of Python libraries such as seaborn, pandas, and matplotlib. These libraries offered robust capabilities for generating insightful visualizations, enabling a deeper exploration and interpretation of the data. Through these visualization tools, intricate patterns, trends, and interrelationships among the selected features were unveiled, fostering a more nuanced comprehension of the dataset and guiding subsequent modeling strategies.

IV. USER INTERFACE

A query-based user interface for PCOS prediction simplifiesthe process of assessing the likelihood of Polycystic Ovary Syndrome (PCOS) development for users. Through a series of targeted questions related to known PCOS risk factors like menstrual irregularity, hirsutism, and weight issues, individuals can input their personal information. Behind the scenes, machine learning algorithms analyze these responses to generate personalized predictions regarding the probability of PCOS occurrence. This interface streamlines the prediction process, offering users a convenient and accessible way to understand their potential risk without the need for extensive medical knowledge or specialized testing. By leveraging the power of data-driven algorithms in a user-friendly format, query-based interfaces empower individuals to take proactive steps towards managing their health and seeking appropriate medical advice if necessary.

V. PCOS PREDICTION

Through meticulous data collection and preprocessing, the PCOS prediction system gathers comprehensive health information, including demographic data, medical history, hormonal profiles, and other relevant variables. Leveraging state-of-the-art ML algorithms such as Random Forest, Logistic Regression, and Support Vector Machine (SVM), the system constructs predictive models that analyze the intricate relationships between these variables and PCOS development. Additionally, bioinformatics techniques may be integrated to enhance predictive accuracy by incorporating genetic markers,gene expression data, and molecular features into the models.

VI. RESULTS

The results obtained from our predictive model demonstrated promising performance in accurately assessing an individual's likelihood of developing PCOS based on pertinent health data. Through rigorous evaluation and comparison of multiple machine learning algorithms,including Random Forest, Logistic Regression, Naive Bayes, Radial SVM, Linear SVM, and KNeighbours Classifier with an accuracy rate of 90.91 , 89.98 , 86.95 , 88.11 ,
88.11 , 87.64 respectively, we identified the most effective approaches for PCOS prediction. The evaluation metrics, including accuracy, precision, recall, and F1-score, showcasedthe robustness and reliability of the predictive model in discriminating between individuals at high and low riskof PCOS. Additionally, the incorporation of bioinformatics techniques enriched the predictive capabilities of the model, providing deeper insights into the molecular mechanisms underlying PCOS pathogenesis and potential biomarkers for improved prediction accuracy. Overall, the results of our project underscore the potential of data-driven approaches to transform PCOS diagnosis and management, offering personalized risk assessments and empowering individuals and healthcare providers with valuable insights for proactive intervention and healthcare decision-making.

VII. FUTURE SCOPE

1) The future scope for PCOS prediction using query-based applications holds considerable promise, with opportu- nities for further refinement and expansion. One avenue for advancement lies in enhancing the accuracy and gran-ularity of predictive models by incorporating additional relevant variables and refining the algorithms' predictive capabilities.

2) Integrating emerging biomarkers or genetic markers as- sociated with PCOS could provide deeper insights into individual risk profiles, thereby improving the precision of predictions. Furthermore, the development of mobileor web-based applications tailored to specific populations or demographic groups could enhance accessibility and uptake, reaching a broader audience and facilitating earlyintervention and management.

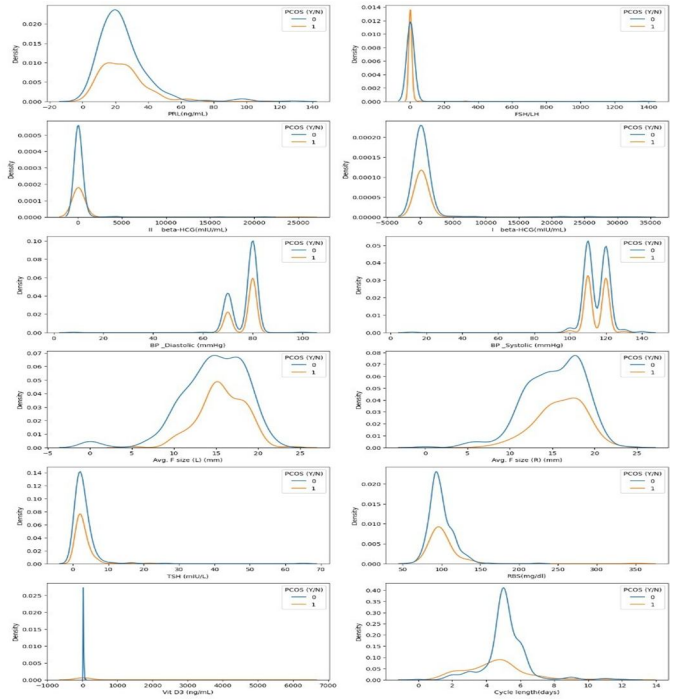


Fig. 2. Comparing Possibilities for PCOS and Selected Features

3) The integration of machine learning techniques such as reinforcement learning or deep learning could enable more dynamic and adaptive prediction models, capable of continuously learning and improving over time. Collabo- rations between data scientists, healthcare professionals, and technology developers will be instrumental in driving innovation and realizing the full potential of query-based applications for PCOS prediction, ultimately empowering individuals to make informed decisions about their health and well-being.

4) The development of predictive analytics tools capable of incorporating longitudinal data from electronic health records (EHRs) and patient-reported outcomes could enable more robust and personalized predictions. By harnessing rich clinical data and patient histories, these tools could provide insights into disease progression and treatment response, guiding personalized interventions and improving patient outcomes.

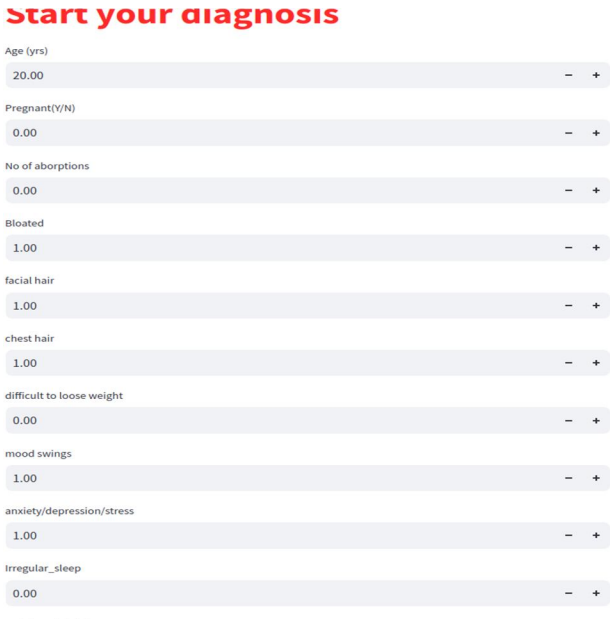


Fig. 3. User Interface for PCOS Prediction

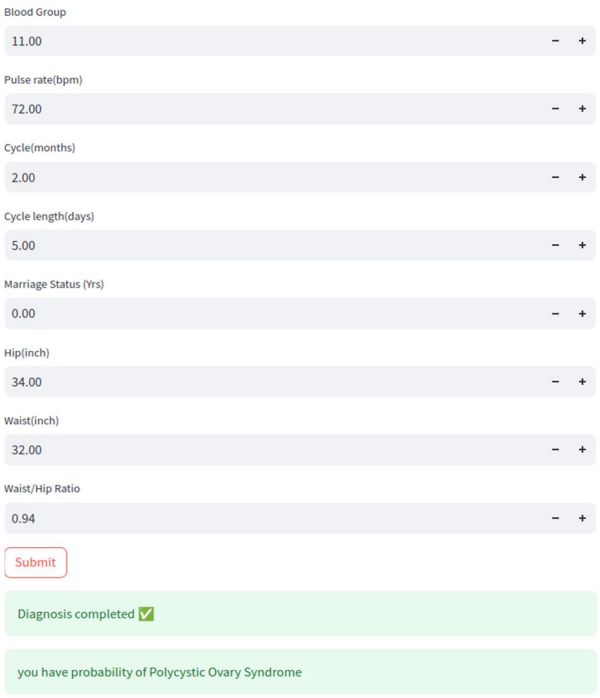


Fig. 4. User Interface for PCOS Prediction - Continuation

5) There is potential for the integration of telemedicine ca- pabilities within query-based applications, allowing users to connect with healthcare providers remotely for further evaluation and guidance based on their predictive results. This integration could bridge gaps in access to special- ized care, particularly in underserved communities, and facilitate timely intervention and management of PCOS- related symptoms.

In summary, ongoing research into novel biomarkers, genetic markers, and imaging techniques associated with PCOS could inform the development of more sophisticated prediction mod- els with enhanced sensitivity and specificity. Collaborative efforts across multidisciplinary teams, including researchers, clinicians, data scientists, and technology developers, will be essential to realize the full potential of query-based applica- tions for PCOS prediction and empower individuals to take proactive control of their health.

VIII. CONCLUSION

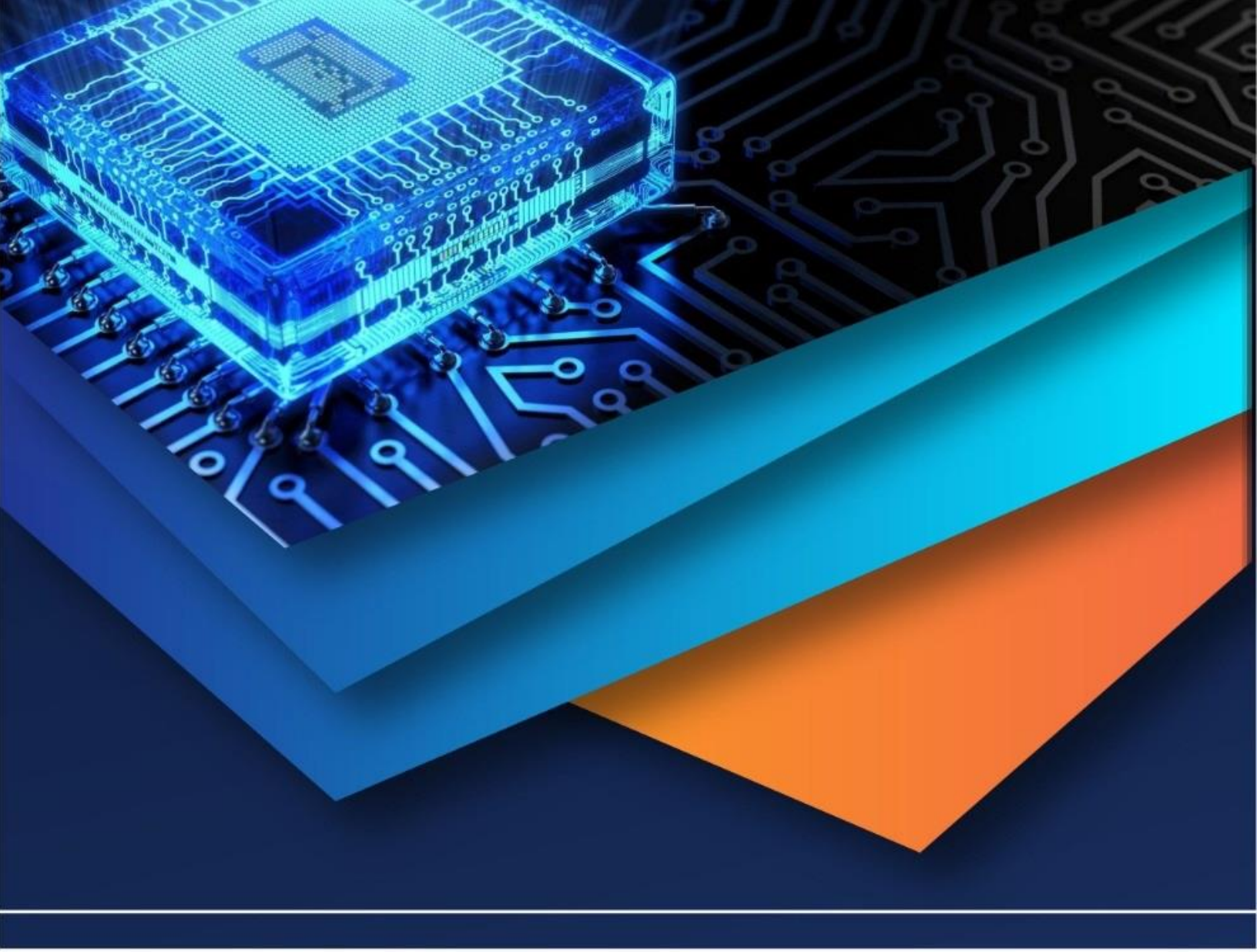
In conclusion, the development of the PCOS prediction system utilizing machine learning algorithms and a user- friendly interface marks a significant milestone in women's healthcare. Through meticulous data collection, preprocessing, and model development, this project has successfully leveraged advanced techniques to provide early detection and personalized risk assessment for Polycystic Ovary Syndrome (PCOS). The integration of state-of-the-art ML algorithms, including Random Forest, Logistic Regression, and Support Vector Machine (SVM), alongside bioinformatics techniques, has enabled the construction of robust predictive models capable of analyzing complex relationships within health data. Furthermore, the implementation of a user-friendly interface empowers individuals to easily input their health data and receive personalized predictions, fostering proactive health management and informed decision-making. By offering accessible and actionable insights into PCOS development, this project has the potential to revolutionize the approach to PCOS diagnosis and management, ultimately improving outcomes and quality of life for affected individuals. As the project continues to evolve and undergoes further validation and refinement, it is poised to make a significant impact on women's health and set a precedent for future advancements in predictive healthcare technology.

REFERENCES

[1] P. Chauhan, P. Patil, N. Rane, P. Raundale and H. Kanakia, "Com- parative Analysis of Machine Learning Algorithms for Prediction of PCOS," 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 2021, pp. 1-7, doi: 10.1109/ICCICT50803.2021.9510128.

[2] Elmannai H, El-Rashidy N, Mashal I, Alohali MA, Farag S, El- Sappagh S, Saleh H. Polycystic Ovary Syndrome Detection Ma- chine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence. Diagnostics (Basel). 2023 Apr 21;13(8):1506. doi: 10.3390/diagnostics13081506. PMID: 37189606; PMCID: PMC10137609.

[3] Suha, S.A., Islam, M.N. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. Sci Rep 12, 17123 (2022). https://doi.org/10.1038/s41598-022-21724-0.
[4] Guixue, G., Yifu, P., Yuan, G. et al. Progress of the application clinical prediction model in polycystic ovary syndrome. J Ovarian Res 16, 230 (2023). https://doi.org/10.1186/s13048-023-01310-2
[5] Shi N, Ma HB. Global trends in polycystic ovary syndrome research: A 10-year bibliometric analysis. Front Endocrinol (Lausanne). 2023 Jan 9;13:1027945. doi: 10.3389/fendo.2022.1027945. PMID: 36699019;PMCID: PMC9868474.
[6] Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records Zahra Zad, Victoria S. Jiang, Amber T. Wolf, Taiyao Wang, J. Jojo Cheng, Ioannis Ch. Paschalidis,
[7] Shruthi Mahalingaiah
[8] YAlam Suha S, Islam MN. Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique. Heliyon. 2023 Mar 16;9(3):e14518. doi: 10.1016/j.heliyon.2023.e14518. PMID: 36994397; PMCID: PMC10040521.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY