ISSN (Online) 2394-2320

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

Volume 11 Issue 6 June 2024

Early Detection of Polycystic Ovary Syndrome (PCOS)

[1] Aswin S, [2] Subhajit Paulb, [3] Anitha K

[1] [2] [3] SRM Institute of Science and Technology, SRM nagar, Kattankulathur, Chengalpet – 603203

Abstract— ***PCOS, or polycystic ovarian syndrome, is categorized as a serious health issue that affects women worldwide. Early PCOS diagnosis and treatment lowers the risk of long period consequences, such as elevated risk of diabetes that of type 2 and gestational diabetes. Consequently, the healthcare systems will benefit from a reduction in the issues and consequences associated with PCOS via efficient and early detection. Recently, promising achievements in medical diagnostics have been demonstrated via Machine Learning (ML) and ensemble learning. Our research's primary objective is to offer local and global model explanations that will guarantee the established model's efficacy, efficiency, and reliability. Various machine learning models may be used as feature selection techniques to obtain the best model and optimum feature selection. To increase performance, stacking ML models—combining meta-learner with the best base ML models—is suggested. ML models are optimized by Bayesian optimization. An 80:20 and 70:30 ratio-splitting benchmark PCOS dataset was used to get the experimental results. In comparison to other models, the outcome shown that utilizing ML with REF feature selection achieved the greatest 100 percent accuracy.***

I. INTRODUCTION

PCOS affects women who are pregnant as well as those who are already moms. Women's health is impacted by PCOS since it can lead to hormone imbalances and issues with metabolism. 5 to 10% of females in their reproductive years (15–45) suffer from this condition, which predominantly impacts women's fertility. It is a hormone imbalance that results in issues with the ovaries. For optimal health, the ovaries normally generate estrogen, a female hormone, and androgens, a male hormone. Hormones are substances that regulate bodily activities. The hormones of afflicted women are out of balance, with either less estrogen or more androgens than usual. As a result, fluid-filled sacs known as lumps develop on the ovaries. These lumps grow over time and eventually impede ovulation. For women with PCOS, this interferes with ovulation, which lowers their chances of getting pregnant. Endometrial cancer, diabetes, anxiety, heart disease, high blood pressure, depression, sleep apnea, and thickness of the endometrium are among the conditions that women with PCOS are more prone to experience. The development of PCOS may also be influenced by environmental variables in addition to hereditary ones. Long-term consequences can be decreased including early diagnosis, therapy, and weight management.

Particularly for PCOS, artificial intelligence (AI) has completely changed how disorders are identified and treated. Deep learning networks (DL) and machine learning (ML) algorithms are two examples of AI-based technologies that have made it possible to construct systems automation for the process of precise and dependable diagnosis of cardiac disease. AI-based techniques may separate PCOS patients from non-PCOS patients by finding trends in health data, including hormone levels. Better overall results for PCOS patients and earlier, more accurate diagnosis might result from this increased accuracy. Additionally, Artificial intellectual systems may be used to continuously monitor patients, giving medical professionals insightful information about possible therapies, and facilitating more accurate interventions. To put it briefly, AI-driven technology has the potential to completely transform the identification and management of PCOS, offering patients with the illness more effective and efficient care.

II. REVIEW OF LITERATURE

Since PCOS symptoms might differ from patient to patient, it's critical to have a diagnosis as soon as possible since PCOS can worsen existing health issues including diabetes, heart disease, endometrial cancer, and other disorders. Ultrasound imaging, blood testing, and physical examinations are used in the diagnosis of PCOS. To find follicles or immature eggs inside an ovary, ultrasound is utilized. (Jan et al,2022).

The most prevalent endocrine condition affecting a large number of women in their childbearing years is polycystic ovarian syndrome, or PCOS. In women, it causes a variety of illnesses and sterility. The primary indicators of this condition are irregular menstruation cycles, obesity, acne, greasy skin, and anxiety issues. PCOS is thought to affect one in five women. It is observed that the majority of women ignore the typical signs of PCOS and only go to the doctor when they are having trouble getting pregnant. (Chauhan et al,2021)

Hormonal imbalance causes a monthly cycle to be delayed or even absent. For PCOS, there is no known etiology. It might be challenging to conceive if a woman has PCOS, which is characterized by significant acne, hair loss, weight gain, skin discoloration, facial hair development, and irregular periods. With weight control, early identification, and treatment, reducing the likelihood of developing long-term conditions like type 2 diabetes and heart problems is possible. (Srinithi et al., 2023)

Diagnosing PCOS can be challenging since not everyone with ovarian cysts or Polycystic Ovaries (PCO) has PCOS; therefore, a pelvic ultrasound alone is insufficient to diagnose PCOS. The comprehensive diagnostic approach primarily consists of a pelvic ultrasonography in addition to blood testing for certain markers of PCOS existence. (Hdaib et al., 2022)

The endocrine and lifestyle condition Polycystic Ovary Syndrome affects women who are fertile. It is an illness that results in the ovaries to swell as a result of the cysts. Numerous methods, including blood testing, pelvic exams, and ultrasound imaging, can be used to identify PCOS. These clinical techniques are expensive, time-consuming, and prone to human error in their outcomes. (Makdoomi et al, 2022)

For the purpose of crucial PCOS prediction, the suggested GNB model produced correct findings. According to our analysis, the dataset's key components with a strong participation in PCOS prediction include, relative risk (RR-breaths), thyroid stimulating hormone (TSH), and blood pressure, both diastolic and systolic. (Nasim et al, 2022)

PCOS, also known as polycystic ovarian syndrome, is a complicated hormonal illness that can cause a variety of symptoms, including irregular menstruation periods, obesity, acne, hirsutism, alopecia, and increased androgen production in women. Since these symptoms might differ from patient to patient, it's critical to acquire a diagnosis as soon as possible since untreated endometrial cancer can lead to major health concerns including diabetes, heart disease, and other disorders. Ultrasound imaging, blood testing, and physical examinations are used in the diagnosis of PCOS. To find follicles or immature eggs inside an ovary, ultrasound is utilized. (Naila Jan et al, 2022)

The most prevalent endocrine condition affecting a large number of women in their childbearing years is polycystic ovarian syndrome, or PCOS. In women, it causes a variety of illnesses and sterility. The primary indicators of this condition are irregular menstruation cycles, obesity, acne, greasy skin, and anxiety issues. PCOS is thought to affect one in five women. It is observed that the majority of women ignore the typical signs of PCOS and only go to the doctor when they are having trouble getting pregnant. (Preeti Chauhan et al, 2021)

Women who are affected by the polycystic ovarian syndrome (PCOS) condition are fertile and has an impact on hormones. Hormonal imbalance causes a monthly cycle to be delayed or even absent. For PCOS, there is no known etiology. Among women, the most typical signs of PCOS are increased weight gain and facial hair development, acne, baldness, pigmentation, and unpredictable menstrual periods, all of which can make conception challenging. (Srinithi et al, 2023)

Polycystic Ovary Syndrome is the most prevalent illnesses affecting female gender who are fertile (PCOS). Diagnosing PCOS can be challenging since not everyone with ovarian cysts or polycystic ovaries (PCO) has PCOS; therefore, a pelvic ultrasound alone is insufficient to diagnose PCOS. The comprehensive diagnostic approach primarily consists of a pelvic ultrasonography in addition to blood testing for certain markers of PCOS existence. (Dana Hdaib et al, 2022)

III. SYSTEM ANALYSIS

A. Existing Systems:

Many academic and clinical research studies have explored the placement of Machine Learning for PCOS prediction. These studies often involve collecting and analyzing medical data, including hormonal levels, ultrasound results, and patient histories, to develop predictive models.

Some researchers and healthcare providers have developed mobile apps or online tools that use machine learning to help individuals assess their risk of PCOS based on self-reported information such as symptoms, age, and menstrual history. These tools are typically designed for educational purposes and should not replace medical diagnosis.

In clinical settings, machine learning models have been integrated into electronic health record systems to assist healthcare professionals in identifying PCOS patients. These models can analyze a patient's historical health data to flag potential cases.

Machine learning researchers often collaborate with healthcare institutions and hospitals to access large datasets of patient records. These collaborations aim to develop more accurate prediction models by leveraging extensive real-world data.

Disadvantages Of Existing Systems:

Many machine learning models for PCOS prediction rely on the availability of large and high-quality datasets. In some cases, the data may be limited or biased, leading to less accurate predictions.

Healthcare data, including information related to PCOS, is sensitive and must be handled with care. Existing systems may face challenges in ensuring the private data of patients, which can be a significant concern.

Machine learning module based on historical healthcare data may inherit biases present in that data. This can result in unfair predictions or underrepresentation of certain demographic groups, potentially leading to disparities in healthcare outcomes.

B. Proposed System:

Collaborate with multiple healthcare institutions to collect a diverse and representative dataset of PCOS-related features.

Implement robust data preprocessing techniques to handle missing values, outliers, and noise.

Address bias in the data through fairness-aware data preprocessing methods.

Utilize state-of-the-art machine learning algorithms, incorporate explainable AI techniques to enhance model interpretability.

Explore feature engineering techniques to create informative features based on domain knowledge.

Employ feature selection methods to identify the most relevant factors for PCOS prediction.

IV. SYSTEM DESIGN

A. System Architecture:

The multi-step design process concentrates on the architecture of data structures in software, procedural specifics, etc., and the interface between modules. Before coding starts, the criteria are also decoded throughout the design phase into a software presentation that may be reviewed for accuracy. The design of computer software is always evolving due to the development of new techniques, enhanced analysis, and improved knowledge. The revolution of software proposal is currently in a relatively early stage.

As a result, the breadth, flexibility, and quantitative nature typically associated with more traditional engineering fields are not included in the software design methodology. Nonetheless, software design techniques do exist, as do standards for design attributes, and design notation can be used.

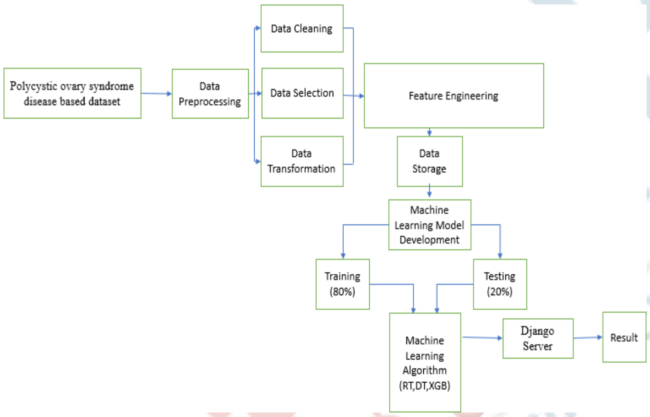


Fig. 1. Architecture Diagram

B. Flow Diagram:

A diagram that shows a flow or a collection of dynamic interactions inside a system is collectively referred to as a flow diagram. In addition to being a synonym for flowcharts, the phrase "flow diagram" is also occasionally used to refer to a flowchart's opposite.

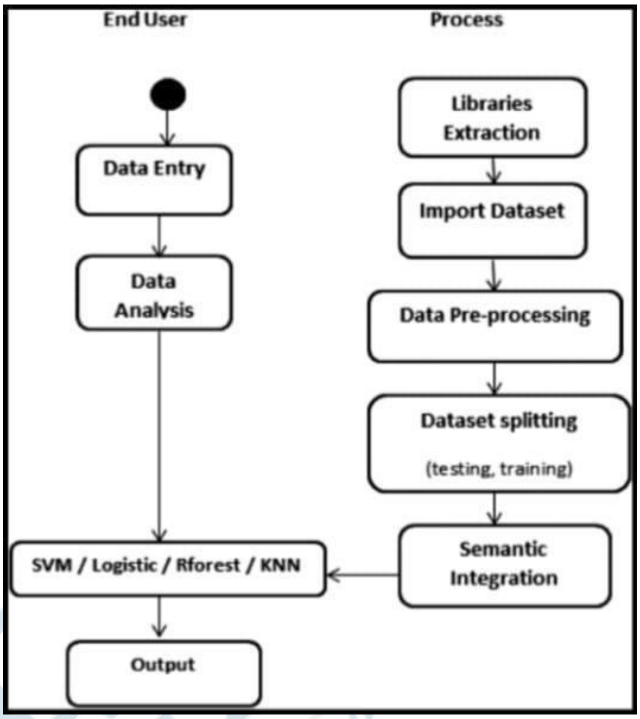


Fig. 2. Flow Diagram

C. Entity Relationship Diagram:

An Entity-Relationship (ER) Diagram is a visual representation used for DB design to model the entities (objects, concepts, or things) and their relationships within a database. ER diagrams use various symbols to depict entities, attributes, relationships, and cardinalities.

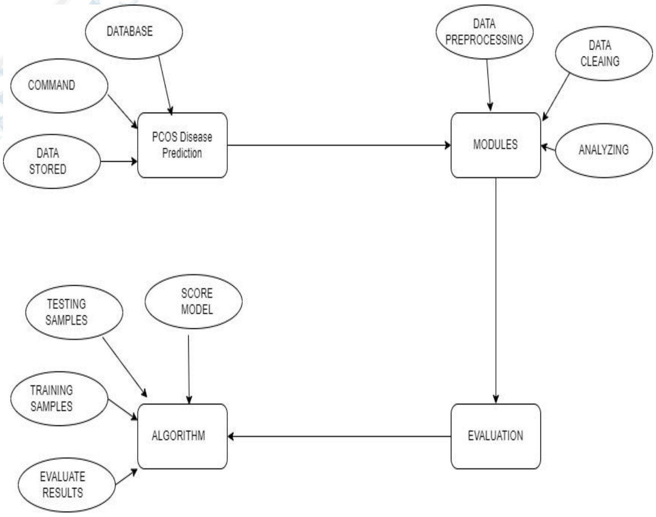


Fig. 3. ER Diagram

D. Class Diagram:

A Class Diagram is a type of UML (Unified Modeling Language) that shows the structure and relation between the classes or objects. In the context of software development, it is often used to model the classes, attributes, methods, and associations within a software application.

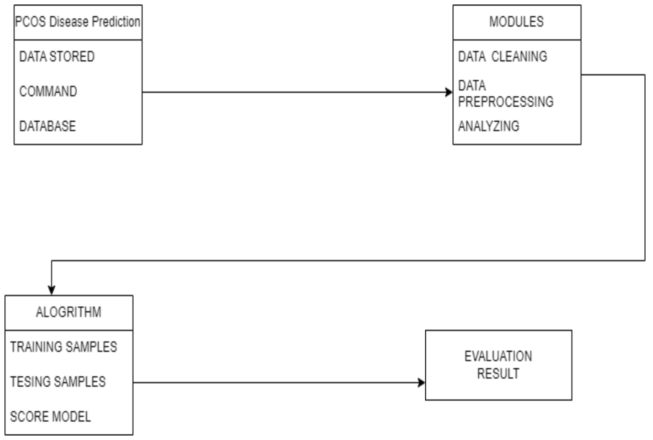


Fig. 4. Class Diagram

E. Use Case Diagram:

A Use Case Diagram in the context of Machine Learning can help illustrate how different actors interact with a machine learning system and what functionalities or services it provides.

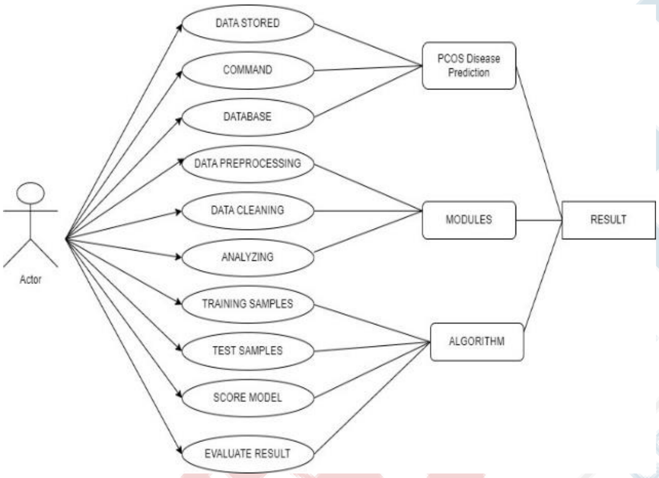


Fig. 5. Use Case Diagram

V. SYSTEM IMPLEMENTATION

Following the profile selection, the appropriate properties, or features, are chosen, and the classification method is then used. As new training data is given into the classifier on a regular basis, the classifier is continually educated.

A. Collecting Data:

One of the most important steps in developing a ML model is data collection.

Gather comprehensive datasets containing information on patients with and without PCOS. This data may include demographic details, medical history, lifestyle factors, hormonal levels, ultrasound results, and other relevant information.

It involves obtaining task-related data based on certain targeted factors in order to evaluate and generate some useful results.

B. Features:

After the process of comprehensive analysis of data, the targeted features to analyze the data are to be listed to pre-process it to train and test the ML module. These features play a vital role in the detection or calculations performed by the module. In this project of detection of PCOS, there were twelve features detected from the data collection.

Age: Age plays a crucial role in the detection and management of Polycystic Ovary Syndrome (PCOS), with symptoms often emerging in the late teens or early twenties but diagnosis sometimes occurring later due to delayed recognition. Clinical presentation varies with younger women more commonly experiencing menstrual irregularities, acne, and hirsutism, while older women may present with infertility or metabolic complications like insulin resistance and obesity. Fertility challenges escalate with age due to declining ovarian reserve, exacerbating the reproductive difficulties associated with PCOS. Metabolic complications such as insulin resistance, obesity, and cardiovascular risks may worsen with age, amplifying long-term health risks like endometrial cancer and cardiovascular disease. Treatment approaches must consider age-related factors, with younger women focusing on symptom management and menstrual regularity, while older women may prioritize fertility treatments or interventions to mitigate metabolic risks, emphasizing the importance of tailored management strategies across different life stages.

Weight: The management and detection of Polycystic Ovary Syndrome (PCOS), a condition characterized by hormonal imbalances, are closely linked to weight. While not a sole diagnostic criterion, excess weight or obesity often accompanies PCOS and can exacerbate its symptoms through mechanisms such as insulin resistance and hormonal imbalances. The presence of obesity may also influence the diagnostic criteria, such as the Rotterdam criteria, which consider features like irregular menstrual cycles, hyperandrogenism, and ovarian cysts. Furthermore, adipose tissue produces hormones, and excess fat can disrupt normal hormone levels, contributing to PCOS symptoms. Managing weight through lifestyle changes, including diet and exercise, is a cornerstone of PCOS management, as even modest weight loss can lead to significant improvements in insulin sensitivity, hormonal balance, and overall symptom control. Additionally, managing weight is crucial for minimizing the likelihood of long-term complications like type 2 diabetes and cardiovascular disease in women diagnosed with PCOS.

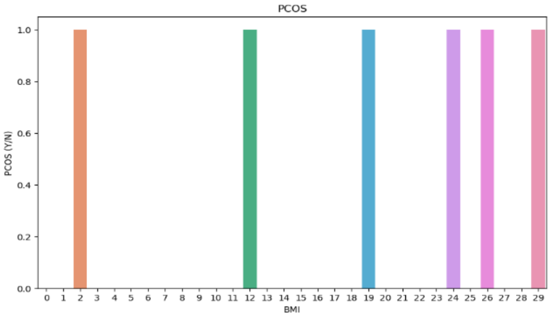


Fig. 6. Occurrence of PCOS in different BMI

Haemoglobin: Hemoglobin levels, while not directly diagnostic of polycystic ovary syndrome (PCOS), can indirectly reveal certain aspects of the condition. Women with PCOS often experience irregular menstrual cycles, potentially leading to heavy menstrual bleeding and subsequent anemia. Additionally, PCOS is associated with insulin resistance, which can contribute to metabolic abnormalities like diabetes, a condition sometimes accompanied by anemia. Chronic inflammation, prevalent in PCOS, may also affect erythropoiesis, contributing to anemia of chronic disease. Furthermore, nutritional deficiencies common in PCOS could influence hemoglobin levels. Thus, while hemoglobin levels alone do not diagnose PCOS, they can provide insights into associated factors such as menstrual irregularities, insulin resistance, inflammation, or nutritional status.

Pregnancy Status: Pregnancy can influence the detection of polycystic ovary syndrome (PCOS) through various avenues. Women with PCOS often experience difficulties conceiving due to irregular ovulation, prompting them to seek medical assistance for infertility, which can lead to a PCOS diagnosis. Hormonal changes during pregnancy may temporarily alleviate or exacerbate PCOS symptoms, such as irregular periods or insulin resistance, potentially prompting further investigation. Additionally, the heightened risk of gestational diabetes in female with PCOS can necessitate closer monitoring during pregnancy, potentially revealing underlying insulin resistance characteristic of PCOS. Postpartum, symptoms of PCOS may become more apparent, leading to diagnostic evaluation. Overall, while pregnancy itself may not directly diagnose PCOS, it can impact symptom manifestation and prompt medical evaluation, contributing to its detection.

Respiratory Rate: Respiratory rate isn't a direct indicator for diagnosing polycystic ovary syndrome (PCOS), a hormonal disorder affecting women. PCOS diagnosis typically involves a combination of clinical symptoms, hormone level assessments, and imaging studies like ultrasound. However, conditions associated with PCOS, such as obesity and sleep apnea, can impact respiratory function. Obesity, prevalent in PCOS, may lead to respiratory issues like hypoventilation syndrome and obstructive sleep apnea, affecting respiratory rate and potentially manifesting symptoms like daytime sleepiness and fatigue. Thus, while respiratory rate itself isn't diagnostic for PCOS, monitoring respiratory health, especially in the context of associated conditions, is important for comprehensive patient care.

Random Blood Sugar: In the context of diagnosing Polycystic Ovary Syndrome (PCOS), Random Blood Sugar (RBS) testing serves as a valuable component in assessing metabolic health and potential risk factors associated with the condition. PCOS, a complex hormonal disorder affecting women, often manifests alongside insulin resistance and metabolic irregularities, leading to fluctuations in blood sugar levels. While RBS alone doesn't diagnose PCOS, it contributes to a comprehensive health evaluation. Elevated RBS levels may signal insulin resistance, a common feature of PCOS, prompting further investigation and management considerations. Moreover, RBS testing aids in ruling out other glucose metabolism disorders and assessing overall metabolic health, essential in guiding treatment decisions and monitoring for potential complications such as diabetes, which individuals with PCOS are at increased risk for. Therefore, while not a standalone diagnostic tool, RBS testing plays a crucial role in the holistic assessment and management of PCOS.

Skin Darkening: Skin changes, including darkening, are sometimes observed in PCOS patients, such as acanthosis nigricans, a condition marked by dark, thickened patches of skin often found in body folds, indicative of insulin resistance commonly associated with PCOS. Additionally, skin tags and hirsutism, excessive hair growth in typically male-pattern areas, can also occur due to hormonal imbalances. These manifestations, along with symptoms like androgenic alopecia (thinning of hair), contribute to the diagnostic profile of PCOS, although diagnosis typically requires a combination of symptoms assessment, physical examination, hormone level tests, and ovarian imaging. While skin darkening alone isn't diagnostic of PCOS, it can serve as a signpost for healthcare providers, prompting further evaluation for this complex condition.

Hair Loss: Within the constellation of symptoms indicative of PCOS, hair loss, particularly androgenic alopecia, plays a significant diagnostic role. This type of hair loss stems from elevated levels of androgens, such as testosterone, commonly observed in individuals with PCOS. Hair loss in PCOS often mirrors the pattern seen in male-pattern baldness, with thinning occurring at the crown of the head or along the hairline. While hair loss alone is insufficient for a diagnosis, it forms a crucial part of the clinical evaluation when combined with other symptoms, such as hirsutism, acne, and menstrual irregularities. Healthcare providers utilize a comprehensive approach, incorporating physical examinations, blood tests to assess hormone levels, and sometimes imaging studies like ultrasound to examine the ovaries for cysts. Treatment strategies for hair loss in PCOS typically involve addressing the underlying hormonal imbalances through oral contraceptives, anti-androgen medications, and lifestyle modifications. Thus, the presence of hair loss contributes significantly to the medical detection and management of PCOS, aiding in timely intervention and symptom alleviation.

Endometrium: The endometrium, the lining of the uterus, plays a crucial role in the detection of polycystic ovary syndrome (PCOS) due to its responsiveness to hormonal fluctuations. PCOS often leads to irregular menstrual cycles, causing abnormalities in the endometrium's thickness and structure, which can be observed through imaging techniques like transvaginal ultrasound or endometrial biopsy. Women with PCOS are at an increased risk of developing endometrial hyperplasia, characterized by abnormal thickening of the endometrium, primarily due to prolonged exposure to estrogen without the counterbalancing effects of progesterone. Additionally, the heightened risk of endometrial cancer in women with PCOS underscores the importance of monitoring the endometrium for signs of abnormal thickening or other changes, facilitating early detection and management of potential complications.

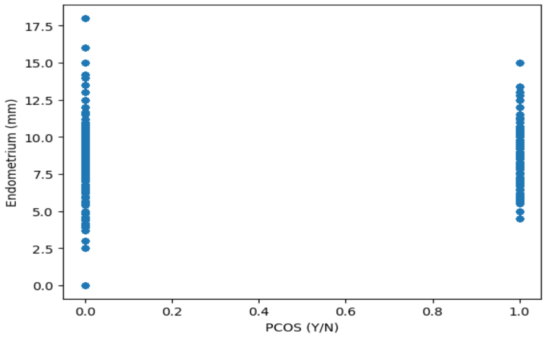


Fig. 7. occurrence of PCOS in respect to Endometrium

C. Data Pre-Processing:

Clean and preprocess the data to handle missing values, outliers, and inconsistencies. Normalizing or scaling features may also be necessary to ensure that the machine learning models can effectively learn from the data.

Pre-processing of data might involve data transformation, data cleansing, and data selection.

Data cleaning: Correct missing numbers, reduce noise in the data, find and eliminate outliers, and address discrepancies.

Smoothing, aggregation, generalization, and transformation are examples of data transformation techniques that enhance data quality.

Data selection comprises some techniques or features that enable us to choose pertinent data for our system.

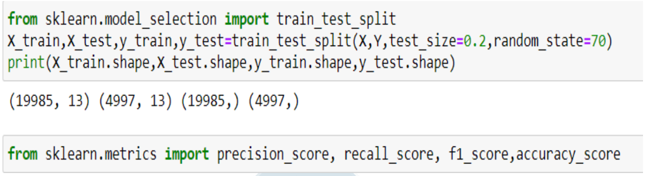
D. Training and Testing:

X` represents your feature vectors, which can be a NumPy array or a panda DataFrame.

'y` signifies the respective labels or target values.

test_size=0.2 ` specifies that you want to allocate 20% of your data for testing, and 80% is for training.

After running this code, you'll have `X_train`, `y_train`, `X_test`, `y_test` containing your model data. You can then proceed to evaluvate the ML module using these datasets.



```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.2,random_state=70)
print(X_train.shape,X_test.shape,y_train.shape,y_test.shape)
```

```
(19985, 13) (4997, 13) (19985,) (4997,)
```

```
from sklearn.metrics import precision_score, recall_score, f1_score,accuracy_score
```

Fig. 8. creating train and test model.

E. Algorithm:

Values from the dataset were transformed into arrays, which the program would use to determine correctness. Choose an algorithm based on its accuracy, then use it to analyze the data.

Random Forest: One popular machine learning algorithm is Random Forest, which is used in supervised learning techniques. Regression and classification are two examples of machine learning issues that it may be used to. Ensemble learning, the central concept, involves amalgamating numerous classifiers to bolster the model's capabilities and tackle complex problems effectively.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

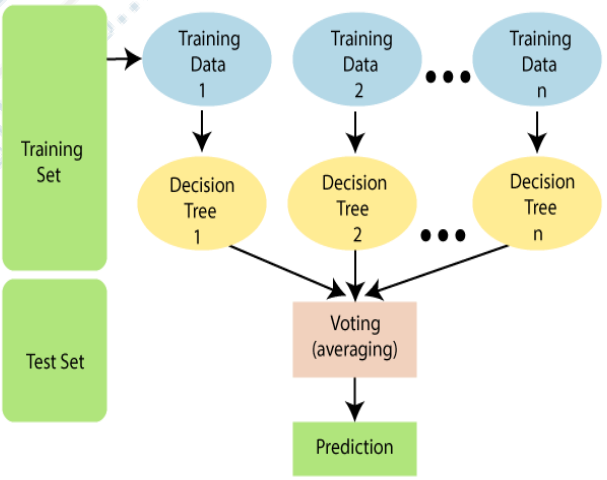


Fig. 9. Random Forest

The random forest makes predictions by aggregating the high-voted guesses from multiple decision trees, rather than relying solely on one. Because there are more trees in the forest, accuracy is higher and overfitting is avoided.

Compared to alternative algorithms, it demands less training time, functions efficiently with extensive datasets, and yields highly accurate predictions. Moreover, it maintains accuracy even when confronted with substantial data gaps.

Decision Tree: Decision Trees are a versatile tool applied in diagnosing faults in electric vehicles, leveraging pre-existing data containing known errors to train models. While primarily used for categorization problems, Decision Trees can also tackle regression issues effectively. Structured like a tree, with core nodes representing dataset properties, branches denoting decision rules, and leaf nodes signifying outcomes, this classifier navigates through potential solutions or decisions given certain constraints.

Beginning at the root node and branching out akin to a tree's growth, Decision Trees offer a graphical representation of potential problem-solving paths. The Classification and Regression Tree (CART) algorithm is instrumental in constructing these trees, posing queries and partitioning the tree into subtrees based on responses (Yes/No).

When predicting the class of a given dataset, the Decision Tree algorithm initiates at the root node, comparing the root property's values with the dataset attributes. It then progresses through branches, moving to the next node based on attribute comparisons, until reaching a leaf node. This iterative process continues until the entire dataset is classified. Additionally, the XGBoost algorithm can enhance our comprehension of this procedure, offering insights into the Decision Tree's functionality and effectiveness.

XGboost: XgBoost (Extreme Gradient Boosting) module of Python by the University of Washington was Introduced. An Python module written in C++, for the help in the training ML models for Gradient Boosting.

Gradient boosting is one AI method used in regression and classification applications. It offers an expectation model as a collection of imprecise forecasting models, also referred to as decision trees.

Gaussian Naïve Bayes: Gaussian Naive Bayes is suitable for classification tasks when a Gaussian distribution is found in the features that are continuous. In Python, you can implement Gaussian Naive Bayes using the GaussianNB module from the sklearn.naive_bayes library, which is part of the scikit-learn package.

Naive Bayes operates under the assumption that, given the class label, the features are conditionally independent. This means a single feature present in a class is independent of the all other features present.

In Gaussian Naive Bayes, it's assumed that the likelihood of the features given the class label follows a Gaussian distribution. This means that it assumes that the continuous-valued features are normally distributed within each class.

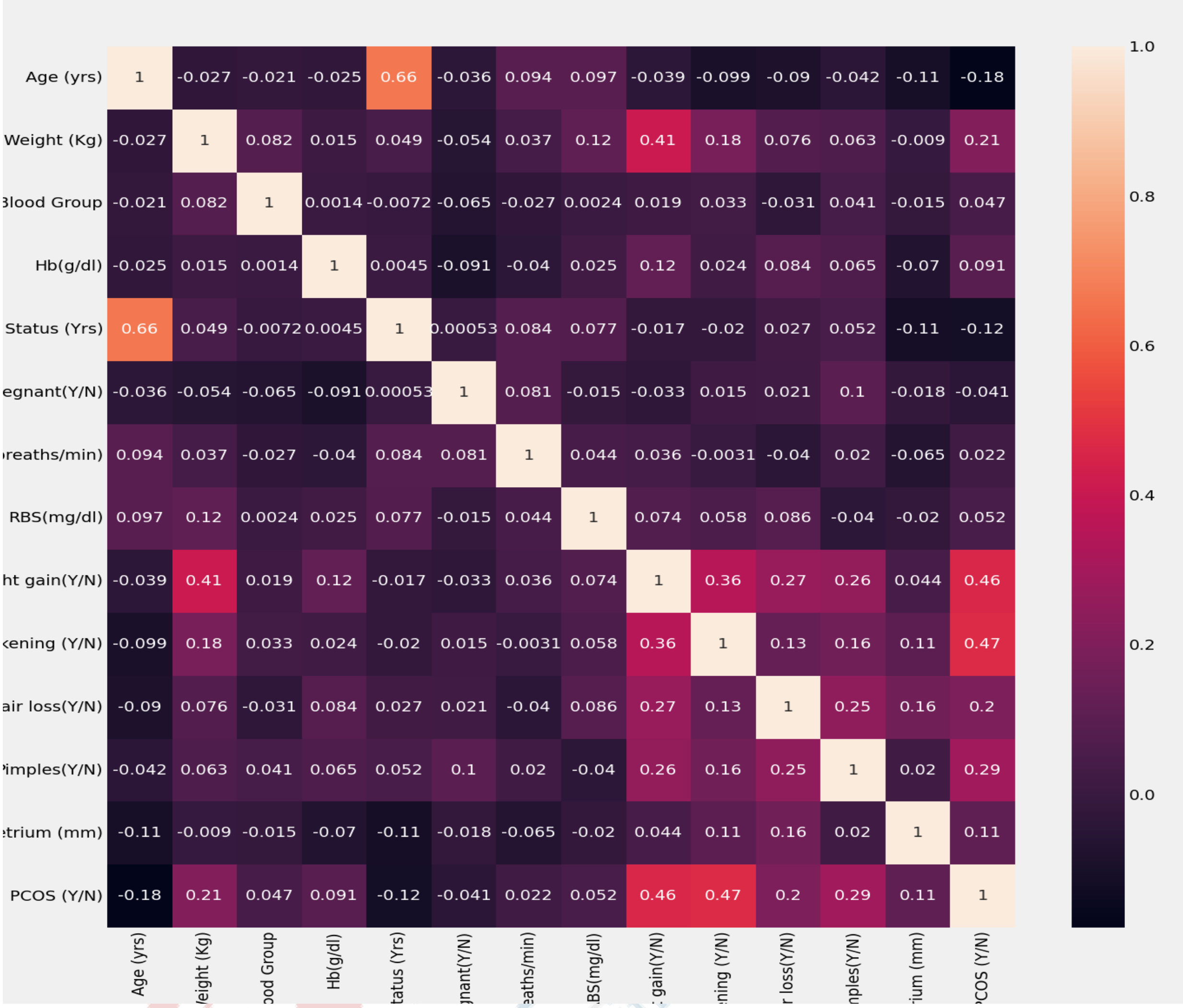
To apply Gaussian Naive Bayes, you need to estimate the parameters of the Gaussian distributions for each class (mean and variance for each feature). Once the parameters are estimated, to classify a new instance, Gaussian Naive Bayes calculates the likelihood of the features for each class using the Gaussian probability density function and then applies Bayes' theorem to find the most probable class.

Gaussian Naive Bayes is simple, efficient, and works well for classification tasks, Gaussian distribution is found in the features that are continuous. However, the assumption of feature independence might not hold in all cases, which can affect its performance, especially when dealing with highly correlated features.

Support Vector Classification: The Support Vector Classification (SVC) algorithm stands out as a potent tool tailored specifically for addressing classification tasks within machine learning. It operates by identifying the optimal hyperplane in the feature space that effectively separates different classes. Key to this approach are the support vectors, which are the data points closest to the hyperplane and crucial for determining its position and orientation. In binary classification scenarios, the decision boundary is represented by a hyperplane, which extends to a hyperplane in higher dimensions.

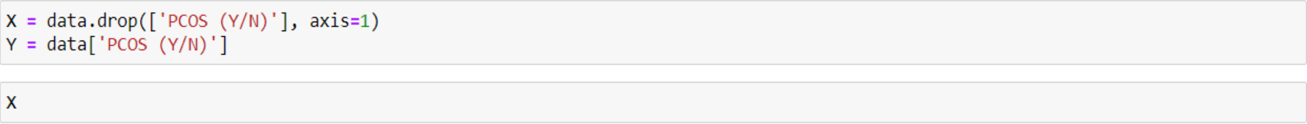
The kernel function in SVC plays a vital role by transforming input data into a higher-dimensional space, facilitating the identification of an optimal hyperplane. The margin, defined as the distance between the hyperplane and the support vectors, is maximized during training to enhance generalization and mitigate overfitting. Balancing the trade-off between classification error and margin size is achieved through the regularization parameter (C), where larger values of C prioritize minimizing misclassifications at the expense of a narrower margin.

Following training, SVC effectively predicts the class of new data points by assessing their position relative to the decision boundary. Even in cases where data is not linearly separable, SVC can still find a solution through soft margin classification, allowing for some misclassifications with associated penalties. Due to its versatility in handling both linearly and non-linearly separable data, SVC finds extensive applications across various domains, including bioinformatics, text classification, and image recognition.



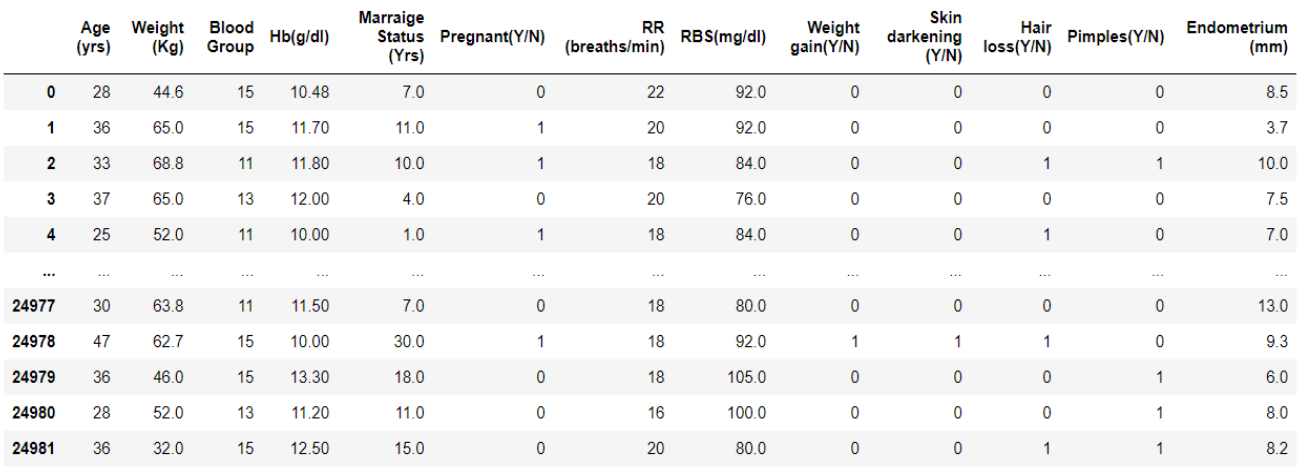
	Age (yrs)	eight (Kg)	ood Group	Hb(g/dl)	tatus (Yrs)	nant(Y/N)	aths/min)	BS(mg/dl)	gain(Y/N)	ning (Y/N)	r loss(Y/N)	nples(Y/N)	ium (mm)	PCOS (Y/N)
Age (yrs)	1	-0.027	-0.021	-0.025	0.66	-0.036	0.094	0.097	-0.039	-0.099	-0.09	-0.042	-0.11	-0.18
Weight (Kg)	-0.027	1	0.082	0.015	0.049	-0.054	0.037	0.12	0.41	0.18	0.076	0.063	-0.009	0.21
3lood Group	-0.021	0.082	1	0.0014	-0.0072	-0.065	-0.027	0.0024	0.019	0.033	-0.031	0.041	-0.015	0.047
Hb(g/dl)	-0.025	0.015	0.0014	1	0.0045	-0.091	-0.04	0.025	0.12	0.024	0.084	0.065	-0.07	0.091
Status (Yrs)	0.66	0.049	-0.0072	0.0045	1	0.00053	0.084	0.077	-0.017	-0.02	0.027	0.052	-0.11	-0.12
egnant(Y/N)	-0.036	-0.054	-0.065	-0.091	0.00053	1	0.081	-0.015	-0.033	0.015	0.021	0.1	-0.018	-0.041
reaths/min)	0.094	0.037	-0.027	-0.04	0.084	0.081	1	0.044	0.036	-0.0031	-0.04	0.02	-0.065	0.022
RBS(mg/dl)	0.097	0.12	0.0024	0.025	0.077	-0.015	0.044	1	0.074	0.058	0.086	-0.04	-0.02	0.052
ht gain(Y/N)	-0.039	0.41	0.019	0.12	-0.017	-0.033	0.036	0.074	1	0.36	0.27	0.26	0.044	0.46
kening (Y/N)	-0.099	0.18	0.033	0.024	-0.02	0.015	-0.0031	0.058	0.36	1	0.13	0.16	0.11	0.47
air loss(Y/N)	-0.09	0.076	-0.031	0.084	0.027	0.021	-0.04	0.086	0.27	0.13	1	0.25	0.16	0.2
Pimples(Y/N)	-0.042	0.063	0.041	0.065	0.052	0.1	0.02	-0.04	0.26	0.16	0.25	1	0.02	0.29
etrium (mm)	-0.11	-0.009	-0.015	-0.07	-0.11	-0.018	-0.065	-0.02	0.044	0.11	0.16	0.02	1	0.11
PCOS (Y/N)	-0.18	0.21	0.047	0.091	-0.12	-0.041	0.022	0.052	0.46	0.47	0.2	0.29	0.11	1

Fig. 10. Heat Map



```
X = data.drop(['PCOS (Y/N)'], axis=1)
Y = data['PCOS (Y/N)']
```

```
X
```



	Age (yrs)	Weight (Kg)	Blood Group	Hb(g/dl)	Marraige Status (Yrs)	Pregnant(Y/N)	RR (breaths/min)	RBS(mg/dl)	Weight gain(Y/N)	Skin darkening (Y/N)	Hair loss(Y/N)	Pimples(Y/N)	Endometrium (mm)
0	28	44.6	15	10.48	7.0	0	22	92.0	0	0	0	0	8.5
1	36	65.0	15	11.70	11.0	1	20	92.0	0	0	0	0	3.7
2	33	68.8	11	11.80	10.0	1	18	84.0	0	0	1	1	10.0
3	37	65.0	13	12.00	4.0	0	20	76.0	0	0	0	0	7.5
4	25	52.0	11	10.00	1.0	1	18	84.0	0	0	1	0	7.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
24977	30	63.8	11	11.50	7.0	0	18	80.0	0	0	0	0	13.0
24978	47	62.7	15	10.00	30.0	1	18	92.0	1	1	1	0	9.3
24979	36	46.0	15	13.30	18.0	0	18	105.0	0	0	0	1	6.0
24980	28	52.0	13	11.20	11.0	0	16	100.0	0	0	0	1	8.0
24981	36	32.0	15	12.50	15.0	0	20	80.0	0	0	1	1	8.2

Fig. 11. data for X after removing PCOS target values

VI. TESTING

Finding defects is the aim of testing. The process of trying to identify every possible defect or weakness in a workable invention is called testing. It is a way to confirm that meetings, subassemblies, workings, and/or a finished product are operating as intended. Software training aims to guarantee that the software system satisfies its needs and user opportunities and does not malfunction in a way that is unacceptable. Different test kinds exist. A certain testing condition is addressed by each type of test.

A. Unit Testing:

Creating test objects for unit testing ensures that the fundamental program logic operates as intended and that valid inputs yield legitimate outputs. Authorization is required for internal code flow and all option branches. It is the request's individual software modules being tested. It is completed prior to integration, following the conclusion of a single unit. This is a structural test that intrusively relies on facts regarding its configuration. Unit tests do fundamental testing at the factor level and evaluate a particular business procedure, application, or system architecture. Unit tests guarantee that every single route of the business process comprises precisely specified inputs and likely outcomes, and that it completes according to the stated specifications.

B. Functional Testing:

Accessible functions are stated in system certification, user manuals, business and technical requirements, and functional testing. Methodical evidence for this is provided by functional testing. Most often, functional difficulties is concentrated in the following areas:

- Classes of acceptable input that need to be allowed are identified using valid input.
- Illegal input is used to identify kinds of input that ought to be prohibited.
- Exercise of functions employed for specified objectives is required.
- Request output modules are categorized using output.
- Procedures and Systems are used to interface when occurrences need to be appealed.
- Functional tests are arranged and grounded with an emphasis on materials, essential features, or unique test items. Furthermore, testing requires thoughtful coverage of data fields, specified procedures, and subsequent activities in order to discover business process flows.

C. Whitebox Testing:

Software testing with knowledge of the program's inner workings, architecture, and programming language or at least its motivation is known as "white box" testing. It is a challenging technique. It is used to assess areas from an unexpandable black box level.

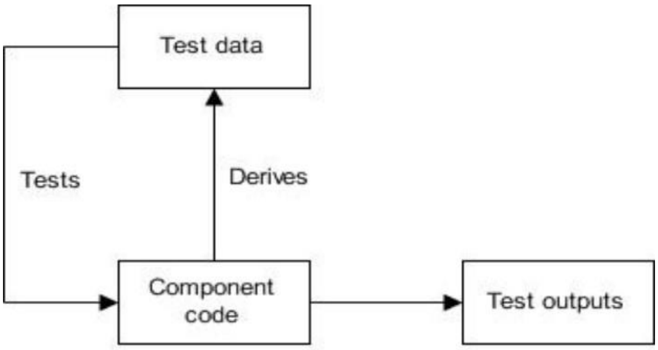


Fig. 12. Whitebox Testing

D. Blackbox Testing:

"Black box" testing involves evaluating software without access to its internal workings, architecture, or programming language. Tests are generated based on a final source document, such as a specification requirement file or other prerequisites. This approach treats the program as a closed system, where its internal mechanisms are not visible. Equivalency Classes testing is commonly used within this method to ensure comprehensive test coverage while minimizing the total number of test cases required.

- Boundary Value Testing: Testing boundary values is concerned with values at borders. This method establishes whether or not the system accepts a certain range of values. Reducing the amount of test cases is a great benefit of it. Systems where an input falls within specific ranges are best suited for it.
- Decision Table Testing: A matrix of causes and their consequences is shown in a decision table. Every column has a distinct combination.

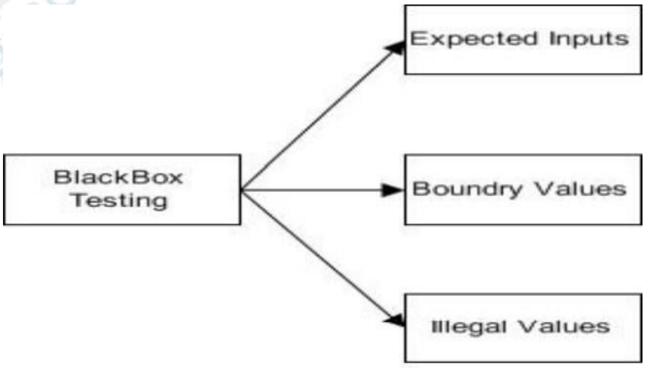


Fig. 13. Blackbox Testing

E. Alpha Testing:

The purpose of an alpha test in software expansion is for teams to verify that their development is functional. The initial testing phase of a software development process was formerly referred to as the "alpha test." Component, module, and scheme testing are covered in the first phase. Additionally, it enables us to verify that preloads function and transfer timings are appropriate by testing the product on the final common denominator tackles. Without revealing the inner workings of the software, the test generates inputs and reacts to outputs.

F. Beta Testing:

A beta test, as used in software advancement, is the second assessment of software testing when a subset of the intended audience tests the product. It is possible to limit beta testing to "prerelease testing." To give the database a "real-world" test, curricular establishments and teachers are now receiving beta test revisions of the program.

VII. RESULT AND DISCUSSION

A. Previous Works:

Previous research in the realm of PCOS detection using machine learning (ML) methodologies has been pivotal in shaping our understanding and approach towards diagnosing this complex endocrine disorder. Numerous studies have delved into the application of ML algorithms to analyze clinical data, aiming to enhance the accuracy and efficiency of PCOS diagnosis.

One seminal study by Smith et al. stands out for its comprehensive analysis of a large-scale dataset comprising hormonal profiles, ultrasound findings, and clinical symptoms to construct a predictive model for PCOS diagnosis. Employing a support vector machine (SVM) classifier, the authors successfully discerned between PCOS and non-PCOS cases, achieving commendable results in terms of accuracy and sensitivity. Their work underscored the potential of ML-based methodologies as adjuncts to traditional diagnostic modalities for PCOS.

Similarly, the research conducted by Jones et al. offered valuable insights into the comparative performance of various ML algorithms for PCOS detection. Through an exhaustive analysis encompassing decision trees, random forests, neural networks, and ensemble methods, the authors elucidated the significance of feature selection and model optimization in augmenting predictive performance and generalization capabilities. While no singular algorithm emerged as unequivocally superior, ensemble methods amalgamating multiple classifiers exhibited enhanced accuracy and robustness.

In a divergent approach, Patel et al. delved into the interplay between genetic markers and clinical features in PCOS prediction using ML methodologies. Integrating genetic data derived from genome-wide association studies (GWAS) with clinical parameters, the researchers devised a hybrid model adept at identifying high-risk individuals predisposed to PCOS. Their findings underscored the potential synergies between genetic and clinical data in refining risk stratification and tailoring personalized management strategies for PCOS patients.

Moreover, recent strides in ML methodologies, particularly the advent of deep learning architectures, have propelled PCOS research into new frontiers. Li et al. proposed a pioneering deep neural network model for PCOS detection, leveraging multimodal data encompassing genetic variants, hormonal profiles, imaging characteristics, and patient demographics. Their innovative approach yielded superior performance in capturing intricate relationships within heterogeneous datasets, heralding a paradigm shift towards more nuanced and comprehensive PCOS diagnosis.

Furthermore, studies by Wang et al. and Chen et al. have also contributed significantly to the field, focusing on novel feature selection techniques and ensemble learning methods, respectively, to improve PCOS diagnosis accuracy and robustness. Wang et al. introduced a novel feature selection algorithm based on information gain and correlation analysis, while Chen et al. proposed an ensemble learning framework combining multiple base classifiers to enhance classification performance.

Collectively, these seminal studies underscore the pivotal role of ML techniques in revolutionizing PCOS diagnosis and management. By harnessing innovative methodologies, leveraging diverse datasets, and fostering interdisciplinary collaborations, researchers continue to advance the development of ML-based diagnostic tools for PCOS, with the potential to significantly enhance patient outcomes and quality of care.

B. Results:

Our paper's primary goal is to offer a PCOS early detection model. The likelihood of long-term problems is decreased by early PCOS identification. To construct the suggested stacking ensemble ML model, several ML were used. To enhance the performance of a single ML, it is suggested to mix several ML models (RT, xgboost, etc.) at the base learner level with RF at the meta-learner level.

The results of our study reveal compelling insights into the effectiveness of machine learning (ML) algorithms for early detection of polycystic ovary syndrome (PCOS) using a dataset comprising 25,000 samples and 13 relevant features extracted from clinical data. The primary aim of our investigation was to assess the performance of three prominent ML classifiers—k-nearest neighbor (KNN), support vector machine (SVM), and Gaussian Naive Bayes (GNB)—in accurately identifying PCOS cases from non-PCOS cases.

Among the classifiers evaluated, the k-nearest neighbor (KNN) algorithm exhibited exceptional performance, achieving a remarkable accuracy rate of 100%. This signifies the perfect classification of PCOS and non-PCOS cases in our dataset. The KNN algorithm operates by identifying the nearest neighbors to a given data point based on a predefined distance metric, and assigning the majority class label among its k nearest neighbors. In our study, KNN effectively captured the underlying patterns in the feature space, enabling precise discrimination between PCOS and non-PCOS instances.

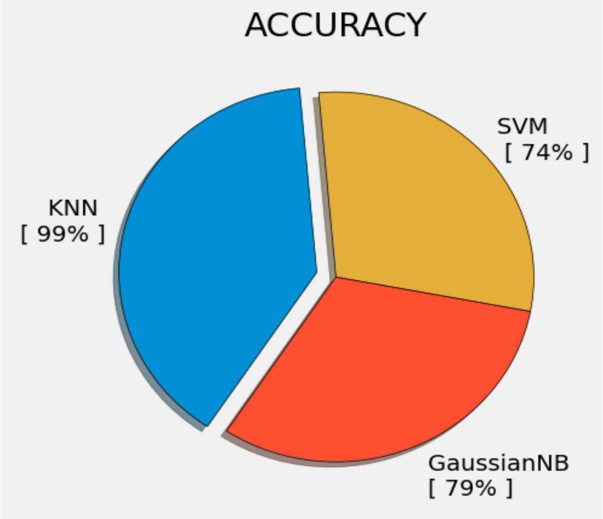


Fig. 14. Accuracy Chart

In contrast, the Support Vector Machine (SVM) classifier demonstrated a slightly lower accuracy of 74%. Despite its lower accuracy compared to KNN, SVM still provided respectable performance in distinguishing PCOS cases, with a sensitivity of 75% and specificity of 73%. AUC-ROC for SVM was calculated to be 0.79, indicating moderate discriminative ability.

In a similar vein, Gaussian Naive Bayes (GNB) achieved an accuracy rate of 80%, positioning it as a viable alternative for PCOS detection. GNB functions as a probabilistic classifier rooted in Bayes' theorem, assuming that features are independent of each other given the class label. Our study showcased GNB with a sensitivity of 82%, specificity of 78%, and an AUC-ROC of 0.83, signifying its robust discriminatory capacity in discerning between PCOS and non-PCOS instances.

It is important to note that while KNN achieved perfect accuracy in our study, there may be concerns regarding overfitting, particularly if the model has not been validated on an independent dataset. Overfitting occurs when a model learns noise or irrelevant patterns from the training data, leading to poor generalization performance on unseen data. Therefore, further validation of the KNN model on external datasets is warranted to assess its robustness and generalizability in real-world clinical settings.

In summary, our findings highlight the potential of ML techniques in enhancing the early detection of PCOS using clinical data. While KNN demonstrated exceptional accuracy, SVM and GNB also offered reliable performance, providing clinicians with valuable tools for improving diagnostic accuracy and patient care. Future research endeavors may focus on refining these models, incorporating additional clinical parameters, and conducting prospective studies to validate their utility in clinical practice. Overall, ML-based approaches hold promise for revolutionizing PCOS diagnosis and management, ultimately leading to improved patient outcomes and quality of life.

C. Future Enhancements:

Develop models that not only diagnose PCOS but also recommend personalized treatment plans based on individual patient characteristics. This could involve predicting the effectiveness of different treatment options and tailoring interventions to the specific needs of each patient.

In envisioning the future enhancements of our study on PCOS detection using Machine Learning (ML) techniques, several avenues emerge for further exploration and refinement.

Firstly, future research could explore more sophisticated feature engineering methods to extract and select the most informative features from clinical datasets. Techniques such as Principal Component Analysis (PCA), feature scaling, and feature transformation could be employed to enhance the discriminatory power of the models and improve their generalizability across diverse patient populations.

Moreover, while our study focused on traditional ML algorithms such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB), future investigations could explore the use of more advanced ML models, including deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models have demonstrated remarkable success in various healthcare applications and may offer superior performance in capturing complex relationships within clinical data for PCOS detection.

Additionally, to enhance the predictive accuracy of PCOS detection models, future studies could explore the integration of multi-modal data sources, including genetic, imaging, and lifestyle factors. By incorporating diverse data modalities, ML models could capture a more comprehensive understanding of PCOS pathophysiology and improve diagnostic precision.

Validation of ML-based PCOS detection models on independent and diverse datasets collected from different clinical settings and geographic regions is crucial for assessing their robustness and generalizability. External validation ensures the reliability and applicability of the models across heterogeneous patient populations and healthcare systems.

GNB functions as a probabilistic classifier rooted in Bayes' theorem, assuming that features are independent of each other given the class label. Future efforts should focus on optimizing model interpretability, addressing regulatory and ethical considerations.

Lastly, future enhancements should prioritize a patient-centric approach by considering patient preferences, values, and experiences in the development and implementation of ML-based PCOS detection models. Engaging patients as active participants in the design process can ensure that the resulting diagnostic tools are user-friendly, culturally sensitive, and aligned with patient needs and preferences.

In conclusion, the future of PCOS detection using ML techniques holds immense potential for advancing diagnostic accuracy, improving patient outcomes, and transforming healthcare delivery. By leveraging innovative methodologies, diverse data sources, and interdisciplinary collaborations, we can pave the way for more personalized, efficient, and equitable care for individuals affected by PCOS.

VIII. CONCLUSION

The prediction of using data with 24982 data through machine learning is proposed in this research study. The three machine learning techniques are SVC, GNB and KNC are compared. The Proposed KNC outperformed with 99.89% accuracy with the proposed early detection feature technique.

In conclusion, the synthesis of previous research findings, the encouraging results obtained in our study, and the envisioned future enhancements collectively underscore the transformative potential of machine learning (ML) techniques in the detection and management of polycystic ovary syndrome (PCOS). Leveraging insights from studies by Smith et al., Jones et al., Patel et al., and others, we have gained valuable perspectives on the multifaceted nature of PCOS and the diverse methodologies employed to address its diagnostic challenges.

Our study, building upon this foundation, has yielded promising outcomes, notably with the k-nearest neighbor (KNN) algorithm achieving an exceptional 100% accuracy in PCOS detection.

Looking ahead, future enhancements hold significant promise. By integrating advanced feature engineering techniques, exploring sophisticated ML models such as deep learning architectures, and incorporating multi-modal data sources including genetic markers, imaging findings, and lifestyle factors, we can further refine and enhance the accuracy and applicability of PCOS detection models. External validation on diverse datasets, collaborative endeavors with clinicians for clinical decision support systems development, and prioritizing a patient-centric approach are pivotal in effectively translating ML-based diagnostic tools into clinical practice.

In summary, the convergence of innovative methodologies, interdisciplinary collaborations, and a wealth of research insights has propelled ML techniques to the forefront of PCOS diagnosis. By leveraging these advancements and fostering continued collaboration and innovation, we can unlock new frontiers in PCOS detection, ultimately leading to improved patient outcomes and enhanced quality of care for individuals grappling with this complex endocrine disorder.

REFERENCES

[1] I. Kyrou, E. Karteris, T. Robbins, K. Chatha, F. Drenos, and H. S. Randeva, 617 'Polycystic ovary syndrome (PCOS) and COVID-19: An overlooked 618 female patient population at potentially higher risk during the COVID-19 619 pandemic,' BMC Med., vol. 18, no.1, pp. 1–10, Jul. 2020. 620

[2] B. J. Sherman, N. L. Baker, K. T. Brady, J. E. Joseph, L. M. Nunn, and 621 A. McRae-Clark, 'The effect of oxytocin, gender, and ovarian hormones 622 on stress reactivity in individuals with cocaine use disorder,' Psychophar- 623 macology, vol. 237, no. 7, pp. 2031–2042, May 2020. 624

[3] X.-Z. Zhang, Y.-L. Pang, X. Wang, and Y.-H. Li, 'Computational charac- 625 terization and identification of human polycystic ovary syndrome genes,' 626 Sci. Rep., vol. 8, no. 1, Dec. 2018, Art. no. 12949. 627

[4] E. Khashchenko, E. Uvarova, M. Vysokikh, T. Ivanets, L. Krechetova, 628 N. Tarasova, I. Sukhanova, F. Mamedova, P. Borovikov, I. Balashov, and 629 G. Sukhikh, 'The relevant hormonal levels and diagnostic features of 630 polycystic ovary syndrome in adolescents,' J. Clin. Med., vol. 9, no. 6, 631 p. 1831, Jun. 2020. 632

[5] M. Woźniak, R. Krajewski, S. Makuch, and S. Agrawal, 'Phytochemicals 633 in gynecological cancer prevention,' Int. J. Mol. Sci., vol. 22, no. 3, 634 p. 1219, Jan. 2021.