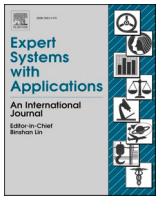
Expert Systems With Applications 217 (2023) 119532

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



# Early identification of PCOS with commonly known diseases: Obesity, diabetes, high blood pressure and heart disease using machine learning techniques

Shivani Aggarwal *, Kavita Pandey *

*Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, India*

ARTICLE INFO

*Keywords:*
Polycystic ovary syndrome (PCOS)
Obesity
High blood pressure
Diabetes
Heart diseases
Machine learning
K-nearest neighbor
Random forest
K-means clustering
Gradient boosting
Decision tree
Support vector machine
Logistic regression
Hybrid random forest logistic regression (RFLR)
Stacking
Feature selection

ABSTRACT

Polycystic Ovary Syndrome (PCOS) is a health disorder that affects around 10 million women worldwide. Due to a lack of awareness in India, it affects one out of every-five women. If not diagnosed at an early stage, it leads to various other harmful diseases as well like diabetes, high blood pressure, obesity, heart disease etc. Thus, PCOS identification at an early stage is essential. Although several expensive tests are available there is no proper treatment and people are not aware of the same. The main objective of this article is to find the diseases which can help early identification of PCOS in a nutshell, we want to answer the question "Can we identify some commonly known diseases which taken as an indication of having PCOS". To achieve this objective, data amalgamation is performed on the identified disease datasets ahead and feature selection methods are applied resulting in 8 parameters and 985 records. To validate the desired hidden relation of PCOS with other diseases, supervised and unsupervised learning algorithms are applied. This article also tries to answer only important features of other diseases that are required for PCOS prediction. The performance metrics analysis with all features, important features and remaining features shows that the important features give the best result for PCOS identification.

## 1. Introduction

PCOS is a complicated hormonal, metabolic and reproductive ailment that influences 1-in-10 women of childbearing age. PCOS is the main reason for infertility in women. Women with PCOS are at higher risk for developing type 2 diabetes and cardiovascular disease. The National Institutes of Health (NIH) reckons that more than 50 % of women with PCOS will become diabetic or prediabetic before age 40 (Wilson and Metink-Kane, 2012). Some researchers have also speculated that women with PCOS are at three times higher risk for endometrial cancer, two times higher risk for ovarian cancer, and two to four times higher risk for breast cancer(Aggarwal & Pandey, 2022). Some studies reveal that due to the symptoms of anxiety and depression in PCOS women, suicidal attempts among these women are up to seven times more than compared to other women. Nowadays, pre-teens and teens are also developing PCOS because of unbalanced diets. Earlier detection can give them the chance to better manage the emotional, internal, and physical consequences of PCOS(Aggarwal & Pandey, 2021). It can also help them to prevent the outbreak of more critical illnesses associated with PCOS. Despite affecting millions of women and the serious health results, PCOS is unknown to most people. It is more shocking when studies reveal that around 50 % of the women living with PCOS are going undiagnosed(Escobar-Morreale, 2018). Recent research from numerous studies, particularly in neonates who have risk factors associated with PCOS development, suggests that PCOS may begin in utero. This may happen to a new born with premature birth and higher birth weight who afterward cope with growth or put on weight postpartum(El Hayek et al., 2016).

In India, PCOS is a particularly common disease with around 1 in every 5 women distressed by the condition. Around 5–10 percent of women of reproductive age are suffering from PCOS (Centers for Disease Control and Prevention, 2020). The greatest knowledge of PCOS's

* Corresponding authors.
*E-mail addresses:* shivaniagarwal850@gmail.com (S. Aggarwal), kavita.pandey@jiit.ac.in (K. Pandey).

https://doi.org/10.1016/j.eswa.2023.119532
Received 15 May 2022; Received in revised form 6 January 2023; Accepted 6 January 2023
Available online 12 January 2023

pathogenesis considers this as a complex condition including abnormal insulin signaling, increased oxidative stress, uncontrollable ovarian steroidogenesis, and environmental and genetic factors. Theca cells in PCOS patients are intrinsically activated to produce steroids, which leads to high levels of androgen secretion even in the lack of trophic elements. By using microarray gene analysis, it was also discovered that the gene expression of various participants in the insulin signalling pathways had changed. In PCOS individuals, oxidative stress can lead to hyperandrogenism and insulin resistance on its own. The genomic identification of PCOS-susceptibility genes and the genetic aggregation of PCOS provide evidence for the involvement of genetics in the genesis of this condition(El Hayek et al., 2016). Unfortunately, only less than 50 percent of women are correctly diagnosed and therefore most women don't receive proper medication. Some studies reveal that insulin resistance is one of the predominant factors in the spread of PCOS as well as for a diabetic person, insulin resistance is a risk factor and around 40 % of women with PCOS evolve either prediabetes (or) diabetes by the age of 40 years (Causes of Sleep Apnea, 2021). Insulin sensitivity can be increased in women having PCOS with the anti-diabetes medication metformin. Patients with PCOS are also at a higher risk for the cardiac ailment, the possibility of a heart attack is 4–7 times more in women with PCOS than in their fellows without PCOS, also around 50 % of women with PCOS suffer from obesity. Obesity is the main reason behind these women commonly developing temporary cessation of breathing during sleep i.e., obstructive sleep disorder. However, obesity in PCOS women can be controlled up to some extent by controlling diet and regular physical activities. It has been observed from the studied literature, that PCOS leading the cause of metabolic syndrome i.e., heart disease, diabetes, obesity and high blood pressure(Chen & Pang, 2021).

The objective of this article is to determine whether the above diseases lead to PCOS or not. Generally, women do not go for PCOS testing because they are unaware and for its testing doctor's prescription is required. Usually, they do some common tests like fasting glucose test, pulse analysis and blood pressure measurement. In our analysis, if a woman has obesity, high blood pressure, diabetes and heart disease. We can predict that she is more likely to have PCOS. To verify, a data amalgamation is performed by merging two different datasets of heart diseases and diabetes. Data amalgamation provides a new dataset on which a feature selection technique is applied to get the eight selected attributes(Zhang et al., 2015). Then, supervised learning classification algorithms are applied to the eight attributes to analyse the performance metrics of the model. Six classification algorithms are used Decision tree, Gradient Boosting, Random Forest, Logistic regression, K-nearest neighbor, Hybrid RFLR and Support vector machine(Osisanwo and O Awodele, 2017). For further refining, the eight features and the four most important features are taken according to the feature selection algorithm ranking which reflects the same diseases that are provided in the literature. The performance metrics analysis of all features, four important features and four remaining features are compared which shows that all features and four important features give good performance outcomes instead of the four remaining features. The same process is also applied to unsupervised K- means clustering algorithms respectively.

The main contributions of this article are listed below which are also explained in Fig. 1:

- Highlights the relevance of some commonly known diseases such as obesity, heart disease, high blood pressure and diabetes for early detection of PCOS.
- Created a new dataset by using different disease datasets which are further utilized for implementing supervised and unsupervised learning algorithms.
- Supervised learning algorithms are applied to analyse the performance metrics of the model for all features, important features and remaining features for the prediction of PCOS.
- To justify the outcome of supervised learning algorithms, an unsupervised learning algorithm has been applied to the amalgamated dataset.
- The findings of this article can help healthcare professionals with the early detection of PCOS by using only a few features.

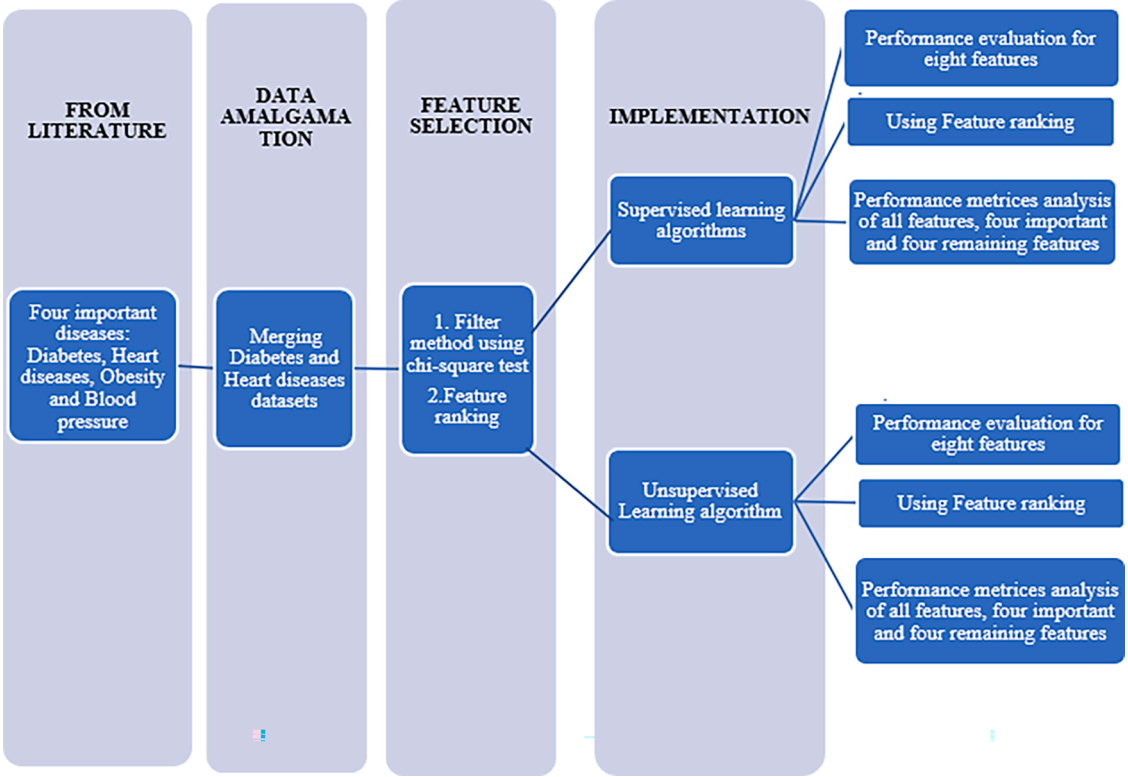



**Fig. 1.** Flowchart of the main contribution.

The rest of the article is arranged as follows, Section 2 presents the literature review as well as associated knowledge on PCOS and related diseases. Section 3 gives a brief description of the data set and software used for our work. This is followed by the proposed work in section 4. The following section 5 describes the results of supervised as well as unsupervised learning algorithms and the performance metrics analysis. Section 6 concludes the article and also enlightens the future work relevant to our study. In the last section, a list of useful references is provided.

## 2. Related work

This section provides an overview of the existing literature about the relationship between PCOS and other diseases. It gives an overview that PCOS leads to obesity, heart disease, diabetes and high blood pressure in women but there is no clear evidence regarding the effect of these diseases on PCOS diagnosis. The whole literature on the relationship of PCOS with other diseases is shown in Table 1. These four sub-sections provide an overview of PCOS and obesity, PCOS and diabetes, PCOS and heart disease, PCOS and high blood pressure.

(Kyrou et al., 2020) provides an overview of pre-existing medical conditions, including long-term cardio-metabolic diseases such as obesity, hypertension and diabetes, which have also been generally acknowledged as fundamental risk features for PCOS patients. PCOS is the most repeated endocrine disorder in women of reproductive age which may increase by 10–15 %, relying on the educated population and the concerning diagnostic criteria. Witchel et al. described in their article that the control and interactions of androgens and AMH during folliculogenesis is an unanswered question (Witchel et al., 2019). AMH and androgens are necessary for a healthy cycle of ovulation. AMH seems to play a dual role in favouring preantral follicle survival while impairing antral follicle growth later. These findings imply that increased prenatal AMH exposure can encourage PCOS-specific abnormal neuroendocrine activity which has been found to have higher LH pulse frequency, amplitude, and LH/FSH ratios. The pulse frequency of GnRH modulates the pulse frequencies of LH and FSH. LH pulse frequency increases with increased GnRH pulse frequency, but FSH pulse frequency decreases. As is seen in PCOS, elevated diastolic GnRH production is probably causing the increased LH pulse amplitude and pulse frequency.

### a. PCOS and Obesity

(Oberg et al., 2019) recommend a treatment for obese women along with PCOS as a lifestyle invention. Although, it is still undefined that lifestyle inventions enhance reproductive functioning. Sixty-eight women of age between 18 and 40 years with a BMI is about 27 attaining all Rotterdam criteria for PCOS diagnosis. (Glueck & Goldenberg, 2019) described that obesity may affect approximately 50 % of women with PCOS. Visceral weight is often found in women with PCOS. Obesity is one of the reasons for increased insulin resistance. (Wang and Wu, 2018) encourage women with PCOS for weight loss to analyze the efficacy of orlistat, liraglutide, metformin and inositol. (Lie et al., 2020) employed PRISMA guidelines as a systematic review for health professionals to originate practical tools and consult women with PCOS about a feasible healthier lifestyle.

### b. PCOS and Diabetes

(Kakoly et al., 2018) represent a *meta*-analysis report which shows that women affected by PCOS have inherent insulin resistance and a 4-fold high prevalence of type 2 diabetes along with obesity. (Condorelli et al., 2017) discussed a primary objective of endocrinologists to analyze the patients having aesthetic disorders reproductive, metabolic and potential oncologic risks as per their phenotypes. Accordingly, they suggest a lifestyle change and suitable pharmacologic treatment to

**Table 1**
Literature analysis of PCOS with related diseases.

| S. No | References | Related Diseases | Approaches |
|---|---|---|---|
| 1. | (Oberg et al., 2019) | Obesity | Strategy for behavioral modification |
| 2. | (Glueck & Goldenberg, 2019) | Obesity | Weight loss through bariatric surgery |
| 3. | (Wang and Wu, 2018) | Obesity | To assess the efficiency of orlistat, metformin, inositol, and liraglutide in causing weight loss in PCOS women. |
| 4. | (Lie et al., 2020) | Obesity | The comprehensive review was carried out under PRISMA standards. |
| 5. | (Saravana et al., 2015) | Diabetes | Diabetes prediction using a predictive analytic technique in a Hadoop/Map Reduce scenario. |
| 6. | (Anagnostis et al., 2018) | Diabetes | To present the most recent information on the metabolic effects in PCOS patients with concurrent cardiovascular disease (CVD) concern. |
| 7. | (Kakoly et al., 2018) | Diabetes | To investigate how factors such as obesity, ethnicity and the Type 2 diabetes mellitus diagnosis method may contribute to the observed variation in Type 2 diabetes mellitus and IGT prevalent in PCOS. |
| 8. | (Zhu et al., 2021) | Diabetes | Focusing on women with high-risk risk traits rather than all PCOS patients will help reduce cardiometabolic problems. |
| 9. | (Condorelli et al., 2017) | Diabetes | The metabolic PCOS may be accelerated by sedentary lifestyles and excessive eating. |
| 10. | (Torchen, 2017) | Heart Disease | Review the research on metabolic risk for PCOS, highlight current advancements in the field and point out critical knowledge gaps that need to be filled in future research. |
| 11. | (Zhao et al., 2021) | Heart Disease | To investigate the possible metabolic features of PCOS-MS and find sensitive biomarkers that can be used for clinical screening, diagnosing, and medication. |
| 12. | (Wekker et al., 2020) | Heart Disease | The analysis assesses an increased chance of cardiometabolic disease, which is crucial for doctors to communicate with patients and account for when evaluating the risk of cardiovascular disease in PCOS-positive women. |
| 13. | (Ali & Ali, 2020) | Heart Disease | To investigate how Glutathione S-transferases (GST) and serum apelin-36 activities relate to metabolic profiles, hormonal and their relationship to the risk of heart disease in both healthy and polycystic ovary syndrome (PCOS) patients. |
| 14. | (Marchesan & Spritzer, 2019) | High Blood Pressure | To evaluate the likelihood of metabolic disorders in PCOS-affected women with SAH as defined by the Joint National Committee on Detection, Identification, Assessment, and Medication of High Blood Pressure (JNC7) and the 2017 ACC/AHA criteria in women with PCOS criteria. |
| 15. | (Fauser et al., 2019) | High Blood Pressure | The goal of the current research was to identify links between women's preconception health and the wellness of their offspring. Before getting pregnant, 74 women who had been given a PCOS diagnosis based on the Rotterdam criteria underwent appropriate results. |
| 16. | (Özkan et al., 2020) | High Blood Pressure | They propose frequent ambulant blood pressure tracking for all PCOS patients to identify hidden |

(*continued on next page*)

**Table 1** (*continued*)

| S. No | References | Related Diseases | Approaches |
|---|---|---|---|
| | | | hypertension and avoid cardiovascular risks. |
| 17. | (Doroszewska et al., 2019) | High Blood Pressure | To briefly describe the research on this disease's impact on postmenopausal women's risk of hypertension and control of blood pressure. |
| 18. | (Mellembakken et al., 2021) | High Blood Pressure | The article's objective was to evaluate blood pressure (BP) between controls and normal-weight women having PCOS who were matched for age and BMI. |

protect them from the risk of cardiovascular disease prevalent in such women.

(Anagnostis et al., 2018) provides a brief description of a woman with PCOS who has a high probability of having type 2 diabetes. Insulin resistance (IR) plays a fundamental role in metabolic syndrome in PCOS patients. It is estimated that 30 % of slim and 70 % of overweight women with PCOS have insulin resistance, even though the term "insulin resistance" has also been defined differently. In comparison to age-related and weight-matched women without PCOS, women with PCOS had a greater risk of insulin resistance and glucose intolerance. (Saravana et al., 2015) employ a Hadoop/Reduce system and a predictive analysis method to anticipate the forms of diabetes that are prevalent, the difficulties affiliated with it, and the method of treatment that will be delivered. According to the findings, this approach is an effective remedy for treating and patient care, with superior outcomes such as accessibility and affordability. (Zhu et al., 2021) undertook two-sample Mendelian randomization (MR) study to evaluate the links between PCOS with coronary heart disease (CHD), stroke and type 2 diabetes.

### c. PCOS and Heart disease

(Wekker et al., 2020) estimate the risk of having heart disease in women with PCOS. Although there have been no recent comprehensive reviews or *meta*-analyses of longitudinal research that evaluate the link between cardiometabolic disease and PCOS. Doctors must know about cardiometabolic disease risk in women having PCOS. (Zhao et al., 2021) prepare a survey on the metabolism aspects of PCOS-MS. This research included 44 PCOS patients with metabolic syndrome (MS), 34 PCOS patients without MS and 32 healthy control patients. The goal was to find accurate biomarkers so that diagnosis, clinical screening and treatment could be targeted.

(Torchen, 2017) summarize those longitudinal investigations evaluating the advancement of metabolic abnormalities and the establishment of cardiovascular events in PCOS are urgently required. Due to a shortage of data on the effect of specific early prevention techniques for T2D and cardiovascular disorders in PCOS, particularly when responses to T2D prevention or therapy differed in women having PCOS compared to women with other types of T2D. To proceed toward the formation and implementation of all such methods of prevention, further research into the early beginnings of PCOS is required. (Ali & Ali, 2020) investigated the relevance of Glutathione S-transferases (GST) and serum apelin-36 activity about hormonal and metabolic profiles as well as their relevance towards the risk of cardiovascular disease (CVD) also a set of fifty-four (PCOS) women and thirty-one healthy women were evaluated as a control condition.

### d. PCOS and High blood pressure

(Doroszewska et al., 2019) provide a brief review of women with PCOS having higher cardiovascular risk factors, such as hypertension. Hypertension is 2.5 times more common in these people than in their healthy colleagues. Furthermore, regardless of the patient's age, hyperandrogenaemia is related to higher blood pressure, insulin sensitivity, being overweight, and hyperlipidemia. Most of the authors believe that PCOS increases the risk of high blood pressure in post-menopausal women, along with all of its implications. In women with PCOS, high blood pressure occurs as a result of poor vascular flexibility and vascular dysfunction. The levels of insulin resistance and androgens are proportional to vascular dysfunction. (Marchesan & Spritzer, 2019) determine that women having SAH, as described mostly by 2017 ACC/AHA guidelines, had a higher chance of cardiovascular co-morbidities. Reducing abnormal blood pressure thresholds appear to be suitable for women having PCOS, offering a simple screening test for cardiovascular co-morbidities and early prevention strategies.

(Fauser et al., 2019) work to figure out if there were any links between the women's preconception health and the health of the child. Before conception, 74 women suffering from PCOS according to Rotterdam criteria were examined comprehensively. Second, researchers examined the links between preconception blood pressure, insulin resistance (HOMA_IR), androgens, and LDL cholesterol in mothers with PCOS, as well as BMI and offspring blood pressure. (Özkan et al., 2020) recommend that all PCOS patients have non-invasive blood pressure tracking to detect concealed hypertension and avoid cardiovascular complications. The goal of this review was to compare the frequency of mask hypertension in patients with PCOS. The prevalence of concealed hypertension was found to be increased in PCOS patients. (Mellembakken et al., 2021) want to see how blood pressure varied between normal-weight PCOS women as well as over-weight PCOS women. They analysed a subgroup of 793 normal body weight women with a BMI of 25 and 512 overweight PCOS women from the Nordic cross-sectional base among 2615 Nordic ethnic women. From the literature, it is clearly shown that PCOS can cause the above diseases but vice-versa has not yet been proven. Motivated by this, this paper, it is demonstrated that the above diseases may also cause PCOS. In the next section, a brief description of data preparation and software used is explained for a better understanding of the dataset.

## 3. Material and methods

### 3.1. Data preparation

#### 3.1.1. Data set

Two types of datasets are used for this study: Diabetes (Pima Indians Diabetes Database, n.d.) and Heart disease (Heart Disease Dataset, 2019). Both datasets are taken from Kaggle, a popular distributed data platform managed by data scientists. Below are some of the key highlights of datasets are mentioned as follows:

1. Firstly, feature selection is applied to both diabetes and heart disease datasets.
2. The diabetes dataset has 8 parameters and 768 records naming Blood pressure, BMI, Insulin, Glucose, Pregnancies, Diabetes Pedigree, Skin thickness and Age. After applying feature selection, the three parameters as BMI, Glucose, and Age have been selected.
3. The heart disease dataset has 13 parameters and 1273 records naming Sex, Chest Pain, Blood pressure, cholesterol, fasting blood pressure, Age, Heart rate, resting electrocardiographic, exercise-induced angina, ST depression, number of major vessels colored by fluoroscopy, Exercise ST segment, Maximum heart rate. After applying feature selection, the five parameters as chest pain, cholesterol, resting blood pressure, fasting sugar and the number of major vessels coloured by fluoroscopy have been selected.
4. Data amalgamation is performed for merging both datasets.
5. After applying data amalgamation and feature selection, the new dataset has 8 parameters and 985 records naming Age, BMI (Body mass index), Glucose (an oral glucose tolerance test, plasma glucose concentration was measured after 2 h), cp (chest pain type), ca

(number of major vessels (0–3) colored by fluoroscopy), chol (serum cholesterol in mg/dl), trestbps (resting blood pressure) and fbs (fasting blood sugar) respectively which is shown in the below Table 2.

6. The target column represents patients diagnosed with PCOS or not. Zero denotes a null score, indicating that the patient has not been diagnosed with PCOS.

#### 3.1.2. Software used

Jupyter notebook is used to conduct this experiment with the composition of the system as 8 GB RAM, Intel Core i5-8265U CPU 1.60 GHz processor and Windows 11 64-bit operating system. It is also used for sharing and creating documents and it is also an open-source web application(Bloice & Holzinger, 2016). Kaggle, a data repository used to compile the medical dataset for the experiment. Python and other programming languages have been used jupyter notebook.

## 4. Proposed methodology

It has been observed from the literature that if a woman is having PCOS, then there is a high chance that she also has heart disease, obesity, high blood pressure or diabetes. Generally, women go for regular check-ups of these commonly known diseases but they are unaware of PCOS tests. So, the objective is too early identification of PCOS from these commonly known diseases. Inspire by this, the paper aims to find out the answer to the question "Can these four diseases predict PCOS". Fig. 2 explains PCOS's relation with other diseases, which hormones get imbalanced and what tests are required. For better understanding labels are also provided on the left-hand side of Fig. 2.

All the above-mentioned diseases are common in PCOS patients and result in an imbalance of hormones which causes insulin resistance and hyperandrogenism. Insulin resistance is a response to the hormone insulin that causes blood sugar levels to rise. On the other hand, Insulin resistance and hyperinsulinemia may play a key role in the pathophysiology of estrogen deficiency in PCOS. In females, hyperandrogenism refers to an excess of the male sex hormone (testosterone) in the bloodstream. Hyperandrogenaemia should be an essential condition for the diagnosis of PCOS, according to an expert panel of the Androgen Excess Society in 2006(Lauritsen et al., 2019).

To detect the level of the imbalanced hormone, the Fasting glucose test (FGT) is used. To analyse this, data amalgamation and feature selection techniques are performed. Feature selection algorithms are applied to find the important parameters. After analysis, it is observed that these parameters are indicators of four diseases. This analysis has also been validated by using supervised and unsupervised learning algorithms. So, it can be stated that after knowing the features from the analysis. These four diseases obesity, diabetes, heart disease and high blood pressure are used to predict PCOS. Detailed implementation and results are explained in the next section.

**Table 2**
A view of the amalgamated dataset.

| Age | BMI | Glucose | cp | ca | Chol | trestbps | fbs | target |
|---|---|---|---|---|---|---|---|---|
| 21 | 28.1 | 89 | 1 | 0 | 210 | 150 | 1 | 1 |
| 30 | 25.6 | 116 | 1 | 0 | 210 | 150 | 0 | 0 |
| 36 | 25.9 | 95 | 0 | 0 | 183 | 138 | 0 | 0 |
| 35 | 32.3 | 116 | 0 | 0 | 183 | 138 | 0 | 0 |
| 35 | 43.4 | 93 | 0 | 0 | 183 | 138 | 1 | 1 |
| 35 | 35 | 136 | 0 | 0 | 183 | 138 | 0 | 0 |
| 38 | 34.1 | 117 | 2 | 0 | 215 | 152 | 0 | 0 |
| 37 | 40.2 | 133 | 2 | 0 | 215 | 152 | 1 | 1 |
| 38 | 32.5 | 109 | 2 | 0 | 215 | 152 | 1 | 1 |
| 37 | 48.8 | 137 | 2 | 0 | 215 | 152 | 1 | 1 |
| 37 | 29.5 | 144 | 2 | 0 | 215 | 152 | 0 | 0 |
| 38 | 34 | 124 | 2 | 0 | 215 | 152 | 1 | 1 |
| 38 | 39.5 | 106 | 2 | 0 | 215 | 152 | 0 | 0 |
| 37 | 21.9 | 114 | 2 | 0 | 215 | 152 | 0 | 0 |
| 37 | 39.1 | 130 | 2 | 0 | 215 | 152 | 1 | 1 |
| 38 | 41.8 | 151 | 2 | 0 | 215 | 152 | 1 | 1 |
| 37 | 25.2 | 119 | 2 | 0 | 215 | 152 | 0 | 0 |
| 36 | 42.9 | 151 | 0 | 0 | 183 | 138 | 1 | 1 |
| 35 | 24.5 | 90 | 0 | 0 | 183 | 138 | 0 | 0 |

## 5. Implementation and results

This section presents the results of supervised and unsupervised learning algorithms. For supervised learning, the first step is calculating the values of performance metrics such as accuracy, precision, recall and F1-score for various classification algorithms like the random forest, gradient boosting, decision tree, K-nearest neighbor, support vector machine, hybrid RFLR and logistic regression. The next step is the creation of feature importance graphs for all the supervised learning algorithms. The same graphs have been used to rank the features based on their scores. Learning curve graphs for all features, important features and remaining features are created. Then, a comparative analysis table has been created by taking the inputs of learning curve graphs and analyzing that the Gradient boosting algorithm gives the best accuracy for all features as well as for important features which indicate that PCOS can be determined by considering only the important features.

For unsupervised learning, the K-mean clustering algorithm has been applied. The Elbow method is applied to find out the number of clusters which is required to implement the model. Then, a scatter plot for all features, important features and remaining features has been created which depicts the overlapping of important and remaining features. For all features, important features and remaining features, a comparative analysis table has been prepared to calculate the Silhouette and Davies Bouldin score. These scores are good for all features and important features which signifies that fbs, BMI, glucose and trestbps are the four most important features for PCOS analysis. A detailed explanation of this work is further elaborated as follows:

### 5.1. Supervised learning classification algorithms results

**Step 1:** To validate the proposed idea, a confusion matrix is utilized. It is evaluated by using performance metrics such as precision, recall, accuracy and F1 score. False-positive (FP), False negative (FN), True positive (TP) and True negative (TN) are used to estimate the confusion matrix.

FP represents an output when the model accurately identifies the negative class labels.

FN represents an output when the model inaccurately identifies the negative class labels.

TP represents an output when the model accurately identifies the positive class labels.

TN represents an output when the model inaccurately identifies the positive class labels.

The prediction outcomes are assessed using the following evaluation criteria, as shown in Eqs. (1) – (4):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{F1} - \text{score} = \frac{2{*}precision{*}recall}{precision + recall} \tag{4}$$

Table 3 depicts the performance of all six supervised classification algorithms on all the above-mentioned performance metrics. Here, all features have been considered. The gradient boosting algorithm gives the best result for all the performance metrics.

**Step 2:** To find out the important features among all the features for PCOS diagnosis, feature importance graphs for all the classification

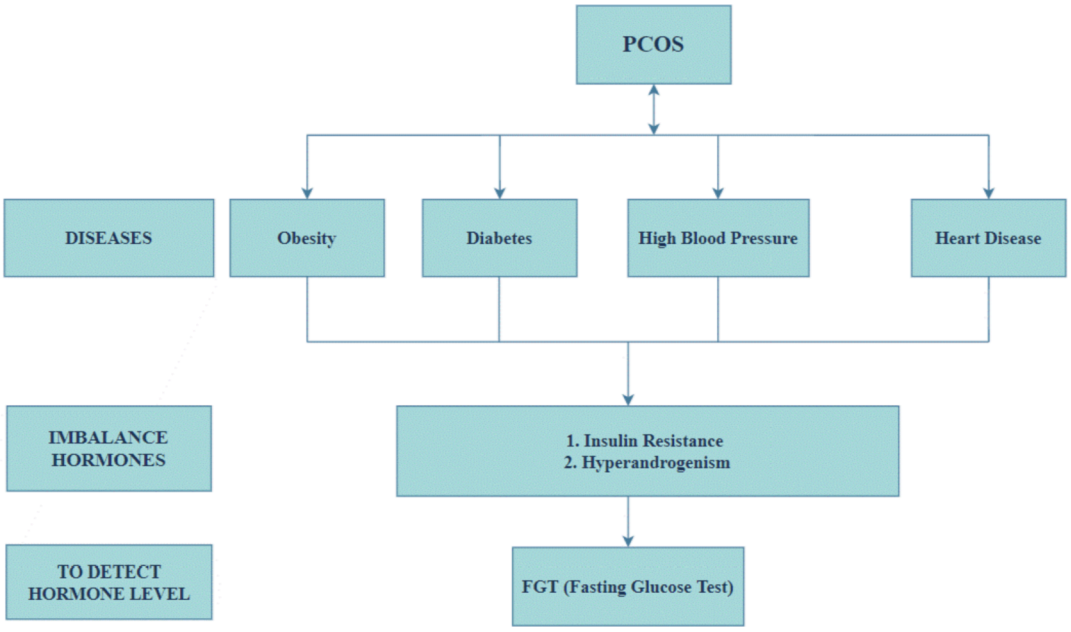



**Fig. 2.** Workflow of the proposed methodology.

**Table 3**
Performance metrics analysis for classification algorithms with consideration of all features.

| Classification Algorithms | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Random Forest | 98.5 % | 98.4 % | 96.7 % | 97.5 % |
| Decision Tree | 98.5 % | 96.8 % | 98.4 % | 97.5 % |
| Gradient Boosting | **98.9 %** | **98.4 %** | **98.4 %** | **98.4 %** |
| KNN | 73.6 % | 61.5 % | 39.4 % | 48.0 % |
| Logistic Regression | 98.5 % | 96.8 % | 98.4 % | 97.6 % |
| SVM | 98.5 % | 96.8 % | 98.3 % | 97.6 % |
| Hybrid RFLR | 81.2 % | 83.3 % | 49.2 % | 61.8 % |

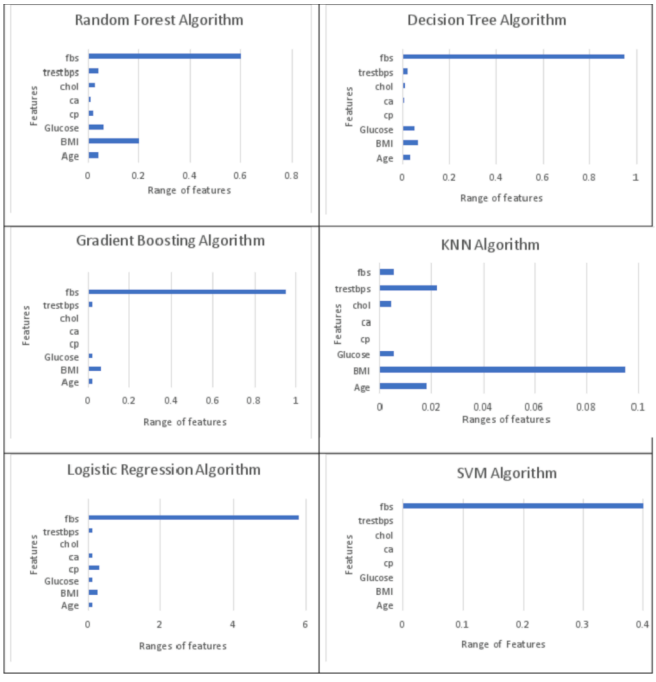



**Fig. 3.** Feature importance graph for different classification algorithms.

algorithms have been prepared as shown in Fig. 3. It can be visualized from Fig. 3 that with all the classification algorithms fbs, BMI, Glucose and trestbps are the important features for PCOS diagnosis.

**Step 3:** To find out the ranking of features for PCOS diagnosis, SelectKBest and chi2 feature selection methods are applied to features. SelectKBest method is used to select the features according to the K highest score. For classification algorithms, the chi2 method is adopted as a scoring function. A chi-squared test is a data analysis based on the findings of a diverse group of variables. Typically, it involves a contrast between two sets of statistical data. The target number of features is defined by K parameters obtained from the SelectKBest method. Ranking and scoring of features are provided by using both of the above-mentioned feature selection techniques and the outcomes of these methods are shown in Table 4. It is represented in Table 4 that BMI, fbs, Glucose and trestbps are the important features for PCOS diagnosis.

**Step 4:** The impact of the set of observations on the evaluation metrics is represented by using a learning curve. Below Figs. 4-6 show the learning curve with all six classification algorithms for all features, important features, and remaining features respectively.

In Fig. 4, the x-axis represents the dataset size and the y-axis represents the accuracy of the model. It is clearly shown in a learning curve with the random forest classifier that the accuracy of the model increases with the dataset size till a threshold point which is 200, after that it becomes constant T. (Vafeiadis et al., 2015) say that a random forest classifier is effective for huge datasets. The learning curve with the decision tree classifier cross-validation score line demonstrated that the accuracy of the model is frequently changed because in the decision tree, a small change in the data size can lead to a large change in the structure

**Table 4**
Feature ranking for all features.

| S.No | Ranking | Features | Score |
|---|---|---|---|
| 0 | 8 | Age | 0.54 |
| 1 | 2 | BMI | 577.46 |
| 2 | 3 | Glucose | 356.59 |
| 3 | 5 | ca | 14.61 |
| 4 | 6 | cp | 6.23 |
| 5 | 7 | chol | 6.04 |
| 6 | 4 | trestbps | 18.26 |
| 7 | 1 | fbs | 614.41 |

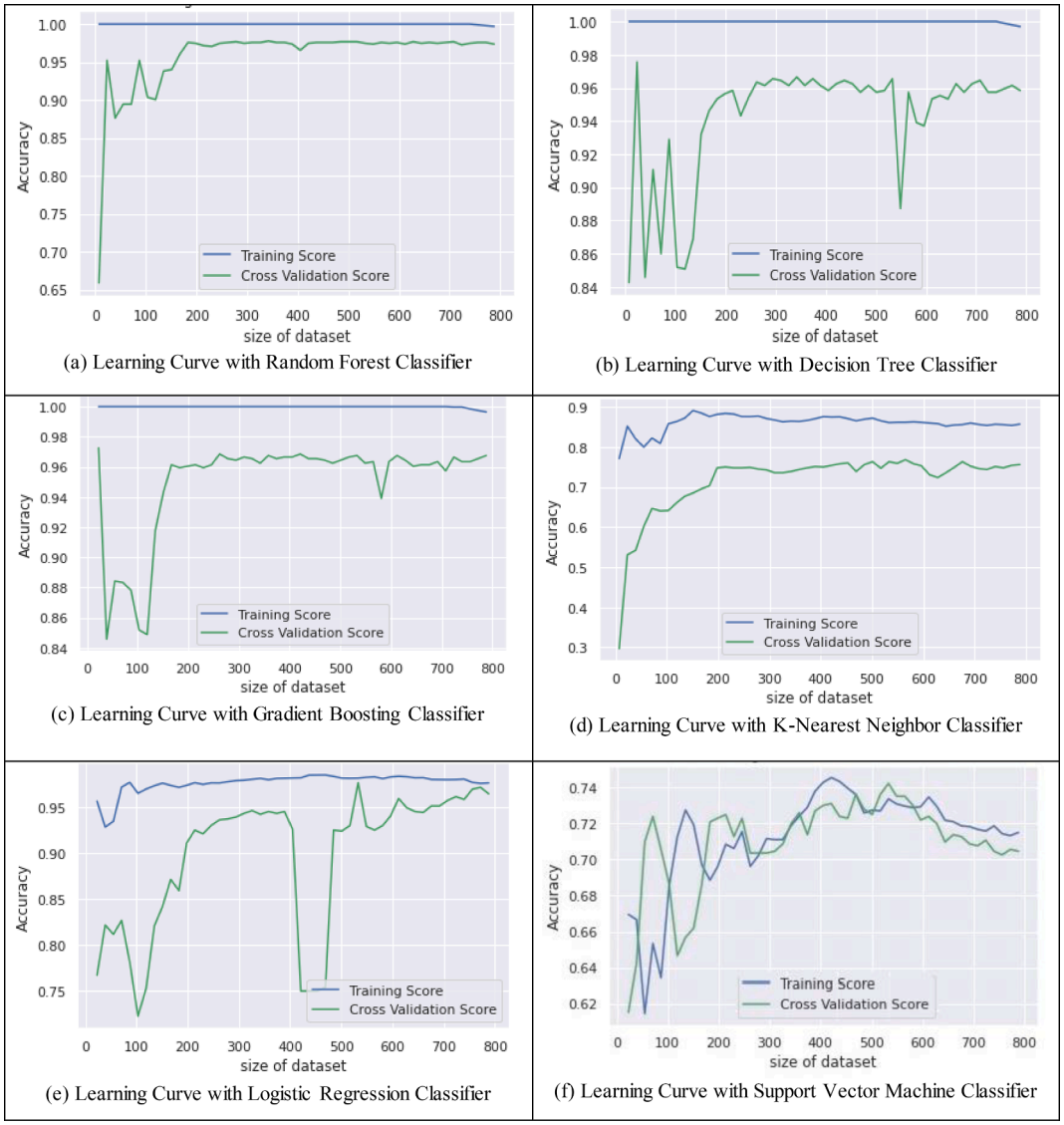



**Fig. 4.** Learning curve graphs for all features.

of the optimal decision tree.

Referring to Fig. 4(c), the learning curve with gradient boosting classifier revealed that at certain point accuracy of the model dropped and after a threshold point it became constant. Literature shows that gradient boosting gives the best accuracy for unbalanced data. This amalgamated dataset is also unbalanced, thus gradient boosting gives us the best accuracy. In Fig. 4(d), the learning curve with K- Nearest Neighbor classifier confirmed that the training score and cross-validation score accuracy changes continuously and it gives less accuracy compared to all five classification algorithms for the reason that K-Nearest Neighbor doesn't work well with large datasets.

In Fig. 4(e), the learning curve with the logistic regression classifier shows that the accuracy of the model for validation score is frequently changed but at a certain point of time it tends to be similar to the training score and it also gives good accuracy. Since it is a linear problem dataset and logistic regression works well on a linear decision surface. In Fig. 4(f), the learning curve with the support vector machine classifier indicates that the training score and cross-validation score vary so much since the support vector machine doesn't perform well when we have a large dataset because the required training time is higher.

Fig. 5, demonstrates the learning curve of all six classification algorithms for important features. It is clearly shown that the pictorial representation of graphs by considering only important features is somewhat similar to the pictorial representation of graphs by considering all features. Then, no need to consider all features same work can be done with fewer features.

Fig. 6 demonstrates the learning curves of all six classification algorithms by considering the remaining features for the identification of PCOS. It can be visualized from Fig. 6, that we get bad learning curves for all six algorithms because the remaining features and characteristics are unimportant for PCOS diagnosis.

**Step 5:** Below Table 5, represents the analysis of learning curve graphs for AF means All features, IF means Important features and RF means Remaining features. This table shows the performance of all algorithms using over fit, best fit and under fit approaches. Learning curve graphs for all classification algorithms depicts training and cross-validation scores shown in Figs. 4-6. Here, most of the algorithms show underfit problems for training score and the cross-validation score below the table represents the performance of all algorithms. It is clearly shown that random forest, gradient boosting and K-nearest neighbor give the best fit for all features and important features respectively.

**Step 6:** To consolidate the results of all classification algorithms with important features, all features and remaining features in Table 6 have been prepared. All results are given in the form of percentages. The

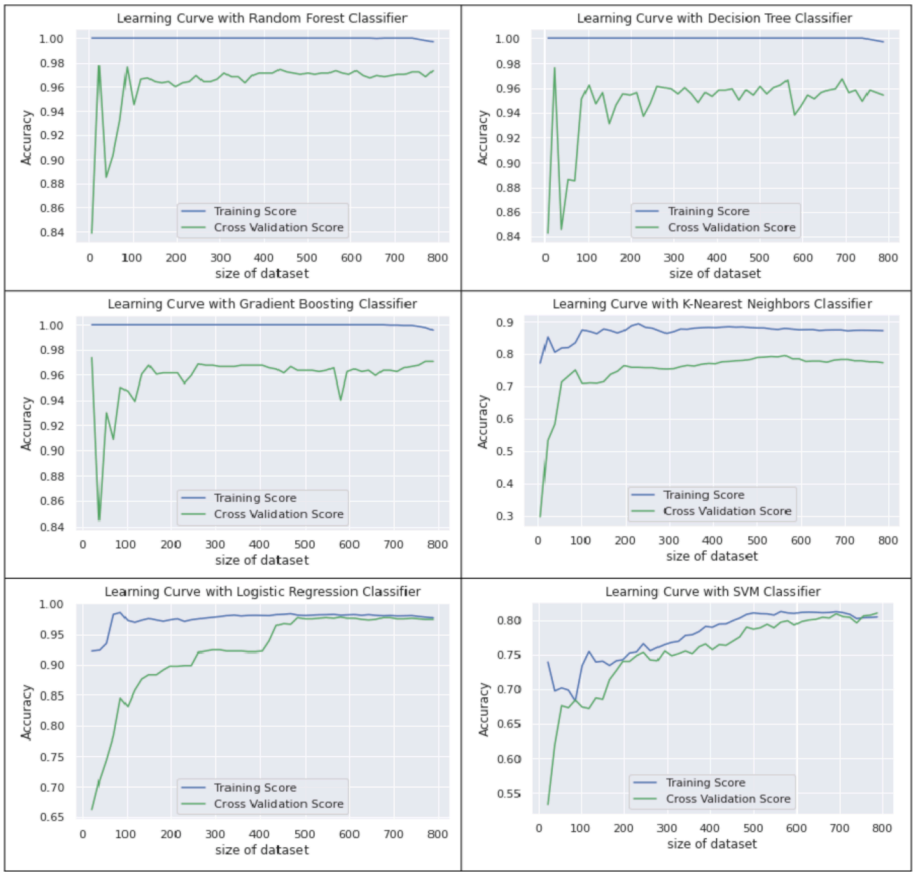



**Fig. 5.** Learning curve graphs for important features.

below table validates that only important features are required for PCOS determination because it provides better accuracy, precision, recall and f1-score as compared to all features and remaining features respectively.

**Step 7:** Below Table 7, represents the accuracy of stacking models for important features. This table shows that stacking with KNN, LR and SVM attain higher accuracy compared with KNN, LR and SVM.

### *5.2. Unsupervised learning clustering algorithm results:*

To validate the above analysis on a realistic dataset, unsupervised learning algorithms are applied. K means the clustering algorithm is applied over the important features of unlabelled data to determine PCOS. K-means clustering is a kind of unsupervised learning, which is used for unlabelled data (i.e., data without defined categories or groups). This algorithm aims to observe correlation within data (grouping of data types), where several groups depict the variable K. Step by step process of K-mean clustering has been applied and explained as follows:

**Step 1:** One of the most prominent approaches for determining the ideal value of K is the Elbow Method. Fig. 7 shows the elbow method graph for K-mean clustering where the x-axis represents the number of clusters and the y-axis represents the distortion. According to the result of Fig. 7, K should be 2,3 or 4.

**Step 2:** For normalizing the dataset, the Gaussian distribution function is used(Normalization, Standardization and Normal Distribution, 2021). A Scatter plot can be drawn using a normalized dataset. Fig. 8 shows the scatter plots by considering all features, important features and remaining features. Fig. 8(a and b) depict that the data points clusters are divided into two groups whereas in Fig. 8(c) data points clusters are overlapped with each other.

**Step 3:** Table 8 represents the results of the K- means clustering algorithm by considering all features, important features and remaining features. Silhouette score and Davies Bouldin score are used to measure the quality of clustering results. In general, the range of silhouette score lies between 0 and 1 i.e., if the output of the model is near 1 it means the quality of clusters is good. Generally, the range of davies bouldin score lies between 0 and 1 i.e., if the output of the model is near 0 it means the quality of clusters is good.

The above Table 8, it is shown that the value of the silhouette score and davies bouldin score are near 1 and 0 respectively by considering all important features. These values indicate good clusters whereas considering only the remaining features gives scatter clusters for both the evaluation metrics. So, after analysis of the unlabelled dataset. Again, it has been found that all four features are important for determining PCOS. These four features i.e., fbs, BMI, glucose and trestbps represent four diseases diabetes, high blood pressure, heart disease and obesity respectively.

## 6. Conclusion

In this article, the primary goal is to identify related diseases that can aid in the early detection of PCOS. To date, no work in this direction can identify PCOS with other diseases. At present, it is shocking to know that in all PCOS-affected women there are 50 % higher chances of having obesity and diabetes. Along with that, these women are highly prone to heart diseases. To achieve the above-mentioned goal, feature selection techniques are applied over the amalgamated data to reduce the number of parameters used for PCOS diagnosis. After that supervised learning algorithms are used to find the performance metrics of the model for all features, important features and remaining features. It shows that important features provide better accuracies of the algorithms compared to all features and remaining features such as random forest(98.5 %),

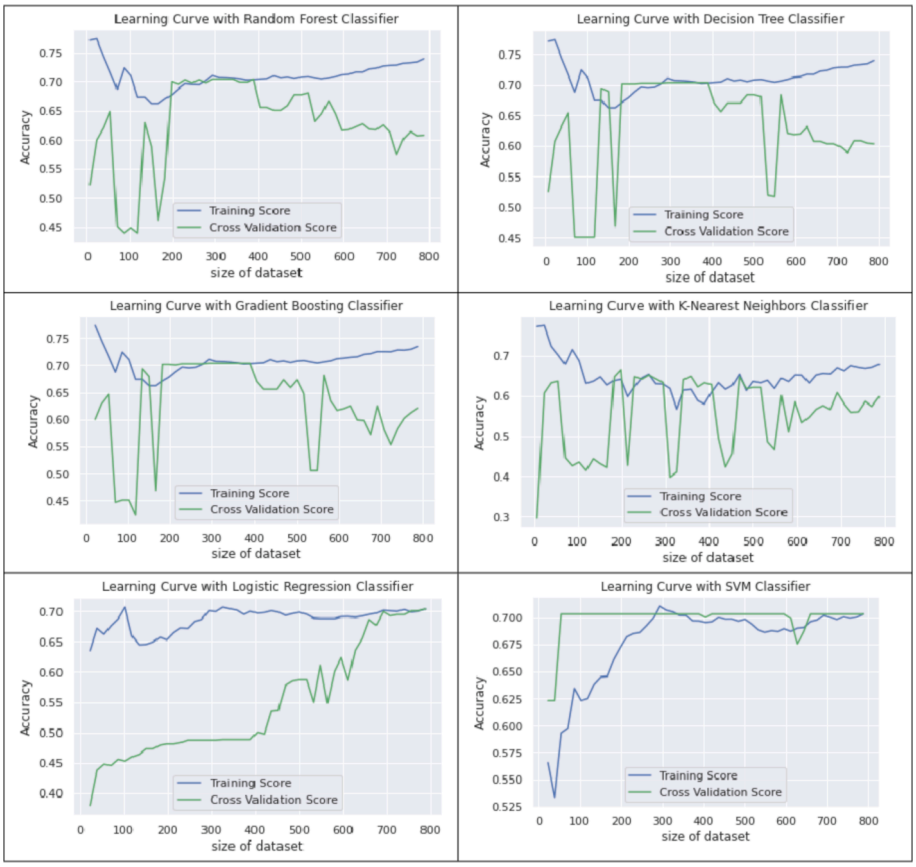



**Fig. 6.** Learning curve graphs for remaining features.

**Table 5**
Performance analysis using over fit, best fit and under fit for all the classification algorithms.

| Classification Algorithms | AF | IF | RF |
|---|---|---|---|
| Random Forest | Best Fit | Best Fit | Over Fit |
| Decision Tree | Over Fit | Over Fit | Over Fit |
| Gradient Boosting | Best Fit | Best Fit | Over Fit |
| K- Nearest Neighbors | Best Fit | Best Fit | Over Fit |
| Logistic Regression | Over Fit | Best Fit | Over Fit |
| Support Vector Machine | Over Fit | Over Fit | Over Fit |

decision tree(98.5 %), gradient boosting(98.9 %), K-nearest neighbor (78.2 %), logistic regression(98.5 %), support vector machine(98.5 %) and hybrid RFLR(98.5 %). It indicates that gradient boosting gives the best accuracy among all these algorithms. To validate the outcomes of supervised learning algorithms, the unsupervised learning K-means clustering algorithm is applied to the unlabelled amalgamated data for all features, important features and remaining features. It provides the best silhouette score (94.4 %) and davies bouldin score (7.8 %) for important features. These approaches justify the relevance of important features i.e., fbs, trestbps, BMI and Glucose for early identification of PCOS which represent diabetes, high blood pressure, obesity and heart diseases respectively. This will reduce the huge financial burden as well as save the time of PCOS patients by reducing the number of clinical tests required for diagnosis. As a future scope, the above-mentioned diseases and their parameters can be used to create a mobile application that can help in the early identification of PCOS.

**Table 7**
Comparative analysis of classification algorithms with stacking classification models.

| Classification Algorithms | Accuracy |
|---|---|
| Stacking With RF | 97.8 % |
| Stacking With DT | 96.3 % |
| Stacking With GB | 97.1 % |
| Stacking With KNN | **97.9 %** |
| Stacking With LR | **98.7 %** |
| Stacking With SVM | **98.7 %** |

**Table 6**
Performance metrics analysis table of all features, important features and remaining features.

| Classification Algorithms | Accuracy | | | Precision | | | Recall/Sensitivity | | | F1-score | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AF | IF | RF | AF | IF | RF | AF | IF | RF | AF | IF | RF |
| Random Forest | 98.5 % | 98.5 % | 68.0 % | 98.4 % | 98.3 % | 43.8 % | 96.7 % | 96.7 % | 11.5 % | 97.5 % | 97.5 % | 18.2 % |
| Decision Tree | 98.5 % | 98.5 % | 68.5 % | 96.8 % | 98.3 % | **44.4 %** | 98.4 % | 96.7 % | **65.4 %** | 97.5 % | 97.5 % | 11.4 % |
| Gradient Boosting | **98.9 %** | **98.9 %** | 68.5 % | **98.4 %** | 98.4 % | 44.4 % | **98.4 %** | **98.4 %** | 6.6 % | **98.4 %** | **98.4 %** | 11.4 % |
| K- Nearest Neighbors | 73.6 % | 78.2 % | 64.9 % | 61.5 % | 71.4 % | 36.7 % | 39.4 % | 49.2 % | 18.0 % | 48.0 % | 58.3 % | **24.2 %** |
| Logistic Regression | 98.5 % | 98.5 % | 68.5 % | 96.8 % | 96.8 % | 0 % | 98.4 % | 98.4 % | 0 % | 97.6 % | 97.6 % | 0 % |
| Support Vector Machine | 98.5 % | 98.5 % | **69.0 %** | 96.8 % | 96.8 % | nan | 98.3 % | 98.3 % | 0 % | 97.6 % | 97.6 % | 0 % |
| Hybrid RFLR | 81.2 % | 98.5 % | 69.0 % | 83.3 % | **98.9 %** | nan | 49.2 % | 95.1 % | 0 % | 61.8 % | 97.5 % | 0 % |

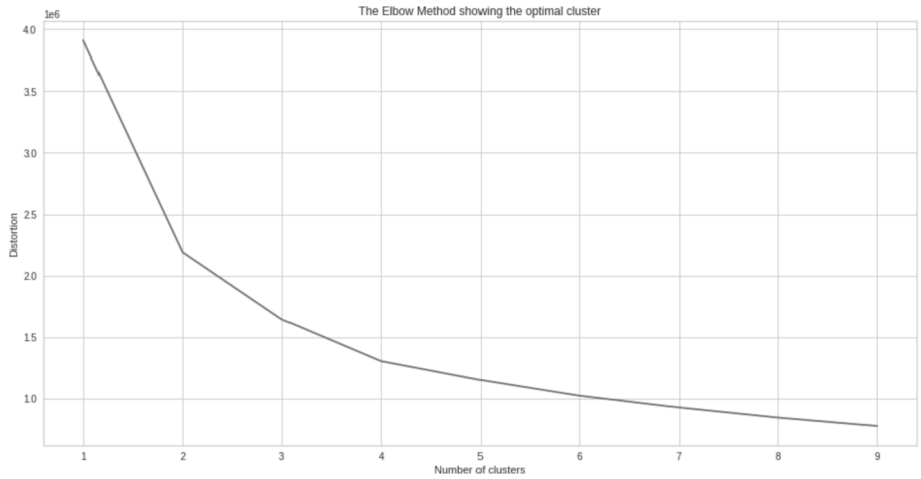



**Fig. 7.** Elbow method graph for K-mean clustering.

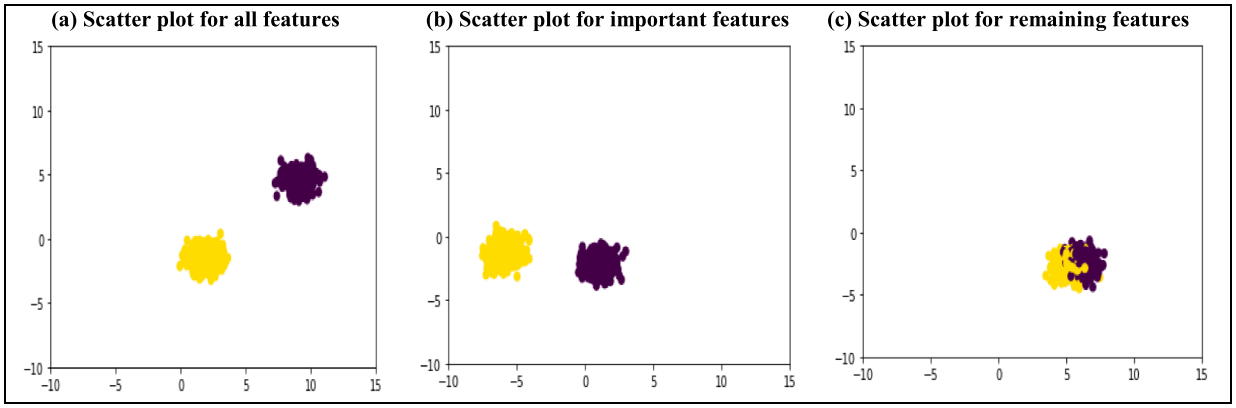



**Fig. 8.** Scatter plots for all features, important features and remaining features.

**Table 8**
Performance analysis of the K-Mean Clustering algorithm with all features, important features and remaining features.

| K-Mean Clustering | Silhouette Score | Davies Bouldin Score |
|---|---|---|
| All features | 90.2 % | 13.9 % |
| Important features | **94.4 %** | **7.8 %** |
| Remaining features | 55.7 % | 65.4 % |

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

I have shared the link in the manuscript on page number 8.

## Acknowledgement

The authors like to extend special thanks to Dr. Gaurav Verma, Department of Electronics and Communication Engineering and Dr. Archana Purwar, Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, for their constant guidance and support in the accomplishment of this work.

## References

Aggarwal, S., & Pandey, K. (2021). An Analysis of PCOS Disease Prediction Model Using Machine Learning Classification Algorithms. *Recent Patent of Engineering, 15*(6), 53–63. https://doi.org/10.2174/1872212115999201224130204

Aggarwal, S., & Pandey, K. (2022). Determining the representative features of polycystic ovary syndrome via Design of Experiments. *Multimedia Tools and Applications, 81*, 29207–29227. https://doi.org/10.1007/s11042-022-12913-0

Ali, S. E., & Ali, F. E. (2020). A Study of Apelin-36 and GST Levels with Their Relationship to Lipid and Other Biochemical Parameters in the Prediction of Heart Diseases in PCOS Women Patients. *Baghdad Science Journal, 17*(3), 924–930. https://doi.org/10.21123/bsj.2020.17.3(Suppl.).0924.

Anagnostis, P., Tarlatzis, B. C., & Kauffman, R. P. (2018). Polycystic ovarian syndrome (PCOS): Long-term metabolic consequences. *Metabolism: Clinical and Experimental, 86*, 33–43. https://doi.org/10.1016/j.metabol.2017.09.016

Bloice, M. D., & Holzinger, A. (2016). A tutorial on machine learning and data science tools with python. *In Machine Learning for Health Informatics*. https://doi.org/10.1007/978-3-319-50478-0_22

*Causes of Sleep Apnea*. (2021). WebMD. https://www.webmd.com/sleep-disorders/sleep-apnea/obstructive-sleep-apnea-causes.

*Centers for Disease Control and Prevention. (2020). PCOS (Polycystic Ovary Syndrome) and Diabetes*. (n.d.). https://www.cdc.gov/diabetes/basics/pcos.html. Accessed February, 2022.

Chen, W., & Pang, Y. (2021). Metabolic Syndrome and PCOS: Pathogenesis and the Role of Metabolites. *Metabolites, 11*(12). https://doi.org/10.3390/metabo11120869

Condorelli, R. A., Calogero, A. E., Mauro, M. D., & La, S. (2017). PCOS and diabetes mellitus : From insulin resistance to altered beta-pancreatic function, a link in evolution. *Gynecological Endocrinology, 33*(9), 665–667. https://doi.org/10.1080/09513590.2017.1342240

Doroszewska, K., Milewicz, T., Mrozińska, S., Janeczko, J., Rokicki, R., Janeczko, M., et al. (2019). Blood pressure in postmenopausal women with a history of polycystic ovary syndrome. *Przegląd Menopauzalny=. Menopause Review, 18*(2), 94–98. https://doi.org/10.5114/pm.2019.84039

El Hayek, S., Bitar, L., Hamdar, L. H., Mirza, F. G., & Daoud, G. (2016). Poly Cystic Ovarian Syndrome: An updated overview. *Frontiers in Physiology, 7*(APR), 1–15. https://doi.org/10.3389/fphys.2016.00124

Escobar-Morreale, H. F. (2018). Polycystic ovary syndrome: Definition, aetiology, diagnosis and treatment. *Nature Reviews Endocrinology, 14*(5), 270–284. https://doi.org/10.1038/nrendo.2018.24

Osisanwo, F. Y., O Awodele, J. E. A., et al. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology, 48*(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

Wang, F. F., Wu, Y. Y. H. Z., et al. (2018). Pharmacologic therapy to induce weight loss in women who have obesity/overweight with polycystic ovary syndrome : A systematic review and network. *Obesity Reviews, 19*(10), 1424–1445. https://doi.org/10.1111/obr.12720

Fauser, B. C. J. M., Van Rijn, B. B., Bekker, M. N., & De Wilde, M. A. (2019). Associations of preconception Body Mass Index in women with PCOS and BMI and blood pressure of their offspring Associations of preconception Body Mass Index in women with PCOS and BMI and blood pressure of their offspring. *Gynecological Endocrinology, 35*(8), 673–678. https://doi.org/10.1080/09513590.2018.1563885

Glueck, C. J., & Goldenberg, N. (2019). CHARACTERISTICS OF OBESITY IN POLYCYSTIC OVARY. *Metabolism, 92*, 108–120. https://doi.org/10.1016/j.metabol.2018.1100%02

*Heart Disease Dataset*. (2019). Kaggle. https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

Kakoly, N. S., Khomami, M. B., Joham, A. E., Cooray, S. D., Misso, M. L., Norman, R. J., et al. (2018). Ethnicity, obesity and the prevalence of impaired glucose tolerance and type 2 diabetes in PCOS : A systematic review and meta-regression. *Human Reproduction Update, 24*(4), 455–467. https://doi.org/10.1093/humupd/dmy007

Kyrou, I., Karteris, E., Robbins, T., Chatha, K., Drenos, F., & Randeva, H. S. (2020). Polycystic ovary syndrome (PCOS) and COVID-19: An overlooked female patient population at potentially higher risk during the COVID-19 pandemic. *BMC Medicine, 18*(1), 1–10. https://doi.org/10.1186/s12916-020-01697-5

Lauritsen, M. P., Svendsen, P. F., & Andersen, A. N. (2019). Diagnostic criteria for polycystic ovary syndrome. *Ugeskrift for Laeger, 181*(15), 671–679. https://doi.org/10.1016/S1701-2163(16)32915-2.Diagnostic

Lie, S., Douma, A., & Verhaeghe, J. (2020). Implementing the international evidence-based guideline of assessment and management of polycystic ovary syndrome (PCOS): How to achieve weight loss in overweight and obese women with PCOS ? *Journal of Gynecology Obstetrics and Human Reproduction, 50*(6), 1–8. https://doi.org/10.1016/j.jogoh.2020.101894

Marchesan, L. B., & Spritzer, P. M. (2019). ACC/AHA 2017 defination of high blood pressure : Implications for women with polycystic ovary syndrome. *Fertility and Sterility, 111*(3), 579–587. https://doi.org/10.1016/j.fertnstert.2018.1100%34

Mellembakken, J. R., Mahmoudan, A., Mørkrid, L., & Sundström-poromaa, I. (2021). Higher blood pressure in normal weight women with PCOS compared to controls. *Endocrine Connections, 10*(2), 154–163. https://doi.org/10.1530/EC-20-0527%0A

*Normalization, Standardization and Normal Distribution*. (2021). Towards Data Science. https://towardsdatascience.com/normalization-standardization-and-normal-distribution-bfbe14e12df0.

Oberg, E., Jakson, I., Mitsell, M., Egnell, P. T., Institutet, K., Medicine, R., et al. (2019). Improved Menstrual Function in Obese Women with Polycystic Ovary Syndrome after Behavioral Modification Intervention - a Randomized Controlled Trial. *Clinical Endocrinology, 90*(3), 468–478. https://doi.org/10.1111/cen.13919

Özkan, S., Yılmaz, Ö.Ç., & Yavuz, B. (2020). Increased masked hypertension prevalence in patients with polycystic ovary syndrome (PCOS). *Clinical and Experimental Hypertension, 42*(8), 681–684. https://doi.org/10.1080/10641963.2020.1772815

*Pima Indians Diabetes Database*. (n.d.). UCI MACHINE LEARNING. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

Saravana, N. M., Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive Methodology for Diabetic Data Analysis in Big Data. *Procedia - Procedia Computer Science, 50*, 203–208. https://doi.org/10.1016/j.procs.2015.04.069

Torchen, L. C. (2017). Cardiometabolic Risk in PCOS : More than a Reproductive Disorder. *Current Diabetes Reports, 17*(12), 137. https://doi.org/10.1007/s11892-017-0956-2

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory, 55*, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003

Wekker, V., Dammen, L. V., Koning, A., Heida, K. Y., Painter, R. C., Limpens, J., et al. (2020). *Long-term cardiometabolic disease risk in women with PCOS : a systematic review and meta-analysis., 26*(6), 942–960. https://doi.org/10.1093/humupd/dmaa029

Wilson, M. S., & Metink-Kane, M. M. (2012). Polycystic Ovary Syndrome and Risk for Long-Term Diabetes and Dyslipidemia. *Obstet Gynecol, 23*(1), 6–13. https://doi.org/10.1097/AOG.0b013e31820209bb.Polycystic

Witchel, S. F., Oberfield, S. E., & Peña, A. S. (2019). Polycystic Ovary Syndrome: Pathophysiology, Presentation, and Treatment with Emphasis on Adolescent Girls. *Journal of the Endocrine Society, 3*(8), 1545–1573. https://doi.org/10.1210/js.2019-00078

Zhang, X., Wu, G., Dong, Z., & Crawford, C. (2015). Embedded feature-selection support vector machine for driving pattern recognition. *Journal of the Franklin Institute, 352*(2), 669–685. https://doi.org/10.1016/j.jfranklin.2014.04.021

Zhao, X., Feng, X., Zhao, X., Jiang, Y., Li, X., & Niu, J. (2021). How to Screen and Prevent Metabolic Syndrome in Patients of PCOS Early : Implications From Metabolomics. *Frontiers in Endocrinology, 12*, 626. https://doi.org/10.3389/fendo.2021.659268

Zhu, T., Cui, J., & Goodarzi, M. O. (2021). Polycystic ovary syndrome and risk of type 2 diabetes, coronary heart disease, and stroke. *Diabetes, 70*(2), 627–637. https://doi.org/10.2337/db20-0800