An Efficient Way of Detecting PCOS Using Machine Learning

Prashant Richhariya, Madhuri Nigam, Bharti Bhattad, Anita Soni and Pankaj Richhariya

April 2, 2022

An efficient way of Detecting PCOS using Machine Learning

***Abstract*— An** endocrine disease that occurs in women of reproductive age is Polycystic Ovary Syndrome or PCOS. This is not necessary to reverse the disease once diagnosed, but medication may help alleviate the symptoms. The exact cause of PCOS is indeed unclear, but the probability of having PCOS is illustrated by some factors. Obesity, addiction to insulin, blood pressure, depression, infection are the causes that occur in this condition. Hirsutism, oligo-ovulation, acne, excessive bruising, and discoloration of the skin are the symptoms. A system is ready to accept them as characteristics and outputs of the inclusion or exclusion of this disorder using the causes and symptoms. K-Nearest Neighbor and Logistic Regression are the machine learning models used for supervised classification. The purpose behind the creation of multiple systems is to discover the perfect one of these in the known information spectrum for the given dataset.

1.1 INTRODUCTION

Infertility is a disease which develops throughout childbearing in women [1]. Female healthcare organs are influenced by ovaries that contain testosterone and estrogen-hormones that control the menstrual cycle. Ovaries often contain small quantities of hormone levels called androgens. PCOS's basic features are:

- Ovarian cysts.
- Elevated hormone levels: androgen.
- Unpredictable times
- Abnormal growth of body hair

It has a series of signs that signify its existence, as the disorder is a syndrome. In identifying this disorder, these signs play a crucial role. Causes that can contribute to the risks of the condition should also be recognized along with these signs. PCOS can be diagnosed as quick as possible since the possibility of miscarriage, diabetes, endometrial cancer and cardiovascular disease is present at the higher level of the disease [2]. Here, to assess the existence of PCOS, a few machine learning models are constructed. Since the dataset classifies whether or not the situation is present, supervised algorithms are used for machine learning called: K-Nearest Neighbor (K-NN) and Logistic Regression. The first is a methodology depending on distance and the other is based on probability, which is why these two methods, which are separated by poles, are used and contrasted with their precision [3]. With 92 percent precision, logistic regression is more precise, while K-NNs are 90.74 percent. Polycystic Ovary Syndrome (PCOS) is an endocrine and metabolic abnormality that affects 5-10% women of reproductive age [13]. It is a condition that is comprised of heterogenous symptoms making it difficult to establish consistent diagnostic criteria based on a single biochemical/clinical assessment. Some of the symptoms exhibited by women with this condition are menstrual irregularity, anovulation or oligoovulation, obesity, infertility, and hyperandrogenism that presents as acne, male pattern baldness, and male pattern hair growth.

The diagnostic criteria for PCOS were originally established by two different bodies; the ESHRE (European Society of Human Reproduction and Embryology), and the ASRM (American Society of Reproductive Medicine). The ESHRE criteria required the presence of polycystic ovaries on an ultrasonographic examination in addition to symptoms of hyperandrogenism and anovulation/oligoovulation. It also required the absence of any pituitary or adrenal disease [14]. The ASRM criteria required clinical signs of hyperandrogenism and ovulatory abnormality in the absence of adrenal hyperplasia (impairment of a key enzyme in the production of cortisol and aldosterone by the adrenal gland). It did not require ultrasonographic examination of the ovaries. This disparity has now been resolved and the diagnostic criteria of PCOS have been jointly redefined by ASRM and ESHRE in an international consensus meeting as the presence of at least two of the following three criteria:

1) Oligo or anovulation;

2) biochemical and/or clinical signs of hyperandrogenism; and,

3) the company of at minimum one polycystic pod

Example of normal and polycystic ovaries trasonographic examination. A polycystic ovary (PCO) is one that is characterized by the presence of 12 or more follicles measuring around 2-9 mm in diameter and/or measured ovarian volume of > 10 cm3 [1]. Analysis of ultrasonographic images for detecting follicle morphology is an important diagnostic marker in the refined definition of PCOS as per the international consensus criteria. Hence, ultrasonographic imaging of the ovaries is included in routine checkups for menstrual abnormalities, infertility treatments, and/or hyperandrogenic symptoms. The ultrasonographic images of a normal ovary and a polycystic ovary. In a normal ovary the follicle count is lesser than in a polycystic ovary, and the follicles exhibit a random distribution within the ovary. Alternatively, polycystic ovaries exhibit more smaller, possibly irregularly shaped follicles, and, in most of the cases, a peripheral distribution of follicles. A decision as to whether a patient has PCO or not is made by counting the number of follicles, determining the size (diameter) of each follicle and/or calculating the ovarian volume. This is a subjective process, and is highly dependent on the experience and proficiency of the ultra-sonographer. Even though expert decisions are reliable, an automation of the process would minimize the subjectivity of the analysis, the interobserver variability that may arise, the need to convert between national/international standardizations, and the fatigue that could happen when reading numerous images every day. It would also increase the number of images that can be analyzed per day thereby increasing the number of patients that can be handled on a day-to-day basis as well as improving the accuracy with which the images can be interpreted. It could provide the patient with a quick response time allowing them to seek medical advice/treatments quickly that may abate or obviate the severe consequences of the disease. The healthcare system could also benefit by this quick response time. This not only off-loads the strenuous work from the imaging specialist, but could also prove to be economically beneficial to the Heath Care systems of federal and provincial Governments.

1.2 Objectives

The hypothesis that PCO morphology can be detected effectively (> 90%) in ultrasonographic images using an automated image analysis technique. Our work aims to automate the analysis of pelvic ultrasonographic images for detecting PCO morphology by examining the number, size, and distribution of follicles within the ovary. Automation of the analysis of pelvic ultrasonographic images was achieved using a three-step process that involved:

- Segmentation of follicles from the ultrasonographic images using image processing methods;
- application of a mathematical methodology called stereology to quantify the attributes of the segmented follicles and store them as feature vectors; and,
- classifying the feature vector obtained from the previous step into one of the two categories: PCO morphology present or normal.

These are the three key steps that form the basis of the automatic PCO morphology detection system. We faced some challenges to accomplish this automatic detection as ultrasonographic images are noisier than images obtained from other modalities due to the physics of acoustic imaging. Also, in ultrasonographic images the borders of follicle regions may not be well defined due to artifacts such as reverberation and acoustic shadowing/enhancement, which makes it difficult for a follicle segmentation algorithm to interpret the exact border. For this reason, the regular thresholding techniques or the edge detection methods that perform segmentation do not give acceptable.

Analysis of normal and PCOS ovaries the other important application of processing ultrasonographic images is to distinguish between normal and abnormal ovaries. That is, ovarian ultrasound is also used for detecting abnormalities such as ovarian cancer, and ovarian cysts (which is an indication of ovarian tumor), although the image speckle of ultrasonographic images makes it difficult to process the image for such abnormalities. The texture-based segmentation of Jiang and Chen [3] uses a texture-based pixel classifier based on four texture energy measures to distinguish between normal and abnormal ovaries for the detection of ovarian cancer. Zimmer et al. [4] have presented a semiautomatic way for quantification of ovarian cysts.

1.2.1 Texture-based pixel Classifier: - Texture-based k-means cluster analysis [3] does not use the standard intensity-direction edge information for segmentation, because tissue edges are blurred in ultrasonographic images and is also indistinct because of speckle noise. This texture-based pixel classifier reevaluates the original Laws' feature masks and derives new texture features to be applied to the k-means clustering as it has been found that the original feature masks does not give satisfactory results on ultrasonographic images.

The original Feature masks [3]. image is convolved with the four feature masks to give the feature image g(i, j), from which the mean and deviation around each pixel is calculated as follows:

$$s\ [i, j] = 1\ (2n+1)\ 2\ i+n \sum k=i-n\ i+n \sum l=i-n\ |g\ [k,l]-mean$$

where, $mean = 1\ MN \sum i\ j\ g\ [i, j]$, and s[i, j] forms the four-dimensional feature space for the k-means clustering process that divides each pixel in the image. This is done by selecting three seed points from three feature vectors that are centers of three regions in feature space. It takes the current seed points, classifies them based on the minimum Euclidean distance between the seed points and the feature vectors. Once a seed point belongs to a particular class, the seed points are recalculated as the mean of the pixels in the class. This process is repeated until the shift in the means becomes less than a preset value. This method achieves a good segmentation of normal and abnormal ovarian tissues.

A septation is long and divides the cyst into separate regions (multilocular). The algorithm of Zimmer et al. does a morphological classification of these structures to determine the malignancy of the ovarian pathology, using scoring systems and a minimum error rate Bayesian classifier. That is, the quantified properties of these cystic structures are scored according to a predetermined table and the resulting values are used for classification. This method starts by separating the cysts from the background as the cysts appear as dark regions in ultrasound B-scans. Next, convex hulls of the cysts are determined to calculate the convex deficiency of the cyst. Convex deficiency is obtained by subtracting the original cyst from its convex hull. This value gives an indication of the upper limit of the area of the cystic structures. Then a new image is formed by applying a morphological closing using a binary disk on the original cyst. This results in closing the convex deficiency indicative of the cystic structures, but which leaves a crater in the outer part of the concavities. Binary disks of different radii are applied and the minimal radius that covers at least 50% of the convex deficiency is chosen. If the disks fill only a small part of the convex deficiency leading to the underestimation of the cystic structures, then the convex deficiency is used for describing the cystic structures, instead of the result of the morphological closing. There can be an underestimation of the cystic structures when morphological closing. This can be corrected by taking the regions for which the morphological closing was applied and replacing them with their convex hulls the initial convex deficiency, in this terminology, the original cyst is white and the convex deficiency is gray.

(a) Initial convex deficiency obtained by subtracting the original cyst from the convex hull

(b) Convex deficiency after morphological closing

(c) Corrected convex deficiency after replacing with the convex hull [4].

The shape features for the classification were chosen to be the area, roundness of the shape (compactness), and a value specifying what percentage of the cystic structure touches the cyst boundary. The last measure is called the "pop" and it is given as: pop = 100 $*$ length of portion touching cyst total length of perimeter.

This value is an indication of the kind of structure present inside the cyst. For example, a value of pop between 75%-95% represents that it is a septation, a value between 60%-80% means it is a papillations, and a value of 50%-60% indicates it is a side structure meaning that it would not be used in the quantification. Also, structures that have an area less than 1% of the area of the cyst are categorized as false structures and are removed. The available data for a specific shape are described as a feature vector in a two-dimensional feature space. Then this shape belongs to class k, if it gives a minimal value for the 30 equations over all classes:

$$F_k = (x - \mu k)\ t - 1 \sum k\ (x - \mu_k) + \ln (|\sum k|) \text{ where, } k = 1, 2, 3,$$

μ_k is the mean vector of class k, and $\sum k$ is the covariance matrix of class k.

The mean vector and the covariance matrix of class k are found using a training sample set. Then the classification technique was applied to a test group of unidentified shapes using the obtained values. Once the structures are identified using the Bayes classifier, the quantitative data is extracted using morphological erosions (Appendix 2 of [4]). The final classifications were compared with that of a human expert in the field. This technique achieved correct classification rates of 72.3% for the papillations (34 out of 47 cases), 75% for the septations (12 out of 16 cases), and 90.2% for the side structures (37 out of 41 cases) [4]. Classifiers The tissue characteristics or the features obtained from the ovarian structures are classified using classifiers such as the linear discriminant classifier, k-nearest neighbor classifier (KNN classify in Python), and the Support Vector Machine Classifier (function SVM classify in Python). The two former implementations arepart of Python statistics toolbox, and the latter is from Python.

1.2.2 Linear discriminant. The method of linear discriminant analysis was originally developed by R.A. Fisher in 1936 and is a classic method used for categorical classification. The basic idea of this method is to classify two or more categories(groups) with n variables by projecting the high dimensional data onto a line and performing classification in this one-dimensional space. If for example there are two groups (classes) involved, the projection should maximize the distance between the mean of these two groups (between-class variance), and minimize the within-class variance of each group (variance within each class). LDA: 3-class feature data projected on two rotated axes [11] linear projections w, the following measure J(w) should be maximized as per theFisher criterion [31].

$$J(w) = |m1 - m2| \; 2 \; s \; 2 \; 1 + s \; 2 \; 2$$

where m1 and m2 are the means of class1 and class2, and s1 and s2 are the variance of class1 and class2.

The projection seeks to rotate the axes so that the when classes are projected onto these axes the differences between the classes are maximized. In the projection of the categories to the lower left axes gives the worst separation of the classes, and the projection to the lower right axes gives the best separation of the classes.

1.2.3 KNN K-Nearest Neighbor Classification K-nearest neighbor algorithm is one of the simplest machine learning algorithms. In this algorithm, an unknown feature (test dataset/pattern) is classified as belonging to the class to which the majority of its K closest training neighbors in the feature space belong. First, all the training datasets/patterns represented as feature vectors are positioned on a multidimensional feature space. Then, the distance between the unknown feature (test feature) and the training feature vectors is computed using a distance metric such as the Euclidean distance to choose the K- nearest neighbors of the unknown feature. If the value of K is 32 KNN: Distribution of the training vectors on a feature space chosen as one, it is called the nearest-neighbor algorithm. The optimal value of K can be determined using cross-validation technique. To eliminate situations of tie, K can be chosen as an odd number. Once the K-nearest neighbors are found, the class of each of those neighbors is determined and the unknown feature is assigned to the class to which majority of its K neighbors belong. This algorithm might become computationally expensive if the training dataset is large as the

Euclidean distance must be computed for every vector in the feature space. The performance of this algorithm is also limited when there is more noise in the training dataset as it might introduce bias during classification, especially when K is chosen as a small integer. The distribution of the training vectors on a feature space [32]. The unknown vector (green dot) is classified to the class of red triangles if K = 2, but would be classified to the class of blue squares if K = 3. It would still be classified to the class of blue squares if K = 5 (3 nearest blue squares versus 2 nearest red triangles). This classification was implemented using the kNN f y function of Python.

1.2.4 Support Vector Machine Classification: The SVM classifier tries to build a maximum-margin separating hyperplane that maximizes the distance between two parallel hyperplanes that separate the data. The vectors that lie on the two parallel hyperplanes are called the support vectors. Mostly, all linear classifiers are based on the idea of building a hyperplane to separate the two sets of data and the difference between the SVM classifier and the other linear classifiers is that the SVM classifier tries to maximize the distance between the two parallel hyperplanes to 33 minimizethe generalization error. The dividing hyperplane takes the form

$$w\cdot x - b = 0,$$

where w is normal to the dividing hyperplane and the offset b allows to increase the margin without which the hyperplane would pass through the origin. The two parallel hyperplanes are described as

$$w\cdot x - b = 1$$
$$w\cdot x - b = -1,$$

where 1 or -1 is the constant denoting the class to which the point xi belongs to, where xi is a p-dimensional vector. Using geometry, it is found that the distance between the two parallel hyperplanes is given as 2 |w|. So, to maximize the distance between the two parallel hyperplanes, |w| has to be minimized. Data points can be excluded by ensuring for each i Equation 3.20 is followed. [12] shows the dividing hyperplane, support vectors, and the two parallel hyperplanes.

$$w\cdot x_i - b \geq 1 \text{ or}$$

$$w\cdot x_i - b \leq -1$$

Equation can be rewritten as $c_i (w\cdot x_i - b) \geq 1, 1 \leq i \leq n$

where ci is a content 1 or −1 representing the class of xi, and n is the number of training patterns used to train the classifier. Thus, the primal form is to minimize |w| subject to the constraint. This is a Quadratic programming (QP) optimization problem. For non-linear classification, the kernel trick is applied to the separating or the maximum margin hyperplanes. The kernel trick transforms the non-linear observations to a higher dimensional space where the linear classifier would be subsequently applied. So, this makes the linear classification in the new space a non-linear classification in the original space. Thus, in the resulting algorithm each dot product of the linear classifier is replaced with a non-linear kernel function to make the maximum margin hyperplane fit in the transformed 34 Maximum-margin hyperplane and margins for a SVM trained with samples from two classes. Support vectors lie on lines w.x - b=1 and w.x - b=-1 [12]. feature space. Some of the non-linear kernels used are polynomial, radial basis function, and sigmoid functions. In Python the SVM classifier is first trained using function svm train that accepts as input the rows of training data and a column vector of the class information for each row of training data. Each row in the training data is an observation and each column is a feature. The default setting of the svm train function is the linear kernel or dot product which was also used for our training dataset.

Validation Once a classifier model has been learned from the training patterns, its ability to classify new patterns can be assessed using cross validation techniques. This is accomplished by using only part of the available patterns for training. The remaining “test” data are used to test the performance of the learned model. Common types of cross validation methods are the holdout method, and the k-fold cross validation method [33]. In the holdout (or half-and-half) method, the data set is randomly split into a training set and a testing set. A model is learned from the training set and the validity of the model checked by determining the classification accuracy of the model using the testing set. Model accuracy is dependent on the particular split of the data. The disadvantage of the holdout method can be avoided by using the k-fold cross validation technique. In this method, the data set is divided into k folds, out of which k−1 folds are used as the training set, and the remaining fold is used as the testing set. The holdout method is performed k times, each using a different fold as the testing set thus eliminating the dependence on the division of the data points among the training and the testing sets. The classification accuracy is averaged over the k trials and the variance decreases as k increases.

2. Related Work

Polycystic ovary syndrome (PCOS) is the most common endocrinological problem affecting women with a prevalence estimated at 4-25% depending on the diagnostic criteria used [4]. Patients with PCOS demonstrate a combination of characteristics which may include anovulation, oligo or amenorrhoea, hirsutism, acne, evidence of increased serum androgen levels and morphological changes in the ovary evident on ultrasonography. Diagnostically, the current practice uses criteria agreed in Rotterdam 2003. Approximately 50% of PCOS patients are obese; a much higher prevalence than the general population. There is also a metabolic element to the condition in the form of insulin resistance that may result in long-term morbidity [5]. South Asian refers to those persons who originate from the Indian subcontinent (India, Pakistan, Sri Lanka, Bangladesh, and Nepal). In a community-based study in the United Kingdom (UK), it had found that polycystic ovaries (PCO) were particularly common among women of South Asian origin (52%), compared to the prevalence of PCO observed in a predominately Caucasian population (22%). The South Asian population, in general, also exhibit a higher prevalence of insulin resistance and type 2 diabetes, which may increase long-term morbidity among those with PCOS [6].

Recent research indicated higher insulin concentrations and lower insulin sensitivity in South Asian women with PCOS compared to Caucasian women with PCOS. This research also concluded that South Asians presenting with anovulatory PCOS were significantly younger, had more severe hirsutism and a higher prevalence of acanthosisnigricans than their Caucasian counterparts. Health-related quality of life (HRQoL) is a concept used to describe the physical, social and emotional effects of a disease and its associated treatments. Research has shown a reduction in the HRQoL of women with PCOS compared with healthy controls. Comparisons with other medical conditions and gynecological populations have also yielded unusually low scores on psychological well-being and quality of life for women with PCOS. Overall, there has been a great paucity of research comparing the influence of ethnicity or cultural background on HRQoL in women with PCOS. Two studies have tentatively explored this relationship.

In the UK, 5.7% of the population of England and Wales identify themselves as Asian or Asian-British. In 2001, 4.0% of the population were South Asian, comprising the largest minority ethnic group. Because of this, it is essential to understand better the impact of PCOS on South Asian women to ensure clinical treatments are well aligned to need. The aim of this study is, therefore, to compare the HRQoL of South Asian and Caucasian women with PCOS, given that it is particularly common among women of South Asian origin and they have been shown to have more severe symptoms. It has been found that South Asian women from the Indian subcontinent with PCOS would show overall lower HRQoL than Caucasian women with the condition. Based on available statistics, approximately 15% of Polish couples suffer from infertility. Some authors suggest the percentage is even 18–20%. This value generally is in the range of 10–20% and reported differences depend on the data collection methods in different countries: in Denmark, it is 11%, in France 16.4% and in the UK 17%. In the United States, using the current duration approach, infertility among women 15–44 years old is 15.5%. Female factors such as endometriosis, PCO or other ovulatory, uterine, or fallopian tube irregularities and malefactors such as oligoasthenospermia, asthenospermia, teratospermia, azoospermia, rare oligospermia, and immunological factors contribute to infertility. Idiopathic infertility is a situation in which the clinical evaluation and laboratory tests are regular (within the range), but the couple is not able to conceive naturally [7].

3. LITERATURE SURVEY ON IMAGE PROCESSING TECHNIQUES FOR THE DETECTION OF PCOS

Subhasish Deb [8], the follicles are detected in the ultrasonic pictures of the ovary. PCOS is an endocrine issue affecting ladies of reproductive age. This syndrome has primarily found in ladies whose age is in the middle of 25 and 35. We are proposing techniques for recognizing whether a man is experiencing Polycystic Ovary Syndrome (PCOS) or not. Ultrasound imaging of the follicles gives essential data about the size, number, and method of course of action of follicles, position, and reaction to hormonal incitement. A thresholding function is connected for denoising the picture in the wavelet space. Before the segmentation process, the high picture is preprocessed utilizing contrast enhancement technique. The morphological approach is used for executing contrast enhancement. This is performed with a specific end goal to enhance the clarity and nature of the picture. Fuzzy c-means clustering calculation is connected to the resultant picture. At last, the cysts are detected with the assistanceof clusters.

Yinhui Deng et. al. [9], Polycystic Ovary Syndrome (PCOS) are a female endocrine issue which seriously upsets ladies' wellbeing. The confusion is characterized by a collection of incomplete created follicles in the ovaries. Manual investigation of PCOS conclusion frequently produces mistakes. Along these lines, in recent years numerous researchers have been enthusiastically working in automatic detection of PCOS.

Amsy Denny et.al. [10] portray that Follicles are liquid filled sac found in ladies' regenerative framework. Follicle detection in ultrasound images of the ovary is imperative for fruitfulness treatment. They are regularly identified physically by Gynecologists for malady conclusion and to track follicular improvement. This procedure is typically furious and inclined to blunder. The current computerized strategies for the detection of follicles are loaded with low detection rates because of the nearness of image curios and commotions coming about because of veins, endometrial, and tissues as caught by the ultrasound machine. Further, Multi-Layer Perceptron (MLP) was utilized to arrange the identified articles in light of the extricated surface highlights into follicles and non-follicles. The created calculation yielded an accuracy of 96%, the sensitivity of 99% and a specificity of 93%. Additionally, Follicle Detection Rate (FDR), the False Acceptance Rate (FAR) and False Rejection Rate (FRR) were computed to be 98.94%, 7.00% and 1.00% individually.

Muhammad Sakib Khan Inan et. al. [11], polycystic ovary syndrome (PCOS) continues to be an essential subject for researches in different medical specialties by its clinic manifestations, biologic complexity, specific imaging procedures, and comorbidities. Although there are at least three groups of standard criteria for the definition of PCOS, they combine the mandatory coexistence of Hyperandrogenism signs and ultrasound imaging for polycystic ovaries differently. For adult women two of the three criteria are sufficient, but for teenagers, all three conditions must have complied, and imaging can determine the stages of the disease and disorder management.

Maryruth J et.al. [13], Polycystic Ovary Syndrome (PCOS) are an endocrine variation from the norm that occurred in the female reproductive cycle. This paper composed an application to classify Polycystic Ovary Syndrome because of follicle detection utilizing USG images. The main phase of this classification is preprocessing, which utilizes low pass filter, balance histogram, binarization, and morphological processes to acquire paired follicle images. The following stage is segmentation with edge detection, marking, and cropping the follicle images. The accompanying stage is highlight extraction utilizing Gabor wavelet. The cropped follicle images are categorized into two gatherings of surface highlights:

(1) Mean,

(2) Mean, Entropy, Kurtosis, Skewness, and Variance.

It recognizes the highlights of PCO and nonPCO follicles in light of the element vectors came about because of highlight extraction. Here, three classification scenarios are planned:

(1) Neural Network-Learning Vector Quantization (LVQ) strategy,

(2) KNN - Euclidean distance, and

(3) Support Vector Machine (SVM) RBF Kernel.

The best accuracy picked up from SVM - RBF Kernel on C=40. It demonstrates that dataset A reaches 82.55% while dataset B that got from KNN-Euclidean distance classification on K=5 reach 78.81%.

Yunlong Li et.al. [14], Knowledge about the status of the female reproductive framework is vital for fertility issues and age-related family arranging. The volume of these fertility asks for in our emancipated society is relentlessly increasing. Transvaginal ultrasound imaging of the follicles in the ovary gives important data about the ovarian maturing, i.e., some follicles, size, position, and reaction to hormonal incitement. Manual examination of numerous follicles is arduous and error-inclined. In this paper, another strategy for recognition of follicles in ultrasound images of ovaries is proposed. This completely digital segmentation strategy depends on active contours without edge technique. The proposed technique is tried on ultrasonography images of ovaries. The trial results are compared with inferences drawn by a medical expert and demonstrate the efficacy of the technique.

J. Madhumitha, et. al. [15], Do the ultrasonography criteria for polycystic ovaries bolstered by the 2003 Rotterdam consensus satisfactorily discriminate between the common and polycystic ovary syndrome (PCOS) condition in light of recent advancements in imaging technology and reliable techniques for assessing follicle populaces in PCOS? The levels of intra-and inter-observer reliability when five observers utilized the proposed criteria on 100 ultrasound cases were additionally decided. Ninety-eight ladies were determined to have PCOS by the National Institutes of Health criteria as having both oligo-amenorrhea and hyperandrogenismand 70 sound female volunteers recruited from the all-inclusive community.

Palak Mehrotra et.al. [16], introduced a novel probabilistic structure for automatic follicle quantification in 3D ultrasound information. The proposed system powerfully gauges the size and location of each ovarian follicle by combining the data from both the global and local context. Follicle candidates at detected locations are then portioned by a novel database-guided segmentation strategy. To efficiently search theory in a high dimensional space for multiple object detection, a clustered outer space learning approach is introduced. Broad assessments conducted on 501 volumes containing 8108 follicles demonstrated that our strategy could detect and portion ovarian follicles with high vigor and accuracy. It is additionally much quicker than the current ultrasound manual work process. The proposed technique can streamline the clinical work process and enhance the accuracy of existing follicular estimations.

Senthil kumar Mohan et. al. [17], the ovarian ultrasound imaging is an effective apparatus in infertility treatment. Checking the follicles is especially critical in human reproduction. Periodic estimations of the size and state of follicles more than a few days are the essential means of assessment by physicians. Today observing the follicles is finished by non-automatic means with human interaction. This work can be exceptionally requesting and inaccurate and, in a large portion of the cases, means just extra weight for medical experts. In this paper, another calculation for automatic detection of follicles in ultrasound images of ovaries is proposed. It has a typical object recognition scheme (preprocessing, segmentation, include extraction and classification).

Amy Neuzil et. al. [18], exhibited a novel technique for mechanized classification of the ovaries in computerized ultrasound images is proposed which utilizes the contour let change for pre-processing, active contours without edge for segmentation and fuzzy logic for classification. The test results are compared with inferences drawn by a medical expert and demonstrate the efficacy of the strategy.

Aroni Saha Prapty et. al. [19], Poly Cystic Ovarian Syndrome (PCOS) is a common illness of the endocrine organ and is generally called as Stein-Leventhal syndrome. For the most part around 5 ladies at the reproductive age are affected by this sickness. The genuine cause of the infection is not exactly known, yet the beginning of the ailment is characterized by the excessive secretion of insulin resistance androgen. There are various strategies to analyze this condition. This PCOS diagnostic apparatus would spare time a physician who needs to invest energy in themanual tracing of follicles.

Bedy Purnama et. al. [20], to assess enthusiastic processing in ladies with insulin-safe polycystic ovary syndrome (IR-PCOS) and its relationship to glucose control and the mu-transfer framework. The plan in this paper is Case-control Pilot, and the setting is the Tertiary alluding medical center. The patients are seven ladies with IR-PCOS and five non-insulin-safe controls, matured 21– 40 years, recruited from the overall public. Four months of metformin (1,500 mg/day) in ladies with IR-PCOS. Appraisal of state of mind, metabolic function, and neuronal activation amid an enthusiastic assignment utilizing functional magnetic resonance imaging (fMRI), and mu-transfer receptor accessibility utilizing positive emanation tomography (PET).

The authors in the paper [21], Patients with polycystic ovary syndrome (PCOS; which is often associated with increased cardiovascular hazard factors) may introduce hemodynamic changes in the cardiovascular framework. The point of the present investigation was to check whether harmonic indexes of the blood vessel circulatory strain waveform (BPWs) can be utilized to discriminate between PCOS patients and sound people.

4. Research Methodology

The data pertaining to a dataset has to be filtered over in order to create an effective machine learning model A phase chain. It is important to transform the algorithm into a filtered and noise-less input. It compares and tests the effects of the algorithms,

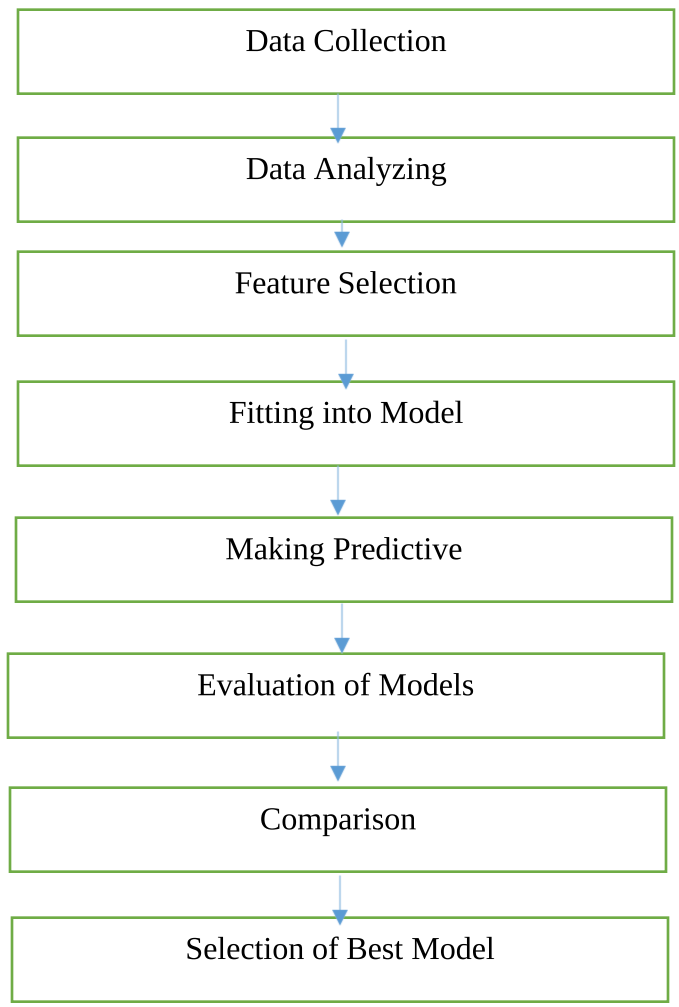

Figure 1: Block diagram of the methodology.

A. Data Collection

Collecting data is the first and critical step. For this reason, diverse channels are accessible. It includes samples from 10 different hospitals in Kerala, India (downloaded from Kaggle). The patients' names just haven't been confirmed.

B. Analyzing Results

It is important to first grasp what the dataset contains in order to go forward. Getting to hear about objects, their properties, and contradictions like: negative values, empty records, unwanted strings, is performed in this step.

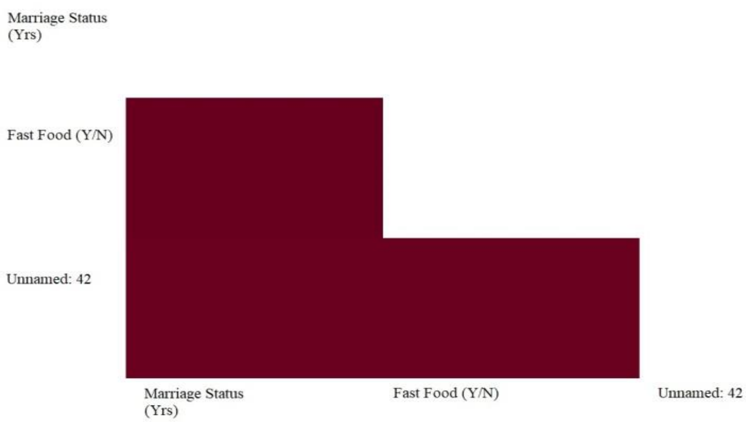

Figure 2: Heat map to localize missing values.

The shape of the dataset is (540, 43) which means that it contains 540 samples and 43 attributes. In figure 1.2 we can see that there are only few records that have missing values. Hence, these records can be dropped. In addition to that, the data type of every value in the dataset should be either float or integer. This is required so that data can be processed by the algorithms.

C. Feature Selection

In order to improve the performance of the model and reduce the computational cost, only selected attributes of the samples act as features. Filter method is used to find the weights of the features in order to determine which of them have high correlation with the target.

Table: 1 displays the features and their correlation to the target, arranged in descending order. The more is the weight of the feature the more is its influence on the target, independently.

Table 1. Feature Weighing

Features Weights	Features Weights
Follicle No. (R)	0.650608
Follicle No. (L)	0.601035
Skin darkening (Y/N)	0.479679
hair growth(Y/N)	0.464623
Weight gain(Y/N)	0.441753
Cycle(R/I)	0.399746
Fast food (Y/N)	0.380246
Pimples(Y/N)	0.28672
AMH(ng/mL)	0.260287
Weight (Kg)	0.206051
BMI	0.195577
Hair loss(Y/N)	0.175055
Hip(inch)	0.156196
Waist(inch)	0.155068
Avg. F size (L) (mm)	0.126586

The features in Table 1 are the top fifteen, amongst the forty-three features. From the above table, we can conclude that the parameter **Follicle No. (R)** and **Follicle No. (L)** have the highest weight that determine the number of follicles in right and left ovaries respectively. Features like **Skin darkening**, **hair growth**, **Weight gain**, **Fast food**, **Pimples**, **Hair loss** contain the values 0 or 1, 0 denoting the absence of that particular feature and 1 denoting its presence. **AMH** or Anti-Müllerian hormone is used as an indicator of egg count. Its unit is nanograms per milliliter. Following, **BMI** refers to Body Mass Index which is the ratio of patient's weight to their height. Along with Hip and Waist sizes in inches, there is also Average Follicle Size of the left ovaries that is measured in millimeters.

D. Fitting into Model

With the data cleaned and selected, it is now ready to be processed by the models. The two models used for supervised machine learning are K-NN and Logistic Regression.

i) K-Nearest Neighbor

KNN is an Instance-Based Learning that compares new instances with instances stored in memory at the time of training of dataset. A new instance is classified by measuring its distances with the instances retrieved from memory, defined in terms of standard Euclidean Geometry, that is, distance between points in n-dimensional space. [4]. the accuracy of this model depends upon two factors: The value of ‘K’ The number of selected features.

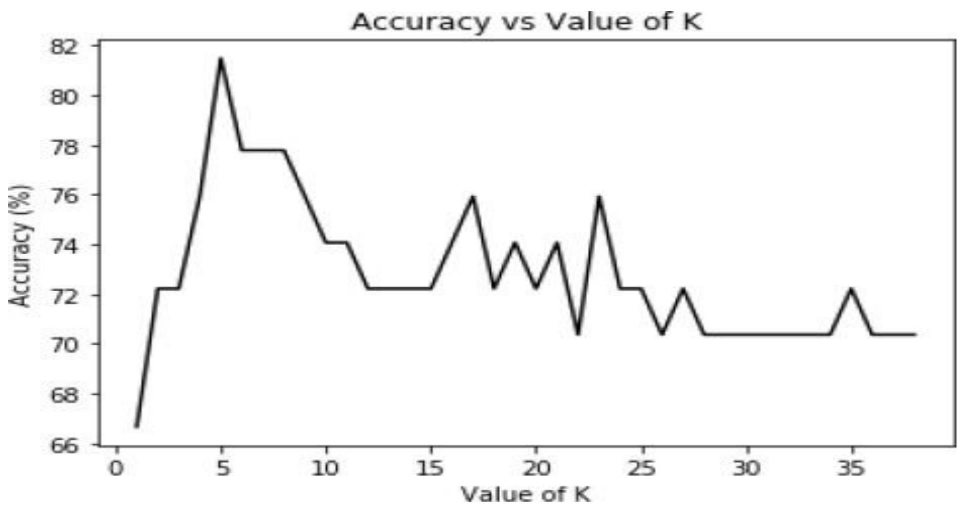

Figure 3 Accuracy Vs Value of K graph

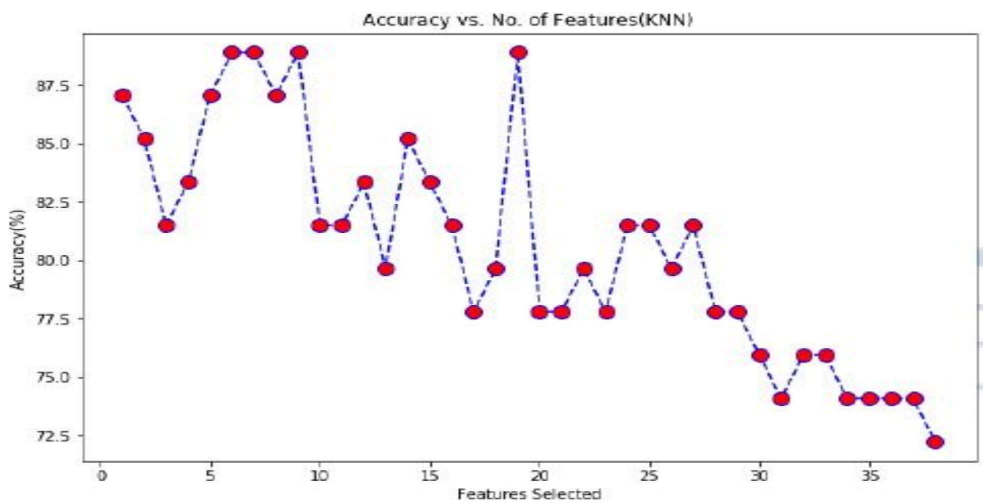

Figure 4 Accuracy Vs. Feature

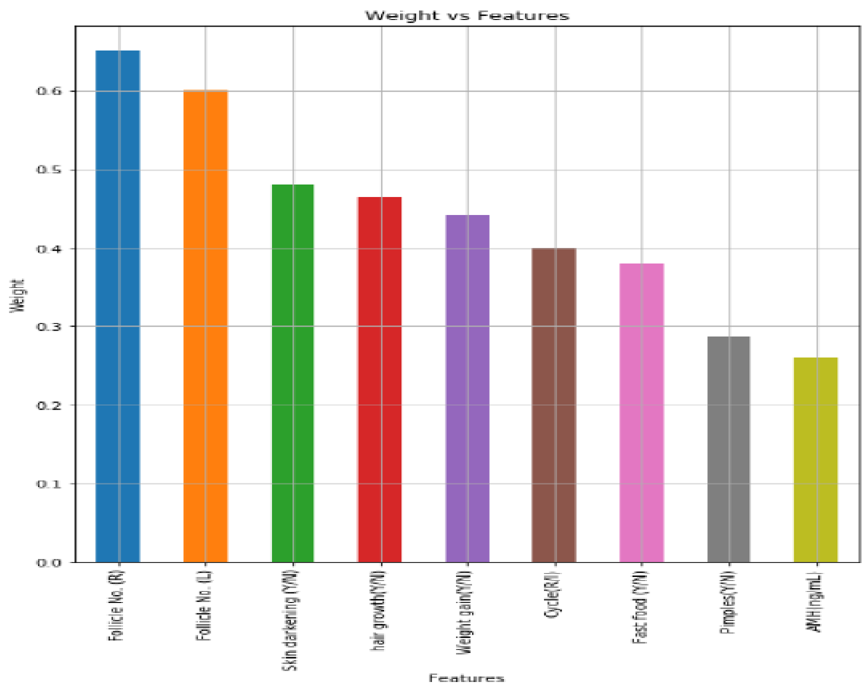

Figure 5 Accuracy vs. No. of Features (KNN) graph

From figure 4.3, it is concluded that highest accuracy occurs when the value of K is 5. It means that when classification needs to be made, the numbers of neighbors whose votes are considered are the five closest ones. From Figure 4.4, we have determined the numbers of features that need to be selected in order to give maximum accuracy. Therefore, the numbers of features included are 9. Following image shows the nine features along with weights.

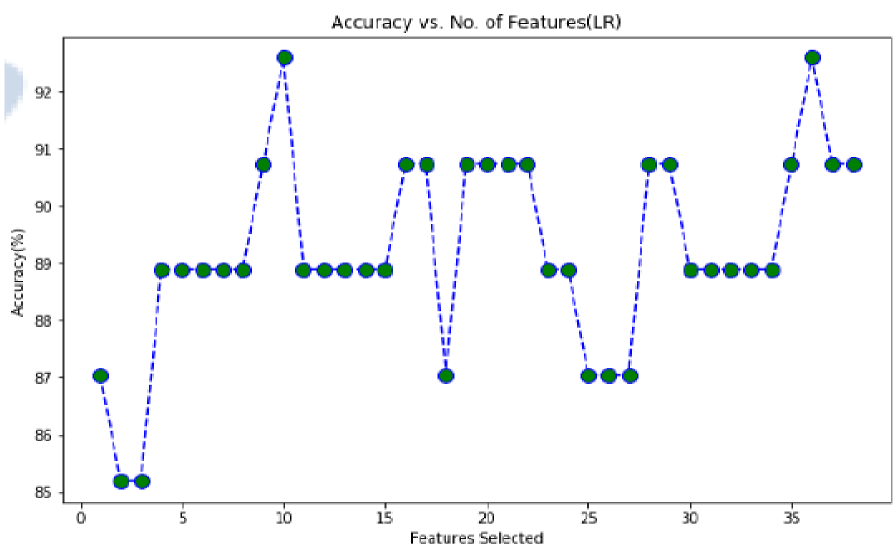

Figure 6: Selected features

ii) Logistic Regression

Logistic Regression is an extension of Simple Linear Regression that rounds off the result of input variables on the result variable as probability. Probability is what determines the relationship between the dependent and independent variables [5]. Here, the accuracy of Logistic Regression depends upon the number of features selected.

Accuracy vs. No. of Features (LR)

We can conclude that the highest accuracy occurs when the numbers of selected features are either 10 or 36. So, naturally, for the sake of computational cost, numbers of features that are selected are 10.

D) Making Predictions

Using the prepared models, predictions are made with testing set. Prediction using KNN model,

5. Result

A total of 538 samples of patients from ten different hospitals from Indore, India made the dataset. There were total of 39 parameters out of which only 9 parameters, with the highest weights were considered for KNN and 10 parameters, were considered for Logistic Regression. A comparison was made between the two different classifiers: linear and nonlinear. The liner classifier is KNN while the nonlinear classifier is the model of Logistic Regression. The F1 score helps determine the best model between the two. The F1 score for KNN is 0.90 and for that of Logistic Regression is 0.92, hence, model of Logistic Regression is selected to determine the absence or presence of PCOS.

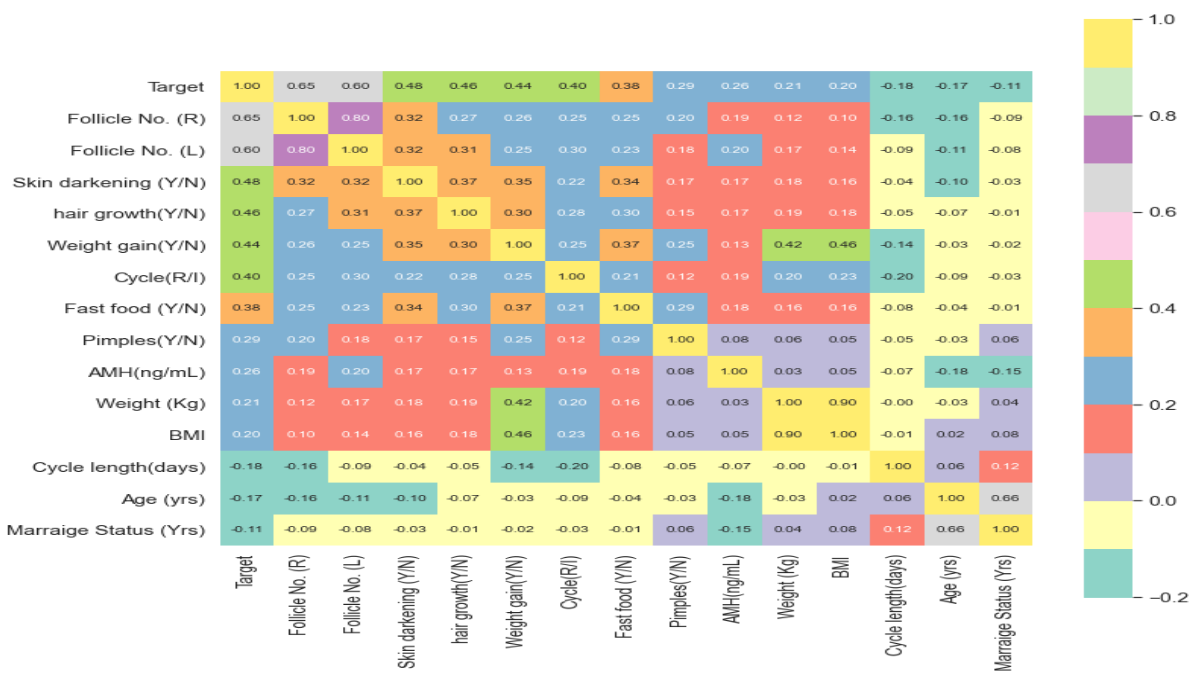

Figure: - 7 Correlation Matrix of all the features

In this section, we discuss complete analysis of machine learning algorithms and novel approach algorithms (XGBRF and CatBoost). After the execution of all the applied model, it is significant to check the performance of each and every model on a training and testing data. The various evaluation metrics have been taken into this research. We will explain the results implications of hyper parameters by using plots for ROC and accuracy of every model. XGBRF and CatBoost is the novelty of this research and it is compared with other machine learning models based on accuracy, precision, recall, f1-score, cross validation accuracy. By using the Algorithm 1 in univariate feature selection method, we determined the ranking of 43 features in our dataset. Table 3 shows the 10 best features.

Table 3: Ranking of 10 best features.

Features Name	**Ranking**
FSH (mIU/mL)	1
FSH/LH	2
Follicle No. (R)	3
Follicle No. (L)	4
AMH (ng/mL)	5
Cycle(R/I)	6
BMI	7
Avg. F size (L) (mm)	8
Cycle length(days)	9
Avg. F size (R) (mm)	10

From the table we can see that Follicle-stimulating hormone (FSH) has the best ranking. On number second there is the ratio of Follicle-stimulating hormone (FSH) Luteinizing hormone (LH) represented as FSH/LH. Follicle No.(R) is on number third respectively. AMU is the Anti-Mullerian hormone and BMI is Body Mass Index. Nowwe will explain the evaluation of our classifiers in next subsections.

5.1 Baseline Approach
Machine learning algorithms

In this section we will discuss the performance of our implemented classifiers. For a comprehensive comparison, the data was presented to models in pre-processed form.

Then we used Gradient Boosting, Random Forest, Logistic Regression, HRFLR, SVM, Decision Tree, MLP individually to produce the precision, recall, f1-score. we did a comprehensive comparison as shown in Table 4. Her N means Normal women and P is For PCOS women. For 10 best features, accuracy obtained by Gradient Boosting, Random Forest, Logistic Regression, HRFLR, SVM, Decision Tree, MLP are 0.82, 0.85, 0.87, 0.87, 0.85, 0.76, 0.83. It can be seen that Logistic Regression and HRFLR have got a good accuracy than other classifiers which means our model is 0.87 accurate. We checked our models

Table 4: Comparison with other classifiers: Here (N) means Normal women class and (P) means PCOS women class.

Baseline Model	**GB**	**RF**	**LR**	**HRFLR**	**SVM**	**DT**	**MLP**
Accuracy	0.82	0.85	0.87	0.87	0.85	0.76	0.83
Precision(N)	0.81	0.82	0.88	0.85	0.85	0.75	0.82
Recall(N)	0.86	0.89	0.88	0.91	0.86	0.80	0.86
F1-Score(N)	0.83	0.86	0.88	0.88	0.85	0.77	0.84
Precision(P)	0.85	0.88	0.88	0.90	0.85	0.79	0.85
Recall(P)	0.80	0.81	0.88	0.83	0.84	0.73	0.81
F1-Score(P)	0.82	0.84	0.88	0.87	0.85	0.76	0.83
AUC Score	0.90	0.90	0.92	0.93	0.91	0.76	0.90

for both Normal women prediction and PCOS women prediction. In precision, Logistic Regression got 0.88 which means it has correctly predicted the normal women. Further HRFLR got the highest Recall value i.e., 0.91 which means it accurately predicted how many truly are Normal women without PCOS. Moreover, Logistic Regression and HR-FLR both have high and same F1-score i.e., 0.88. In prediction of PCOS women, HRFLR got the highest value i.e., 0.90 which means it has correctly predicted the PCOS women and Logistic Regression got the highest Recall value i.e., 0.88 which means those who truly has PCOS are predicted. F1-Score for LR and HRFLR is 0.88, 0.87 respectively. Overall, we can say that Logistic Regression and HRFLR has worked better than other classifiers. From the Table 4, AUC score obtained by GB, RF, LR, HRFLR, SVM, DT, MLP are 0.90, 0.90, 0.92, 0.93, 0.91, 0.76, 0.90. It is quite clear that HRFLR has high AUC score than other classifiers. Adding to this, it means HRFLR has clearly distinguish which women is with and without PCOS. The higher the value of AUC higher is the performance of the model. Figure 8 shows the ROC Curve plot comparison of all models.

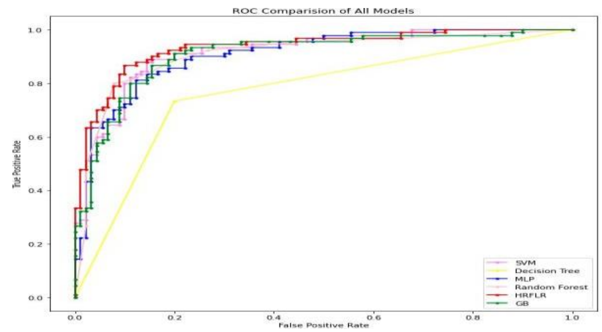

Figure 8: ROC comparison of all models

This ROC curve shows the tradeoff between false positive rate and true positive rate. It is quite evident from ROC curve plot that all classifiers are on the top left corner which means except Decision Tree all are performing better. It is quite clear from ROC curve plot that HRFLR performed better than other classifiers which means it accurately classify the Norman women and PCOS women and predicted which women is having PCOS.

5.2 Novel Approach

XGBRF and CatBoost Models

For the same experimental setting we tested these two models on our trained data. In XGBRF hyper parameters were taken as we take max depth = 3 and random state = 8 while in CatBoost 199 iteration were taken to get the results. Figure 9 shows the results of our novel models.

Baseline Model	XGBRF	CatBoost
Accuracy	0.89	0.95
Precision(N)	0.84	0.83
Recall(N)	0.90	0.90
F1-Score(N)	0.87	0.86
Precision(P)	0.89	0.89
Recall(P)	0.82	0.81
F1-Score(P)	0.86	0.84
AUC Score	0.92	0.90

Figure 9: Results of XGBRF and CatBoost

It is quite evident from the Figure that XGBRF and CatBoost got 0.89 and 0.95 accuracy which means how much our model accurately predicted both the classes. Moreover, XGBRF and CatBoost worked better in predicting the Normal women class with 0.84 and 0.83 precision value. They also worked better in predicting how many women are truly without PCOS having same 0.90 Recall value. 0.87 is the F1-score of XGBRF and 0.86 is CatBoost. Further, in terms of PCOS class prediction both models have same 0.89 precision value. Those who truly has PCOS were predicted by Recall score, 0.82 for XGBRF and 0.81 for CatBoost. F1-score of XGBRF is 0.86 and 0.84 for CatBoost.

From Figure 9, AUC score of XGBRF is 0.92 and 0.90 for CatBoost. It means both models have accurately classified Normal and PCOS women. Higher the value of AUC higher is the performance of the model. On the basis of accuracy CatBoost has performed better than XGBRF while on the basis of AUC score XGBRF has performed better than CatBoost. Now we will see the ROC curve plot of these two models. Figure 10 shows us ROC curve of both models.

The paper aim is to evaluate the newly proposed XGBRF and CatBoost models performance motivated by the perceived studies in the medical domain discussed in literature review. This approach was used in Giloma segmentation and Online transaction detection,we carried comprehensive comparison between baseline approach and newly proposed approach for solving the problem of early detection of PCOS. Table 5 shows the overall comparison and summary of the results of all models.

Table 5: Comparison of models based on accuracy.

Models	**Accuracy**
Gradient Boosting	0.85
Random Forest	0.85
Logistic Regression	0.87
HRFLR	**0.87**
SVM	0.85
Decision Tree	0.76
MLP	0.83
XGBRF	**0.89**
CatBoost	**0.95**

It is quite clear from the Table 5 all models give promising results. Our novel approach performed well from baseline approach. CatBoost outshined here with 0.95

accuracy which means it 0.95 model is accurate. Along, XGBRF also performed better than rest of the classifiers and achieved 0.89 accuracy. It also means 0.89 model is accurately predicting the classes.

Table 6: Comparison based on confusion matrices

Models	GB	RF	LR	HRFLR	SVM	DT	MLP	XGBRF	CatBoost
TP	77	80	79	**82**	77	72	77	**81**	**81**
FP	18	17	11	15	14	24	17	16	17
TN	72	73	79	**75**	76	66	73	**74**	**73**
FN	13	10	11	8	13	18	13	9	9

Next, we evaluate the confusion matrices of all classifiers. Table 6 shows the overall comparison of confusion matrices of all models. Confusion matrix shows us how our model is performing. It further tells us is our model predicting the right thing and what error is our model doing such type-1 error and type-2 error. If we have a type-2 error then our model is failing to predict. In our research all models such as Gradient Boosting, Random Forest, Logistic Regression, HRFLR, Support Vector Machine, Decision Tree and Multi-Layer Perceptron predicted very well which women is with and without PCOS.

It is clear from Table 6 that models like CatBoost, XGBRF, and HRFLR is superior as number of false positive PCOS and non-negative PCOS are less with the rest of the classifiers. All models performed well but HRFR, XGBRF and CatBoost performed better because of less false positive and false negative. Moreover, HRFLR has TP=82, that means 82 patients is having PCOS and TN=75, that means patients not having PCOS. Adding to this, FP=15, that means they don't have PCOS disease and FN=8, that means they predicted no but they have the disease. Overall, HRFLR, XGBRF and CatBoost performed well in confusion matrix.

Next the impact of cross validation accuracy is investigated and compared with other models. Table 7 shows the summary of cross validation accuracy of all models.

It is quite evident that here all models performed in K-fold cross validation accuracy. Table 7 shows the comparison based on K fold cross validation. In our case we used 10-, 20-, 30-, and 40-fold cross validation. First when we use k=10, results shows that CatBoost achieved 0.89 accuracy. When k=20, HRFLR out-shined CatBoost and achieved 0.89 accuracy. When k=30, CatBoost again achieved 0.89 accuracy and at last when k=40 HRFLR again achieved 0.90 accuracy. So overall both were performing better from other classifiers. Hence, this is an important improvement as they successfully detected the true PCOS. An early and correct detection of PCOS will begin the mandatory treatment. As per the state-of-art methods in our research, the accuracy achieved 0.96 for glioma segmentation and 0.98 for online transaction detection [Bhatele and Bhadauria, 2020] [Li et al., 2020], which justify our research in detecting the PCOS using XGBRF and CatBoost model.

Table 7: Comparison based on K fold cross validation.

Models	**Value of K**	**Cross Validation Accuracy**	**Standard Deviation**
Gradient Boosting	10	0.85	0.5
	20	0.86	0.6
	30	0.85	0.8
	40	0.86	0.10
Random Forest	10	0.86	0.5
	20	0.88	0.7
	30	0.87	0.7
	40	0.88	0.9
Logistic Regression	10	0.85	0.4
	20	0.84	0.7
	30	0.85	0.8
	40	0.84	0.9
HRFLR	10	0.89	0.5
	20	0.89	0.5
	30	0.89	0.6
	40	0.90	0.8
SVM	10	0.84	0.4
	20	0.85	0.8
	30	0.84	0.9
	40	0.84	0.11
Decision Tree	10	0.83	0.6
	20	0.84	0.8
	30	0.85	0.9
	40	0.84	0.10
MLP	10	0.84	0.4
	20	0.83	0.7
	30	0.84	0.8
	40	0.83	0.10
	10	0.83	0.4

XGBRF	20	0.83	0.7
	30	0.84	0.8
	40	0.84	0.10
CatBoost	**10**	**0.89**	**0.4**
	20	**0.88**	**0.6**
	30	**0.89**	**0.7**
	40	**0.89**	**0.8**

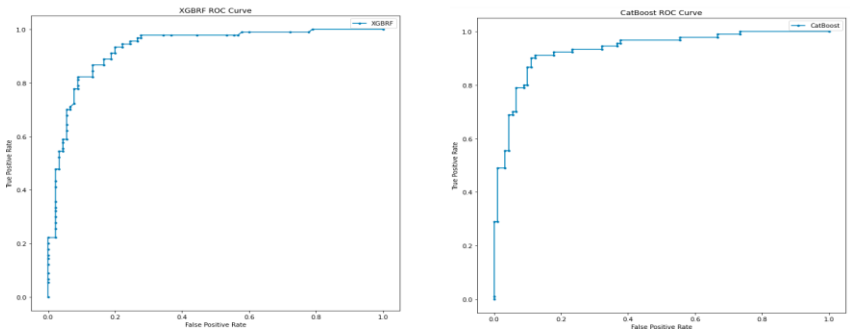

Figure 10: ROC of XGBRF and CatBoost

In Figure 10 it is quite evident from ROC curve plot of both classifiers are on the top left corner which means their performance is better. XGBRF has high true positive rate and less false negative rate which means XGBRF is accurately distinguishing the Normal women and PCOS women. While CatBoost has slight less true positive rate then XGBRF. So, on comparison XGBRF performed better than CatBoost in ROC curve plot.

6. Conclusion

Detection PCOS at an early stage enhance the early treatment of the patients. An automated system which can be beneficial for detecting the PCOS based on clinical and metabolic parameters. The research aims to detect the PCOS using Hybrid XGBRF and Catboost models. we also used machine learning algorithms such as Gradient Boosting, Random Forest, Logistic Regression, Hybrid Random Forest and Logistic Regression, SVM, Decision Tree, MLP. The dataset obtained from Kaggle repository contains 541 patients with 43 attributes. Results showed that attribute FSH is the most important attribute than other attributes. Results also indicated that if we take 10 features only then good accuracy can be achieved which takes less computation time. we implemented nine classifiers on 10 features. It is shown in research paper that our novel approach models such as XGBRF and CatBoost achieved i.e 0.86 and 0.95 accuracy which outperformed other classifiers. On comparison with other classifiers CatBoost got 0.89 K fold cross validation accuracy. The results of XGBRF and CatBoost were compared with other classifiers reported in related work and overall, it proved that CatBoost outperformed all the classifiers.

In future work, the result achieved in this research can be validated if we have large dataset by having more patients like one thousand. Furthermore, a new hybrid algorithm can be produced and if we have larger dataset then deep learning algorithms like optimized form of CNN can be implemented to increase the classification accuracy. There is a huge scope for this research as cases of PCOS are increasing day by day.

ACKNOWLEDGMENTS

The dataset used here is downloaded from Kaggle from the following link: https://www.kaggle.com/prasoonkottarathil/polycysticovary-syndrome-pcos

REFERENCES

1. Dana AL-Dlaeen and Abdallah Alashqur. Using decision tree classification to assist in the prediction of alzheimer's disease. In *2014 6th International Conference on Computer Science and Information Technology (CSIT)*, pages 122–126, 2014. doi:10.1109/CSIT.2014.6805989.

2. Subrato Bharati, Prajoy Podder, and M. Rubaiyat Hossain Mondal. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 1486–1489, 2020. doi:10.1109/TENSYMP50017.2020.9230932.

3. Kirti Raj Bhatele and Sarita Singh Bhadauria. Glioma segmentation and classification system based on proposed texture features extraction method and hybrid ensemble learning. In *2017 2nd International Conference for Convergence in Technology (I2CT)*, volume 37, pages 989–1001, 2020. doi:10.18280/ts.370611.

4. Rok Blagus and Lara Lusa. Evaluation of smote for high-dimensional class imbalanced microarray data. In *2012 11th International Conference on Machine Learning and Applications*, volume 2, pages 89–94, 2012. doi:10.1109/ICMLA.2012.183.

5. B. Cahyono, Adiwijaya, M. S. Mubarok, and U.N. Wisesty. An implementation of convolutional neural network on pco classification based on ultrasound image. In *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, pages 1–4, 2017. doi:10.1109/ICoICT.2017.8074702.

6. Chandan Chakraborty, Palak Mehrotra, Biswanath Ghoshdastidar, Sudarshan Ghoshdastidar, and Kakoli Ghoshdastidar. Automated ovarian follicle recognition for poly-

cystic ovary syndrome. In *2011 International Conference on Image Information Processing*, pages 1–4, 2011. doi:10.1109/ICIIP.2011.6108968.

7. Chih-Yu Chen, Chi-Wei Chang, Sheng-Hung Wang, and Gin-Chung Wang. The effects of polycystic ovary syndrome on cardiovascular system in women of childbearing age. In *BIBE 2019; The Third International Conference on Biological Information and Biomedical Engineering*, pages 1–4, 2019.
8. Subhasish Deb, Arup Kumar Goswami, Rahul Lamichane Chetri, and Rajesh Roy. Prediction of plug-in electric vehicle's state-of-charge using gradient boosting method and random forest method. In 2020 *IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, pages 1–6, 2020. doi:10.1109/PEDES49360.2020.9379906.
9. Yinhui Deng, Yuanyuan Wang, and Ping Chen. Automated detection of polycystic ovary syndrome from ultrasound images. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4772–4775, 2008. doi:10.1109/IEMBS.2008.4650280.
10. Amsy Denny, Anita Raj, Ashi Ashok, C Maneesh Ram, and Remya George. i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 673–678, 2019. doi:10.1109/TENCON.2019.8929674.
11. Muhammad Sakib Khan Inan, Rubaiath E Ulfath, Fahim Irfan Alam, Fateha Khanam Bappee, and Rizwan Hasan. Improved sampling and feature selection to support extreme gradient boosting for pcos diagnosis. In *2021 IEEE 11th Annual Comput- ing and Communication Workshop and Conference (CCWC)*, pages 1046–1050, 2021. doi:10.1109/CCWC51732.2021.9375994.

12. H. Prasanna Kumar and S. Srinivasan. Segmentation of polycystic ovary in ultrasound images. In *Second International Conference on Current Trends In Engineering and Technology - ICCTET 2014*, pages 237–240, 2014. doi:10.1109/ICCTET.2014.6966294.

13. Maryruth J. Lawrence, Mark G. Eramian, Roger A. Pierson, and Eric Neufeld. Computer assisted detection of polycystic ovary morphology in ultrasound images. In *Fourth Canadian Conference on Computer and Robot Vision (CRV '07)*, pages 105–112, 2007. doi:10.1109/CRV.2007.18.

14. Yunlong Li, Yingan Mai, Zijian Lin, and Shufen Liang. Online transaction detection method using catboost model. In *2020 International Conference on Communica- tions, Information System and Computer Engineering (CISCE)*, pages 236–240, 2020. doi:10.1109/CISCE50729.2020.00053.

15. J. Madhumitha, M. Kalaiyarasi, and S. Sakthiya Ram. Automated polycystic ovarian syndrome identification with follicle recognition. In *2021 3rd International Con- ference on Signal Processing and Communication (ICPSC)*, pages 98–102, 2021. doi:10.1109/ICSPC51351.2021.9451720.

16. Palak Mehrotra, Jyotirmoy Chatterjee, Chandan Chakraborty, Biswanath Ghoshdastidar, and Sudarshan Ghoshdastidar. Automated screening of polycystic ovary syndrome using machine learning techniques. In *2011 Annual IEEE India Conference*, pages 1–5, 2011. doi:10.1109/INDCON.2011.6139331.

17. Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542– 81554, 2019. doi:10.1109/ACCESS.2019.2923707.

18. Amy Neuzil. What is polycycstic ovary syndrome (pcos). URL www.austinfitmagazine.com/November-2014/What-is-Polycycstic-Ovary-Syndrome-PCOS/.

19. Aroni Saha Prapty and Tanzim Tamanna Shitu. An efficient decision tree establishment and performance analysis with different machine learning approaches on polycystic ovary syndrome. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5, 2020. doi:10.1109/ICCIT51783.2020.9392666.

20. Bedy Purnama, Untari Novia Wisesti, Adiwijaya, Fhira Nhita, Andini Gayatri, and Titik Mutiah. A classification of polycystic ovary syndrome based on follicle detection of ul- trasound images. In *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, pages 396–401, 2015. doi:10.1109/ICoICT.2015.7231458.

21. Sandy Rihana, Hares Moussallem, Chiraz Skaf, and Charles Yaacoub. Automated algorithm for ovarian cysts detection in ultrasonogram. In *2013 2nd Interna- tional Conference on Advances in Biomedical Engineering*, pages 219–222, 2013. doi:10.1109/ICABME.2013.6648887.

22. Gabriel Rushin, Cody Stancil, Muyang Sun, Stephen Adams, and Peter Beling. Horse race analysis in credit card fraud—deep learning, logistic regression, and gradient boosted tree. In *2017 Systems and Information Engineering Design Symposium (SIEDS)*, pages 117–121, 2017. doi:10.1109/SIEDS.2017.7937700.

23. Hotaka Saito and Hiromistu Ohmori. Control of an abnormal human menstrual cycle in pcos by speed gradient algorithm. In *SICE Annual Conference 2011*, pages 1436–1441,2011.

24. E. Setiawati, Adiwijaya, and A. B. W. Tjokorda. Particle swarm optimization on follicles segmentation to support pcos detection. In *2015 3rd International Confer- ence on Information and Communication Technology (ICoICT)*, pages 369–374, 2015. doi:10.1109/ICoICT.2015.7231453.

25. Gurpreet Singh and Manoj Sachan. Multi-layer perceptron (mlp) neural network technique for offline handwritten gurmukhi character recognition. In *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–5, 2014. doi:10.1109/ICCIC.2014.7238334.

26. Ranjitha Sitheswaran and S. Malarkhodi. An effective automated system in follicle identification for polycystic ovary syndrome using ultrasound images. In *2014 International Conference on Electronics and Communication Systems (ICECS)*, pages 1–5, 2014. doi:10.1109/ECS.2014.6892634.

27. Palvi Soni and Sheveta Vashisht. Exploration on polycystic ovarian syndrome and data mining techniques. In *2018 3rd International Conference on Communication and Elec- tronics Systems (ICCES)*, pages 816–820, 2018. doi:10.1109/CESYS.2018.8724087.

28. Pilar Vigil, Manuel E. Cortés, Ana Zúñiga, Jessica Riquelme, and Francisco Ceric. Scan-

ning electron and light microscopy study of the cervical mucus in women with polycystic ovary syndrome. *Microscopy*, 58(1):21–27, 2009. doi:10.1093/jmicro/dfn032.

29. Ning-Ning Xie, Fang-Fang Wang, Jue Zhou, Chang Liu, and Fan Qu. Establishment and analysis of a combined diagnostic model of polycystic ovary syndrome with random forest and artificial neural network. *BioMed Research International*, 2020:1–13, 08 2020. doi:10.1155/2020/2613091.

30. Guojun Zhang and Jixiong Chen. A svm classifier research based on rs reducts. In *2009 International Conference on Information Management, Innovation Management and Industrial Engineering*, volume 1, pages 227–230, 2009. doi:10.1109/ICIII.2009.62.

1. Mariana StuparuCretu and Aural Nechita, "Ovarian Ultrasound Imaging for Polycystic Ovary Syndrome in Adolescents- Types of Approach," pp.127-132, 2013.
2. E. Setiawati, Adiwijaya and A.B.W Tjokorda, "Particle Swarm Optimization on Follicles Segmentation to support PCOS Detection," 3rd International Conference on Information and Communication Technology, pp.371-376, 2015.
3. Bedy Purnama, Untari Novia Wiseti, Adiwijaya, Fhira Nhita, Andini Gayatri and Titik Mutiah, "A Classification of Polycystic ovarian Syndrome Based on Follicle Detection of Ultrasound Images," 3rd International Conference on Information and Communication Technology, pp.399-403, 2015.
4. P.S. Hiremath and Jyothi R. Tegnoor, "Automatic Detection of Follicles in Ultrasound Images of Ovaries using Active Contours Method," IEEE International Conference on Computational Intelligence and Computing Research, pp.1-7, 2010.
5. Marla E. Lujan, Brittany Y. Jarrett, Eric D. Brooks, Jonathan K. Reines, Andrew K. Peppin, Narry Muhn, Ehsan Haider, Roger A. Pierson, and Donna R. Chizen, "Updated Ultrasound Criteria for Polycystic Ovary Syndrome: Reliable thresholds for elevated follicle population and Ovarian Volume," Human Reproduction, pp. 1361- 1368, 2013.
6. Terrence Chen, Wei Zhang, Sara Good, Kevin S. Zhou, Dorin Comaniciu, "Automatic Ovarian Follicle Quantification from 3D Ultrasound data using Global/Local with database Guided Segmentation", pp.1-8.
7. P.S. Hiremath and Jyothi R. Tegnoor, "Automatic Detection of Follicles in Ultrasound Images of Ovaries using Edge based Method," IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition, pp.120-125,2010.
8. P.S. Hiremath and Jyothi R. Tegnoor, "Automated Ovarian Classification in Digital Ultrasound Images," International Journal of Biomedical Engineering and Technology, pp.46-65, 2013.
9. B. Padmapriya and T. Kesavamurthy, "Diagnostic Tool for PCOS Classification," 7th WACBE World Congress on Bioengineering, pp 182-185, 2015.
10. Courtney A. Marsh, Alison Berent Spillson, Tiffany Love, Carol C. Persad, Rodica Pop Busui, Jon-Kar Zubieta and Yolanda R. Smith, "Functional neuroimaging of emotional processing in women with polycystic ovary syndrome: a case-control pilot study," 10.1016/j.fertnstert.2013.02.054, 2013.
11. Hsu Chia-Liang, Hsiu Hsin, Hsu Wei-Chen and Chen Chih-Yu, "Characteristic of Harmonic Indexes of the arterial blood pressure waveform in Polycystic Ovary Syndrome," Blood Pressure Monitoring, Volume 19 - Issue 4 – pp. 226–232, August 2014.
12. A Patel, S MacMahon, J Chalmers, B Neal, L Billot, M Woodward and et al. "Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes”, N Engl J Med, pp. 2560-72, 2008.