Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms

MSc Research Project
MSc. Data Analytics

Shakoor Ahmad Bhat
Student ID: x19236280

School of Computing
National College of Ireland

Supervisor: Dr. Rashmi Gupta

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shakoor Ahmad Bhat
Student ID:	x19236280
Programme:	MSc. Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Rashmi Gupta
Submission Due Date:	16/08/2021
Project Title:	Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms
Word Count:	9321
Page Count:	30

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms

Shakoor Ahmad Bhat
x19236280

Abstract

This research focus on data driven detection of PolyCystic Ovary Syndrome (PCOS) which is an medical disorder causes female fertility, affecting women in their childbearing age, even steady far off the reproductive age. This medical disorder leads to risk of complex long-term complications. Considering the supreme identification abilities of boosting and begging algorithms, especially in the medical domain. We combined Extreme Boosting with Random Forest(XGBRF). We proposed a new method, such as XGBRF and CatBoost model for early identification of PolyCystic Ovary Syndrome. To completely support this effective classification performance, data were re-sampled based on Synthetic Minority Over-sampling Techniques(SMOTE) to solve outliers and data imbalance issues. By exploiting univariate feature selection method, we identified top 10 important clinical and metabolic parameters which classify PolyCystic Ovary Syndrome conditions. We found that FSH(Follicle-stimulating hormone) is one of the significant parameter followed by LH(Luteinizing hormone). We tested models based on evaluation matrices such as Accuracy, Precision, Recall, F1-score, ROC curve plot, AUC score and K Fold Cross validation. At last, we investigate our model on a PCOS dataset collected from Kaggle repository to justify our novel approach. The other classifiers such as Gradient Boosting, Random Forest, Logistic regression, HRFLR, SVM, Decision Tree and MLP were applied as baseline approach to compare the results. Findings show that CatBoost and XGBRF outperformed all other models with an accuracy score of 0.95 and 0.89 respectively applied on top 10 parameters. Hence, CatBoost is suitable for detecting PolyCystic Ovary Syndrome.

Keywords– PCOS, Machine learning, Medical domain, Feature selection, Tuning, Boosting.

1 Introduction

Innovation and human beings working together can create a way for better services in terms of health care. Machine learning a subsection of artificial intelligence where it gives the system a potential to learn and enhance automatically irrespective of being programmed clearly. It primarily focuses on developing new machine learning algorithms which give access to given datasets and use the data for study and research purposes of the openwork. Machine learning applications accompany huge transformation mainly in industries like health that involve diagnosis, image recognition, identification and prediction of data, etc.

Polycystic ovary syndrome is an endocrine medical disorder which affects mainly women's throughout their period of adolescence. It was first narrated by Leventhal and stein in 1935. Women which are affected by polycystic ovary syndrome agonize from the imbalance of hormone level. It causes critical health problems such as irregular menstrual periods and problems related to getting pregnant. A woman while bearing a child faces hormonal irregularity in the middle of 15yr- 40yr age group. According to a research, the vulnerability of PCOS is 4.8% in white Americans, in African Americans it is 8% , in spain 6.8% and in Asia it is 31.3% . Women having PCOS suffer from diseases like hypertension, cardiovascular disease, type 2 diabetes, obesity, gynecological cancer, hazardous pregnancy and Mellitus. Symptoms for PCOS are like acne problem, high blood pressure, irregular menstruation, increase in body weight, increase in androgen hormone levels etc. we examine PCOS as the main reason for infertility as it hold back the real evolution of follicle which anatomize the maturity of ovaries [Prapty and Shitu, 2020].

Recent research shows the high risk of first Trimester miscarriage. 12-21% women in their reproductive age suffer from PCOS and out of them, 70% remains undetected. This disease can be treated by taking medication prescribed by doctors and changes in lifestyle habits. Medication includes birth control pills, diabetes tablets, medicine for anti androgen, fertility and ultrasound scan. Diagnosis of PCOS is done by barring of immaterial symptoms or test outcomes, mostly because of uneducated composite pathomechanism. These various symptoms force doctors to do a heavy number of clinical tests outcomes and irrelevant radio-logical imaging course of action [Denny et al., 2019].

The reproductive system of women wholly depend on unmatched hormones and required to be balanced for the processes which are needed for conception, ovulation and forming of a child in women's womb. Four hormones are needed namely progesterone, luteinizing hormone(LH), estrogen and follicle stimulating hormone(FSH). FSH and LH hormones are generated from Pituitary gland while as inside ovaries progesterone and estrogen are produced. For a good balanced reproductive system of women both progesterone and estrogen are very well important. Women with PCOS will have risks like sleep apnea, infertility, abnormal uterine bleeding, high cholesterol, elevated lipids, nonalcoholic fatty liver, liver disease, depression and anxiety, high blood pressure, metabolic syndrome, Miscarriages and Cardiac risks. PCOS can be predicted by symptoms like unwanted or excess growth of hairs in the body or face, Amenorrhea in 30 to 40% of women, increase in weight around the waist, before period swollen of breasts, during periods Neuralgic pain occurs, Hysteria, Itchy vagina and vulva and Cysts on ovaries [Soni and Vashisht, 2018].

According to the doctors, women having cysts in ovaries is not one of the main reason or parameter for diagnosis of PCOS. Studies suggest that 30-70% women having PCOS suffers mostly from obesity and it proves that there is a bidirectional relationship between PCOS and obesity. Although highly secretion of androgens, irregular menses and huge number of cysts in the ovary are declared as primary criteria for detection of PCOS. Studies show that these clinical features and can be used as important parameters for the early detection of PCOS. Figure 1 shows the difference between the Normal Ovary and PolyCystic Ovary [Khan Inan et al., 2021].

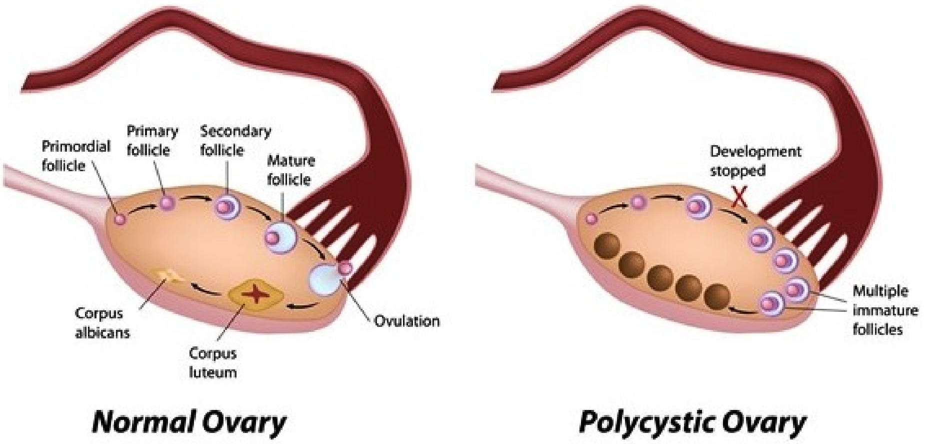


Figure 1: Difference between Normal Ovary and PCOS Ovary [Neuzil]

The principal aim of this paper is to check the state of art methods for the detection of PCOS. The state of art methods integrate the Hybrid machine learning algorithms based on hybrid binary classification. Here, approach of a state of art methods carrying out Hybrid XGBRF classification on low and high grade Glioma and online transaction detection using CatBoost method [Bhatele and Bhadauria, 2020] [Li et al., 2020]. The outcomes performed well than other methods used in these papers and focusing on the success of this state of art approach, the primary goal is to implement these two algorithms on PCOS dataset for the detection of PCOS and evaluate the model on various metrics. It is still unknown that which algorithm will perform better and which algorithm will give best results in order to detect the PCOS. The vital challenge is to select the best attributes like which one is the crucial attribute for detection of PCOS in terms of FSH, LH, AMH, BMI, weight gain, cycle, Follicle NO, cycle length days, etc. Hence, we will implement the machine learning classifier algorithms, i.e. Hybrid Extreme Gradient Boosting with Random Forest (XGBRF) ensemble method and CatBoost Model, which includes physical and clinical parameters to detect the PCOS.

The major contributions of this research is as: (i) pre-processing, which will include a selection of salient features using the feature selection method; (ii) plotting an architecture for hybrid machine learning model; (iii) implementing the Hybrid Extreme Gradient Boosting with Random Forest (XGBRF) ensemble method and CatBoost Model and comparing their results with baseline approaches by providing the reasonable comparison to the research.

We outline the research like this: Section 2 provides the related work on the supplication of stats tools and machine learning models on the detection of PCOS; Section 3 represents the entire process of methodology applied in the given work; Section 4 discuss the architecture of the hybrid machine learning models; Section 5 describes the state of art methods and novel approach for the detection of PCOS; Section 6 provides the evaluation of the implemented models; Section 7 will bring forth the concluding remarks.

2 Related Work

Detection of polycystic ovary syndrome has become a new topic for researchers since the last decade. Researchers have implemented various techniques to diagnose PCOS at an early stage. PCOS is an endocrine medical disorder with many diagnostic criteria because of its heterogenic manifestations. One of the primary diagnostic criteria includes examination of ovaries seen by ultrasound images in terms of number, size, and follicle distribution inside the ovary. This process includes manual tracing and follicle counting on the ultrasound images to decide PCOS. [Lawrence et al., 2007] presented a novel method which automated the identification of PCOS. The algorithm included the follicle segmentation from ultrasound images, computing the features of the instinctively segmented follicles using follicle stereology, storing the follicle features as feature vectors, and at the last classification of feature vectors. They made two categories: one is PCO present and other is PCO absent. This automated tool saved a lot of time consumed with a manual tracing of follicles and calculating the width and length of each follicle. Three classifiers were used, namely linear discriminant classifier(LDC), KNN and SVM. Results were very promising, as LDC achieved an accuracy of 0.92 , 0.91 by KNN and 0.91 by SVM, respectively. Overall, LDC outperformed SVM and KNN, but all three classifiers gave promising results. They reduced the risk of serious complications which can be caused by PCOS from delayed detection.

The development of various follicular cysts inside the ovary distinguishes this PCOS disorder. These-days detection of PCOS performed by medical expertise is to manually calculate the number follicle cysts that can lead to complication of the low efficiency, reproducibility and variability. To control these complications [Deng et al., 2008] proposed an automated scheme to identify the PCOS. First, the ultrasound image of the ovary is taken as input and filtered by using the adaptive morphological filter. Then adapted labeled watershed model is used to bring out the contours of targets. At last, clustering method is implemented to detect the expected follicle cysts. This investigation verifies the efficiency of the implemented automated scheme, which achieved the 0.84 accuracy. However, because of the high apposite of the PCOS, this automated scheme can not be implemented to other various targets identification problems.

There are two types of cervical mucus that are recognized, namely gestagenic and oestrogenic. [Vigil et al., 2009] goal was to detect the characteristics of crystallization and ultrastructure related to the cervical mucus in women which is suffering from PCOS and to contrast these characteristics with normal women. They took 10 samples of cervical mucus from women, out of which 4 belongs to normal women and 6 were from women suffering from PCOS. Because of crystallization and ultra-structure, they characterized mucus. When the samples were taken, the levels got related to the type of mucus were progesterone and oestradiol. To consider mucus ultra-structure, they established differentiation between the women with PCOS and controlled women and anovulatory cycle of menses. These variations were obvious in the mesh and average mucus diameter of poses. In controlled women, Mucus crystallization showed the regular disposition of oestrogen like fern, hexagonal shape or rectilinear. While women having PCOS, unknown mucus crystallizations were established, apart from that patches of crystallization which resembles gestagenic-like mucus and oestrogen. This research shows that characteristics of crystallization and ultra-structure of the cervical mucus in PCOS women are unlike then controlled women.

The PCOS can cause severe problems like anovulation and infertility. The criterion for PCOS detection includes metabolic and clinical parameters which are crucial for early pointer for this disease. [Mehrotra et al., 2011] described a new method which automated detection of PCOS because of these early markers. The model entails feature vector formulation based on the parameters of metabolic and clinical attributes.Further, significant statistical features for discerning between PCOS and normal groups are determined because of two sample t-test. Bayesian Classifier and Logistic Regression were implemented. This automated system acted as an assistant tool for the medical expertise which saved precious time in analysing the patients and thus reduced the delay the risk inof detection of PCOS. Results were very promising, as Bayesian Classifier achieved accuracy of 0.93 and 0.91 were achieved by Logistic Regression.

PCOS affects 10% women mainly when they are in reproductive age and leads to infertility. [Saito and Ohmori, 2011] proposed a new dosing system that is build on mathematical method for PCOS, a type of sterilities. This proposed system entitle us to initiate a new treatment for PCOS suited for sole PCOS patients. Performance is based on computer simulations. It produce sufficient LH surges unusual cycle by speed gradient model. If this system will be applied, it will be possible to give this dosage and helps in minimizing the side effects caused by PCOS especially pregnancy of women. [Chakraborty et al., 2011] described an automated scheme where the diagnosis of pathognomonic pattern and follicle arrangement is proposed to control this problem. The data is collected from GDIFR (GD Institute for Fertility Research) located in Kolkata. Patients from age group 25-35years diagnosed from PCOS are included. Ultrasonographer were used to get the ultrasounds of the patients by using the 7MHz transducer (General Electricals, Milwaukee, USA) and later verification were done by Gynaecologist. The imaging were pre-processed as per the methodology used. First, ultrasound image were pre-processed by the approach of multiscale morphological for contrast enhancement. Thenthresholding scaline is used for the extraction of follicle contours. The findings are then compared with the manual selection results to verify the potency of scheme.

PCOS disorder only affect women's health and it can be treated either by medication or surgery. Manual examination of PCOS detection frequently produce errors. [Rihana et al., 2013] described an algorithms which is capable of identifying cysts from the ultrasound images of ovaries and of alter between the kind of cysts. 25 digital recording with ultrasound images were taken for the model. Images were provided by MD. Barakat who is a gynaecologist. Images pre-processing were implemented in which images were converted into grey-scale and contrast enhancement were performed as well which are totally based on operation of morphology. After this feature extraction were used to characteristics follicles based on standard parameters. SVM classifier were used for evaluation and validated by using ROC. 0.90 accuracy were achieved which is very promising.

Nowadays, PCOS is a common disorder seen in women's which is caused by the development of various follicles in ovary. [Sitheswaran and Malarkhodi, 2014] described an effective model for the computer aided detection of PCOS by using the ultrasound images of ovary. The model is identifying follicles by using the object growing algorithm. It goes through two stages one is pre-processing phase and another is follicle identification on the basis of object growing algorithm. The speckle noise found in ultrasound image is decreased by using the median filter. After this, labeled watershed algorithm is used on possible follicles and extraction of local minimum takes place. Region of attentiveness is picked to build the segmentation part effortless. The use of object growing algorithm is to select and detect the follicles. It is significant to recognize and detect the many forms of ovarian failure which can lead to infertility. PCOS can be determined by using images of ultrasounds which provides significant information related to size and number of follicles presented in ovary. [Kumar and Srinivasan, 2014] used Improved Chan-Vase Method which can detect follicles and can do fast segmentation. 50 images were collected from JB Diagnostic Center Bangalore in the age group of 25-35 having the PCOS. Pictures were taken from LOGIQ P3 (General Electricals, Milwaukee, USA). The model were implemented in Mat Lab version 7.10 and results shows that ICV algorithm is working faster then CV algorithm with low iteration and less computational time.

Currently, digital medical images are frequently used for patient's medical records in hospital assistance. [Setiawati et al., 2015] proposed an image clustering method for the segmentation of follicles by using the PSO(Particle Swarm Optimization) as well as non-parametric fitness function. This fitness function utilize Normalized Mean Square Error and Mean Structural Similarity Index to create more convergent and close packed cluster. Results showed that fitness function of PSO produced extra convergent solution than preceding fitness function used in other research mostly on ultrasound images. Further, it explored that impact of contrast enhancement to the PSO performance on image clustering and follicle size extraction. Findings showed that PSO image clustering with contrast enhancement build quantization error, and intra-cluster distance than PSO image clustering without contrast enhancement. PSO with contrast enhancement put together near Region of interest to ROI reference.

According to a report published by National Institutes of Health (NIH) says at around 6-10% women are suffering from PCOS and Rotterdam defined 15% have PCOS prevalency. [Purnama et al., 2015] described an application which will classify PCOS based on identification of follicle by using the images of USG. First the stage is pre-processing which engage equalization histogram, morphological processes, low pass filter, and binarization to acquire binary images of follicles. Then, next phase is segmentation with cropping, edge detection, and labelling of the follicle images. Further phase is feature extraction by using the Gabor wavelet. They used two datasets one is dataset A contains 40 images which has 26 images of normal women and 14 images of women having PCOS. The dataset B contains 40 images out of which 34 images of normal women and 6 images of women having PCOS. On the basis of feature vector evolved from feature extraction identified attribute of PCOS and Non-PCOS follicles. Results were very promising SVM-RBF kernel having C=40 shows that dataset A achieved 0.82 and dataset B which used KNN-euclidean distance classifier having k=5 achieved 0.78 accuracy.

PCOS is a big concern especially in married people based on rate of infertility of PCOS women. In previous papers feature extraction of ultrasound images were done manually to overcome this [Cahyono et al., 2017] proposed an solution by using the CNN algorithm. They used an dataset which was labelled by expert. Dateset were categorised into three groups, 0.60 of training data, 0.20 of test data, 0.20 of validation data. Dataset contains 40

Non-PCOS data and 14 PCOS data in which 24 Non-PCOS data and 8 PCOS data were put into train data. 8 Non-PCOS data and 3 PCOS data were put into test and validation data. Data were duplicated in order to make it balanced. CNN were implemented at the end and 1 of F1-score were achieved with 0.76 of 5-Fold cross validation.

In this modern era, the data is explored and assembled in one place in data warehouse and to take out useful information out of it is process of data mining. Data mining uses many set of rules which produce predictions, classification, clustering, and associations etc. Today women suffer from many diseases and PCOS is one of them. [Soni and Vashisht, 2018] highlighted symptoms and many risks associated with PCOS, treatment convenient for PCOS, and many Data mining methods that can be used to detect the PCOS with high rate of accuracy. First, feature extraction and anlysis can be used to extract top ranking features out of the dataset. Then, classification algorithm like Naive Bayes classifier, Decision Tree, SVM can be implemented to get the high accuracy. Clustering methods like Partition, Hierarchical, Density based can also applied on significant elements.

It is estimated that 5 million women in their reproductive age are affected from PCOS worldwide. It is getting difficult to detect PCOS because of its heterogeneity symptoms and existence of many gynecological disorders. The time involving in clinical tests and scanning of ovary has become freight for PCOS patients. In order to counter this problem [Denny et al., 2019] proposed an system which will detect PCOS at an early stage. The dataset selected for this research contains survey of 541 patients mainly women based on medical expertise consultations and clinical analysis. Out of 23 attributes, 8 significant features which has the potential are picked by using the SPSS V 22.0. PCA is mainly used with classification algorithms like KNN, Logistic Regression, Naive Bayes Classifier, Random Forest, SVM in Spyder Python IDE. Results shows that Random Forest is the accurate and outperformed other algorithms with an 0.89 accuracy.

[Chen et al., 2019] there aim was to study the effects caused by PCOS on cardiovascular system by parameters of hemodynamic like radial pulse spectroscopy, and blood pressure. A data of 242 women in their childbearing age took part in this study and gain standard measurement of blooed pressure and non-obtrusive radial pulse wave measurement. It were calculated by Fourier model. Findings shows no important difference in diastolic blood pressure, and systolic between PCOS women and without PCOS women. Further, the study also revealed that women having PCOS have high body index C2 and C4, and heart rate on comparison with normal women. Logistic Regression revealed that C2 and high body index are most significant risk factors for polycystic ovary syndrome.

PCOS is now considered a very serious disease for which women pays a lifelong damage like infertility because women do not know that she is affected by PCOS at an early stage but if she gets to know this disease it can be treated under medical supervision. [Prapty and Shitu, 2020] applied many machine learning algorithms and efficient classification tree is confirmed based on performance. They used a data having 542 women out of which 177 are suffering from PCOS. 31 feature are available in this in which 7-8 features were selected like irregular menstrual cycle length, BMI index, hemoglobin rate in blood, pregnancy, respiratory rate, FSH, thyroid hormone, endometrium, prolactin level, and anti mullerian hormone level. Further, noticing the changes in the women body and hormonal tests clasification trees were implemented like SVM, KNN, Naive Bayes Classifier and Random Forest. Results were very promising as Random Forest outperformed with 0.93 accuracy.

Approximately, 5-10% womens in their reproductive age(15-49 years old) suffers from PCOS. Further, it was documented that in African American women 8% were impacted by this outbreak and 4.8% in White women. [Bharati et al., 2020] implemented many machine learning models. Dataset was taken from Kaggle repository Dataset contains 43 features and 10 top ranking features were selected using the feature selection algorithm 1 that can detect PCOS at an early stage. It was seen that FSH and LH hormones are the important features. After that, next holdout and cross validation model were implemented to divide the dataset into train and test. Classifiers like Random Forest, Gradient Boosting, Logistic Regression, and Hybrid Random Forest and Logistic Regression were applied on the dataset and were evaluated. Findings shows that RFLR got the best results 0.91 accuracy and 0.90 recall value by using cross fold validation model on 10 attributes.

PCOS has become a common reproductive endocrinopathies disorder. Various studies contructed models on the based of gene biomarkers. [Xie et al., 2020] implemented computational model by joining the two machine learning models namely Artificial Neural Network and Random Forest to detect the gene biomarkers and make diagnostic model. Data were collected from the database of Gene Expression Omnibus in which 57 were Normal and 76 were PCOS samples. Moreover, 5 datasets were used, including 1 data of screening differently gene expression, 2 train datasets and 2 validation datasets. At first, In DEG's 12 key genes were detected on the basis of Random Forest to be significant for Normal and PCOS samples. Furthermore, these key genes weights were calculated by using Artificial Neural Network having RNN-seq train dataset and microarray. Moreover, the detected models for these kind of 2 datasets were made and call as neuralPCOS. At last, 2 validation dataset were utilized to test and contrast the execution of neuralPCOS with other 2 gene marker datasets by area under curve. 0.72 accuracy were achieved by the model in microarray dataset, in RNN-seq 0.64 respectively.

Use of sonography has also played a role in the detection of PCOS and care of infertility patients. Ultrasound images gives important information about ovary like high range of follicles, size of follicles response to imbalance of hormones, and type of cysts. [Madhumitha et al., 2021] described image segmentation that bring forth more details related to Region of interest in ultrasound images and also detects the background and object from images. Thus, segmentation of sonography images can be difficult to due to noise, so by mixing the photograph with morphological operations detection of follicles can be form. Machine learning algorithms like KNN, SVM, and Logistic Regression were implemented considering the each and every specification of Normal and PCOS ovaries. All three classifiers were combined and hybrid model were made and 0.98 accuracy were achieved.

PCOS is considered a medical disorder which leads to long term complications like irregular periods, high blood pressure, type 2 diabetes and it mainly occur in women. [Khan Inan et al., 2021] described probabilistic recognition abilities of ensemble-based gradient boosting algorithms especially in the area of medical domain. XG Boost, SVM, KNN, NB, RF, MLP and AdaB were proposed in this paper. To effectively support the performance data were re-sampled by using the SMOTE and ENN, to solve outliers and class imbalance issues. Moreover, utilizing the admired statistical correlation models like Chi-Square Test and ANOVA Test, 23 important clinical and metabolic parameters were identified which will classify PCOS conditions. Data were collected from KAGGLE repository and findings shows that Extreme Gradient Boosting came out to be best performer than other classifier 0.96 score of 10 Fold Cross-validation accuracy and 0.98 Recall. Below table shows the summary table of the relevant related work related to PCOS (see Table 1). Many work has been carried out to detect the PCOS using machine learning and deep learning algorithms. Both of the techniques has shown good results The use of Hybrid Extreme Gradient Boosting with Random Forest (XGBRF) ensemble method and CatBoost Model resulted successful on both Glioma segmentation and online transaction detection[Bhatele and Bhadauria, 2020][Li et al., 2020]. The results on Glioma segmentation and online transaction detection using Hybrid XGBRF and CatBoost are remarkable. Implementing Hybrid classifier XGBRF and CatBoost to detect the PCOS at an early stage is the novelty of this paper as there is no work done for PCOS detection using Hybrid machine learning models. After the related work, the paper is inspired to create the Hybrid model framework by combining XGBRF and use of CatBoost.

Table 1: Summary details of the related work for PCOS detection

Author(s)	**Objectives**	**Research Design**	**Results**
Lawrence et al. [2007]	Follicle Segmentation, Use of Stereology,Storing follicles as Feature vectors and classify them into two categories.	Measurement of Length and Width of each Follicle by implementing LDC,KNN, SVM.	All three classifiers performed well with an accuracy of 0.93, 0.91 , 0.91, along LDC outperformed than KNN and SVM
Mehrotra et al. [2011]	Clinical and Metabolic feature formulation as feature vector and statistical feature selection based on two sample t-test.	Classification of features based on Logistic Regression and Bayesian Classifier.	0.93 accuracy were achieved by Bayesian and 0.91 by Logistic Regression.
Rihana et al. [2013]	Cysts detection and classification in ovary ultrasound images based on geometrical features of the cysts texture.	Image pre-processing, Feature extraction, SVM classifier and Validation were used by ROC.	0.90 accuracy were achieved and cysts were detected in ovary ultrasound images
Kumar and Srinivasan [2014]	Identification of small Follicles and fast segmentation using active contour edge model.	Chan Vase (c-v), Active Contour model and Improved Chan Vase model(icv) were used.	ICV algorithm got better results than CV algorithm on despeckled images. It can also be used to check the follicle is real or fake.

Purnama et al. [2015]	Detection of Follicles based on USG images by using the binary follicle images, feature extraction and segmentation.	Three classification scenarios were designed Neural Network - LVQ, KNN - Euclidean distance, SVM - RBF kernel.	On C=40 SVM-RBF kernel achieved 0.82 accuracy and on K=5 KNN achieved 0.78 accuracy.
Cahyono et al. [2017]	Checking the number and diameter of Follicles on ultrasound images.	Feature extraction were done automatically using CNN.	CNN provided best performance with 1 F1-Score and 0.76 5-Fold cross-validation.
Denny et al. [2019]	To overcome the time and cost involved in various clinical tests and ovary scanning.	PCOS features transformed with PCA were done by many machine learning algorithms like KNN, SVM, RF, CART.	The best and accurate model for the PCOS detection came out Random Forest with 0.89 accuracy.
Bharati et al. [2020]	Detection of PCOS at an early stage based on feature selection.	Holdout and cross validation were applied on dataset and algorithms like gradient boosting, RF and LR, RFLR.	RFLR proven out to be the best model for detection of PCOS with 0.91 accuracy and 0.90 recall using 40-fold cross validation.
Madhumitha et al. [2021]	Details of the ovary like large range of follicles, type of cysts, folicle size, using image segmentation	Based on pre-processing and morphological operations SVM, KNN and Logistic Regression were used.	All three algorithms were combined and hybrid model were made and 0.98 accuracy were achieved.
Khan Inan et al. [2021]	Conducting a probabilistic approach to select statistically relevant features which contribute to PCOS instances.	SMOTE, ENN and ANOVA Test, Chi-Square Test were used to identify important features. Classififers such as XG Boost, SVM, KNN, NB, MLP, RF, AdaB were used.	XG Boost outperformed all other classifiers with 0.96 accuracy and 0.98 Recall.

3 Methodology

Manual detection of PCOS is very time consuming. In this paper, proposing a new automated system can be a game changer for early detection of PCOS. This study focuses on implementation of this proposed automated system to help medical expertise in PCOS detection and further helping in the treatment of the PCOS patients. As mentioned in the literature work, hybrid approach is very much effective in detection of PCOS. In this research, a novel approach is proposed that is detection of PCOS using hybrid machine learning model and CatBoost model. We implemented baseline approach that is Gradient Boosting, Logistic Regression, Random Forest, Hybrid Random Forest and Logistic Regression to detect the PCOS. In Figure 2, we illustrate the work flow of research methodology which comprises the analogous steps which will be carried out during this research. As clear from the Figure 2, first we collect the PCOS data. Secondly, data preparation takes place in which we clean the data first followed by dropping unwanted columns. Further, feature selection is implemented were important features will be selected to detect the PCOS at an early stage. Next, data splitting and unsampling will take place by using the SMOTE function. Followed to this, we build the model and apply the machine learning algorithms which we will evaluate based on results matrices. At last models will be compared and novelty will be justified for this research work.

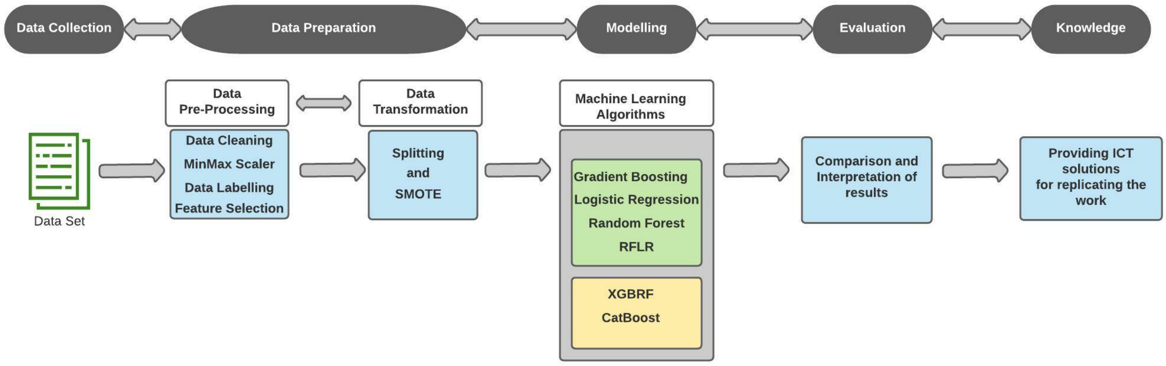


Figure 2: Research Methodology for detecting the PCOS

3.1 Data Collection

We performed research on a public dataset published in Kaggle repository by prasoon kottarathil named as Polycystic ovary syndrome (PCOS) [1]. In this dataset, the data publisher has changed the patients file number for privacy issues. Dataset has 45 attributes based on clinical and metabolic parameters which will help us to detect the PCOS. Data is collected from 10 different hospitals in Kerala, India. The dataset has important attributes like Follicle-stimulating hormone(FSH), Luteinizing hormone(LH), Cycle Length, Cycle, Endometrium, Follicle Number, BMI etc. Because of the 45 attributes the challenge here is to select top 10 highest ranking attribute for the early detection of PCOS.

3.2 Data Preparation

The data obtained in data collection is in csv format which will be used in python. For this, anaconda distribution packages, Jupiter notebook, Scikit-learn library, Matplotlib, are used in the python development. In data preparation, it was important to check the missing values first then removal of unwanted columns which we don't needed in modelling phase. In the initial pre processing correlation were checked between attributes by using the corr plot. Before building the model we removed the unwanted attributes so that feature selection can be applied and significant features will be selected by using SelectKbest and chi2. The dataset contains the 541 women patients in which 177 are those patients suffering from PCOS. Further, we implemented the data pre-processing process as follows:-

- **Data Cleaning:-** It is the process of identifying and removing of imperfect data from the dataset and then replacing the data by modifying or deleting the coarse data. In our paper the dataset was checked for missing values first by using the Pandas and Numpy[2]. As the there were zero misiing values the unwanted attributes were removed by using the drop() function so that new dataframe will be created for further process.

[1] https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[2] https://realpython.com/python-data-cleaning-numpy-pandas/

- **MinMaxScaler** It is simple method used for recalling of the attributes and removal of median and scaled in data between the range of 0 and 1 or -1 and 1. For this selection of target range depend on the nature of the dataset. In this research MinMax Scaler is used as

$$p' = \frac{p - p_{min}}{p_{max} - p_{min}} \tag{1}$$

  Here, P represents the primal value of the attribute, Pmax is the maximum value of P and Pmin is the minimum value of P. P' represents the normalized value of P. MinMax Scaler[3] mainly here transform attributes by scaling every attribute to range. Standard deviation of attributes is low because of scaling.

- **Data Labelling** It is a method of allocating the enlightening tags to subsets of data. Labelled data is significant for the performance of the machine learning algorithms. In this work non-numerical categorical features are transformed into numerical values[4].

- **Feature Selection** It is a process of selecting the significant and informative features from vast number of features. This can bring out preferable interpretation of multiple classes. Taking unrelated features can decrease the classification accuracy of models. The advantage of feature selection is that it decrease the over-fitting and improve the classification accuracy.

  A filtering based approach Univariate feature selection is one form of the feature selection approach. In this approach features are independently selected in regard's with the dependent variable. every feature is scored separately and selected based on higher ranks or high scores. In this research, Univariate feature selection[5] is taken for selecting the important features and finding their scores based on selecting k features from the PCOS dataset, Algorithm 1 is used.

- **Data Transformation** The data transformation is the procedure where data is extracted based on model requirement. The dataset collected contain 541 women's patient in which 177 are PCOS patients and a .csv (comma- separated values). contains the attribute name and its category (0 - for Normal women and 1 - women having PCOS). After selecting the important features we created a new dataframe which contains only 10 features. We performed data split by using the sci-kit learn library available in python.

  Splitting:- The data was spilt into train and test set using train_test_split() function[6]. The split is executed by splitting ratio and random state on machine learning models. The split ratio is 70% on train data and 25% on test data with random sate = 27. For dissociation the train and test data samples cross validation model is also applied using cross_val_score() function[7] available in scikit-learn library.

  SMOTE:- Stands for Synthetic Minority Oversampling Technique. It is an oversampling method in which we generate synthetic samples from minority class. syn-

[3] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

[4] https://www.datalabeler.com/how-to-label-data-for-machine-learning-in-python/

[5] https://github.com/solegalli/feature-selection-for-machine-learning/blob/master/05-Filter-Statistical-Tests/05.3-Univariate-selection.ipynb

[6] https://www.pythonforbeginners.com/dictionary/python-split

[7] https://scikit-learn.org/stable/modules/cross_validation.html

thetically balanced class or nearly balanced class training set is obtained by SMOTE which in-turn is used to train the classifier. It first select a minority class a taken as random instance then finds its nearest neighbour of minority class k. Then, the synthetic instance is developed by selecting one of the nearest neighbours k at random instance b and later connect a and b to create a line segment in the feature space. It is generated by the convex combination of these two instances a and b [Blagus and Lusa, 2012]. In this work, imblearn packages is used for SMOTE() function having Random state = 23 with class minority were applied on both training set and test set[8]

3.3 Modelling

In this step nine different models is used on the pre-processed data. We implemented machine algorithms such as Gradient Boosting, Logistic Regression, Random Forest, Hybrid Random Forest and Logistic Regression, SVM, Decision Tree, Multi-layer Perceptron as baseline approaches on the pre-processed PCOS dataset. Hybrid Extreme Boosting and Random Forest(XGBRF) and CatBoost is the novelty of this paper for detecting PCOS.

3.3.1 Baseline Approach: Machine Learning Algorithms

The related work proved that machine learning models and hybrid model have been effective to detect the PCOS [Bharati et al., 2020], [Khan Inan et al., 2021]. Thus, we used these baseline approach's first in this research. These models are used on trained model on PCOS data. The pre-trained methods are applied on 10 features to classify the Normal women or PCOS women. These are already trained on target variable PCOS(Y/N) to detect the PCOS. Figure 3 show the concept of machine learning models.

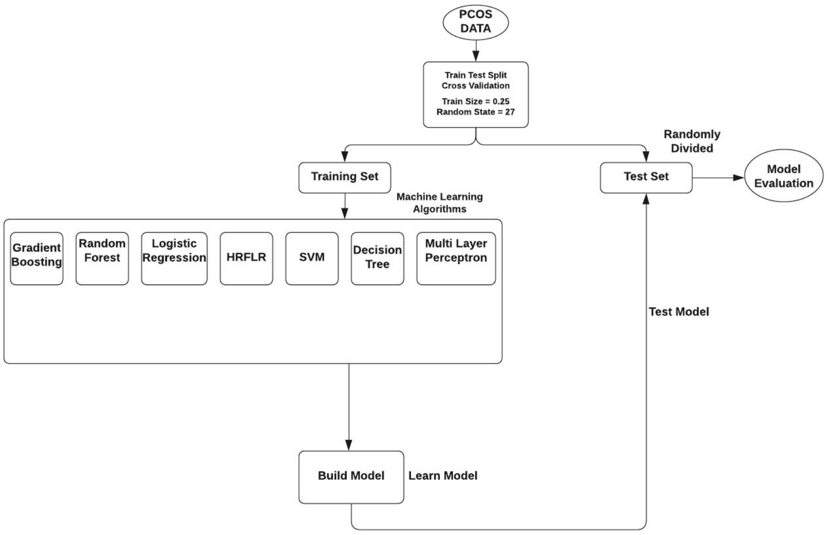


Figure 3: Pre-trained models (Machine learning algorithms) implemented in this research[Bharati et al., 2020],[Khan Inan et al., 2021]

[8] `https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html`

- **Gradient Boosting:-** We select the Gradient Boosting model in this research. This model re-sample the training data to give us the useful information. In Gradient Boosting, additional methods are integrated to decrease the definite sum of losses. The loss function control the deviation amount from predicted value. The approach of on-wards stage insert new methods sequentially without changing the parameters. In 2002, Friedman introduced a new modified version of GBM which improved model method quality based on fixed size of regression tress [Deb et al., 2020] which is as follows:

$$\mathrm{f_m(x) = f_{m-1}(x) + J.} \sum_{\mathrm{j=1}}^{\mathrm{J}} \rho_{\mathrm{jm}} \mathrm{I(x \in R_{jm})} \tag{2}$$

  Here J denotes the learning rate, C is the maximum interaction, and M is the number of trees. J, C, and M controls the performance of GBM.

- **Random Forest:-** This model was developed by Breiman in 2001 [Deb et al., 2020]. It initiates both procedure of random feature selection and bagging idea. The construction of Bagging method is done to calculate the distribution of estimator based on sampling and with replacing from real dataset. In bagging model, n sample size is taken from training data, bagging model produce new data using the sampling and replacing the actual dataset with n sample size. On the other hand, procedure of random feature selection authorize random feature subsets in every node during splitting in the trees in such a way that diversity of base method may be observed. Both, Bagging and Random feature selection improves accuracy during prediction. The variance of Random Forest is calculated as follows:

$$\rho\sigma^2 + \frac{1-\rho}{\mathrm{K}}\sigma^2 \tag{3}$$

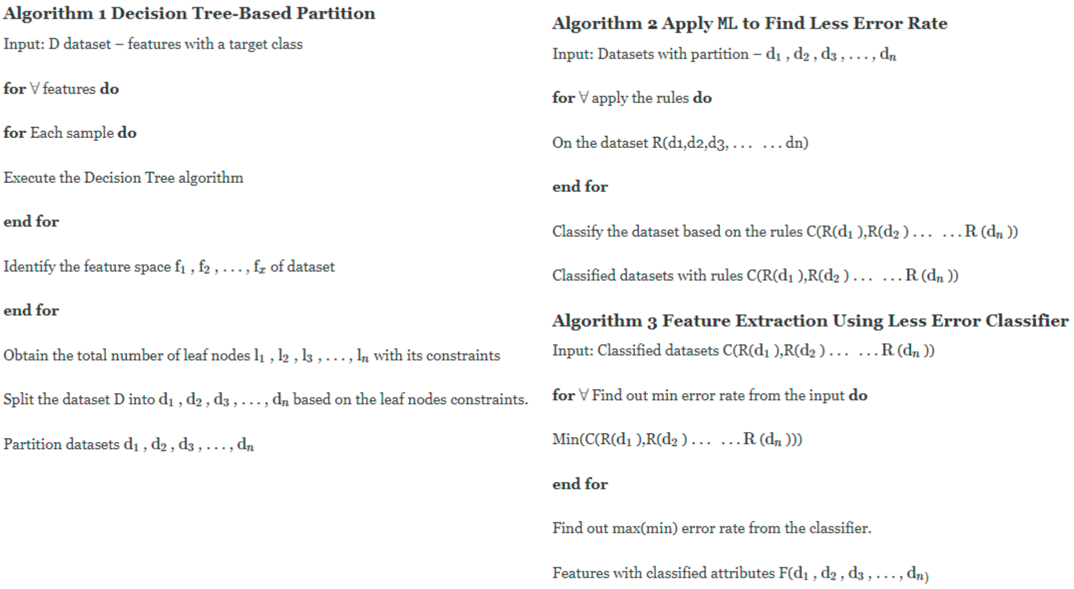
  Here σ^2 denotes tree variance, ρ denotes correlation between trees, K represents total trees.

- **Logistic Regression:-** This model regulate the relationship among independent variables and binary outcomes based on probability as forecast value of dependent variable. In this paper, every feature is tested and allocated a probability which is used to classify the PCOS as Normal women or PCOS Women. If the probability is higher than threshold it is PCOS women else Normal women [Rushin et al., 2017]. The equation of Logistic Regression is as follows:

$$\Pi(\mathrm{x}) = 1/1 + e^{-\mathrm{y}} \tag{4}$$

  Here y represents coefficients of variable and e is Euler's number. If $\Pi(\mathrm{x})$ is higher than 0.5 then it is considered as Home win else as Away win.

- **Hybrid Model of Random Forest and Logistic Regression(HRFLR):-** In this research, we created a hybrid model using a Random Forest and Logistic Regression algorithm. Combined models works on Random forest probabilities. These probabilities from RF are added to training dataset and fed to the logistic regression algorithm. Similarly, Logistic Regression probabilities are fed into the test datset. At last, values are predicted [Mohan et al., 2019]. Implementing HRFLM using the pre-trained data designed for PCOS detection. The input given by women patient's helped us to detect the PCOS. Detection is based on classification as binary prediction type such as 0 is Normal women and 1 is PCOS women. The development were designed using TkInter in python. Below Figure 4 is the process of 3 algorithms used in the construction of HRFLR:



```
Algorithm 1 Decision Tree-Based Partition
Input: D dataset – features with a target class
for ∀ features do
for Each sample do
Execute the Decision Tree algorithm
end for
Identify the feature space $f_1, f_2, \dots, f_x$ of dataset
end for
Obtain the total number of leaf nodes $l_1, l_2, l_3, \dots, l_n$ with its constraints
Split the dataset D into $d_1, d_2, d_3, \dots, d_n$ based on the leaf nodes constraints.
Partition datasets $d_1, d_2, d_3, \dots, d_n$
```

```
Algorithm 2 Apply ML to Find Less Error Rate
Input: Datasets with partition – $d_1, d_2, d_3, \dots, d_n$
for ∀ apply the rules do
On the dataset R(d1,d2,d3, . . .   . . . dn)
end for
Classify the dataset based on the rules $C(R(d_1), R(d_2) \dots \dots R(d_n))$
Classified datasets with rules $C(R(d_1), R(d_2) \dots \dots R(d_n))$
```

```
Algorithm 3 Feature Extraction Using Less Error Classifier
Input: Classified datasets $C(R(d_1), R(d_2) \dots \dots R(d_n))$
for ∀ Find out min error rate from the input do
$Min(C(R(d_1), R(d_2) \dots \dots R(d_n)))$
end for
Find out max(min) error rate from the classifier.
Features with classified attributes $F(d_1, d_2, d_3, \dots, d_n)$
```

Figure 4: Three Algorithms for constructing HRFLR [Mohan et al., 2019]

Algorithm 4 Apply Classifier on Extracted Features
Apply the hybrid method based on the error rate

$$\sum\nolimits_0^n F(n) = d + m_1 x_1 + m_2 x_2 + \dots. + m_n x_n \tag{5}$$

$$\sum\nolimits_0^n F(0) = Gain + \sum\nolimits_0^n w_i x_i \tag{6}$$

- **Support Vector Machine:-** Support Vector Machine is totally based on VC dimension theory and structural risk minimum principle. The idea here is searching hyper plane optimum which satisfies classification request then uses this model to made the margin of dissociation alongside the hyper-plane maximum optimum while make sure the correct classification accuracy [Zhang and Chen, 2009]. Based on kernel trick, SVM is as follows:

$$S(x) = \text{sign}(\sum_k \alpha_k y_k k(x_k, x) + b) \tag{7}$$

Here $K_{\mathbf{x}} \in \mathrm{R}^N$ are support vectors and $k(x_k, x)$ is the kernel function. In this work, radial bases kernel function is used. The decision SVM function is formed on dot product of input feature vector having support vectors which means it has no dimension requirements of the feature vector.

- **Decision Trees:-** In the development of Decision tree process [AL-Dlaeen and Alashqur, 2014], classifier learns from labelled class trained tuples. In DT structure, every non-leaf node act for test on features and every branch act for test outcome.

Oppositely, if test outcome that consists of tuples which belong to single class, then leaf node is add on to end that branch and indicates that tuples which satisfies all tests across the path from root to leaf node belong to only single class. A key thing here is the use of feature selection measure which is to select the feature that foremost partitions the tuples into well defined classes. We applied three step process, in first step the information is computed to classify tuple into the PCOS dataset. This step distributes the tuples into the classes. Second step considers the features values since it tries to calculate the impact of every feature on the class. Below three equation were used in this process:
Equation 8 calculate the information which we used to classify the tuples in PCOS dataset.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{8}$$

Equation 9 is used to get the information after using the important features to split the PCOS dataset into v partitions to claddify the PCOS dataset.

$$Info_X(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j) \tag{9}$$

Equation 10 is the last step where information gained by every feature by subtracting the equation 9 result from the equation 8 result

$$Gain(X) = Info(D) - Info_X(D) \tag{10}$$

Equation 8 is applied on the start of the dataset than equation 9 and 10 are applied to every features. The process continues to every level of constructed tree until we reach to leaf nodes which will represent the pure classes.

- **Multi-Layered Perceptron:-** MLP algorithm is like pattern recognition algorithm [Singh and Sachan, 2014]. In MLP, one neuron in a network has the ability to link with other 10,000 neurons at one time to produce new information. Neurons are interconnected with each other through links known as synapse. Neurons receives information from other neurons as actual data. Figure 5 is the architecture of MLP. There are three layers input, hidden and output layer. Each layer is consists of neurons with weight associated for further pre-processing.

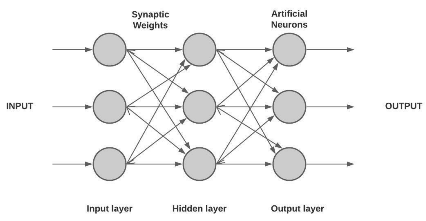


Figure 5: Basic architecture of MLP [Singh and Sachan, 2014]

The Post Synaptic Potential is calculated by the sum of input weights subtracted from neuron threshold. Network produces its output by flow of activation signal throughout the transfer function. In MLP, each input perform a weighted sum and passes them to activation signal by transfer function to produce output. Hyperbolic and logistic sigmoid functions are the most common activation function used in MLP. In our work, for the purpose of detection we used the hyperbolic tangent sigmoid function.

3.3.2 Novel Approach: Hybrid XGBRF classifier and CatBoost Model

Detecting a disease on the basis of clinical and metabolic parameters are vital in medical field. Both XGBRF and CatBoost classifier models can do the task of handling labelled data. Both used the feature selection algorithm which comprises of selecting the significant features and does the detection of PCOS. XGBRF and CatBoost proved to be unique solution in medical field [Bhatele and Bhadauria, 2020] [Li et al., 2020] and beneficial in detecting the patterns in medical data. Hence, we used XGBRF and CatBoost as novel approach in this research. Figure 6 shows the concept of XGBRF and CatBoost implementation in this research. After importing the dataset feature selection is applied in which significant feature are selected followed by splitting the train and test data. Both XGBRF and Catboost are implemented to get the classification results.

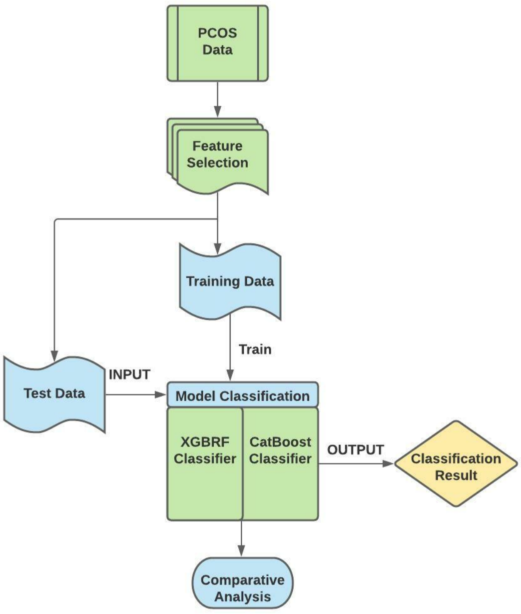


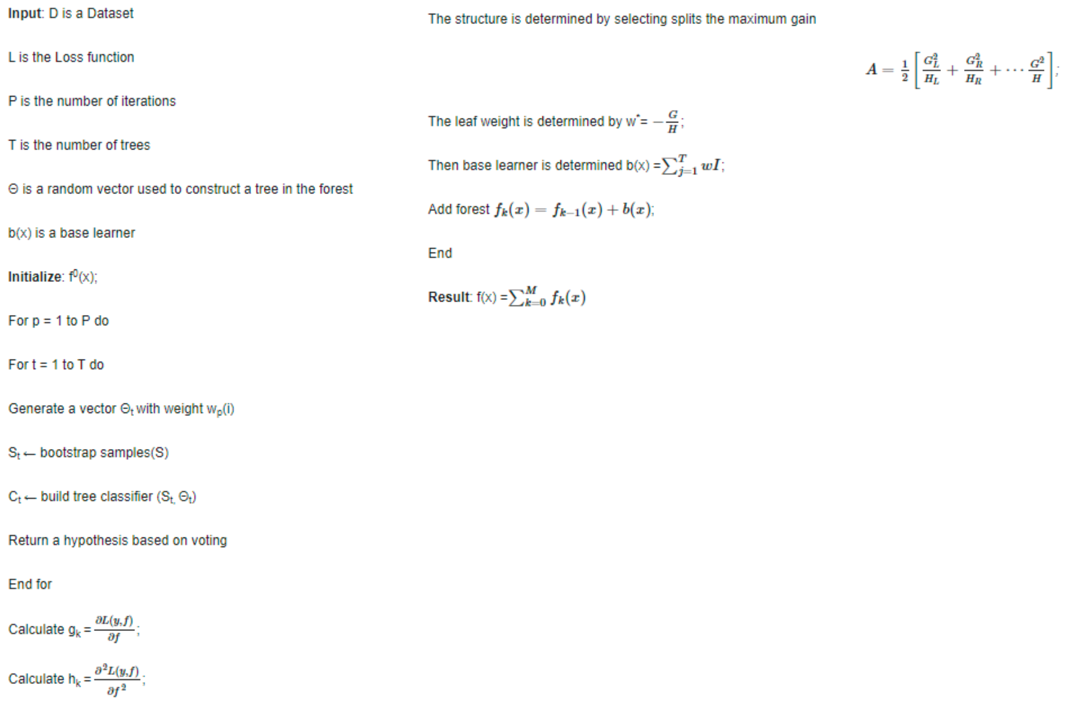
Figure 6: Concept of XGBRF and CatBoost model [Bhatele and Bhadauria, 2020] [Li et al., 2020]

- **XGBRF classifier:** In this research, XGBoost with Random forest (XGBRF) hybrid ensemble method is used for classification of PCOS. Both are popular decision tree based models. XGBoost is an illustration of boosting algorithm while as Random Forest is illustration of begging algorithm. So they can be combined and entitle as hybrid ensemble learning model [Bhatele and Bhadauria, 2020]. XGBRF is considered as modified version of XGBoost classifier. In this work, decision tree are simply replaced by Random Forest as base estimator. Random Forest are trained in XGBRF rather than gradient boosting decision trees and high accuracy are rendered on various datasets. The best advantage of XGBRF is that it harness both XGBoost and RF to get the stability, accuracy and to overcome the over-fitting problem. Here, input is taken to the boosting model on a pre-trained data consists of significant features with (0,1) as class label. The class 0 represents Normal women and 1 is PCOS women. As we know that XGBoost is end-to-end scalable tree boosting structure, that involve better tree method into the present classification model and thus predict the outcomes. This approach is executed with the help of python.

Random Forest is magnified begging model that offer advantage such as handling of numerous input variable and feature importance calculation for classification to check it is noise or robust, and outlier detection [Bhatele and Bhadauria, 2020]. Rf is a fusion of tree structure(classifiers) in which every classifier depend on independent values of sampled random vector and with the same distribution of all trees in the forest. The following points are required for the construction of every tree:

1. From bootstrap samples forest trees are constructed. The sample size of bootstrap sample which is N are same as the N number of samples taken in the training data.
2. If R are the features than number r<R are taken as if every node r features are back out of R feature at random.
3. Each tree is enlarged to the great extend without pruning.

Below Figure 7 shows the algorithm of XGBRF hybrid ensemble method:



```
Input: D is a Dataset
L is the Loss function
P is the number of iterations
T is the number of trees
Θ is a random vector used to construct a tree in the forest
b(x) is a base learner
Initialize: $f^0(x)$;
For p = 1 to P do
For t = 1 to T do
Generate a vector $\Theta_t$ with weight $w_p(i)$
$S_t \leftarrow$ bootstrap samples(S)
$C_t \leftarrow$ build tree classifier ($S_{t,}$ $\Theta_t$)
Return a hypothesis based on voting
End for
Calculate $g_k = \frac{\partial L(y,f)}{\partial f}$;
Calculate $h_k = \frac{\partial^2 L(y,f)}{\partial f^2}$;
The structure is determined by selecting splits the maximum gain
$A = \frac{1}{2}\left[\frac{G_L^2}{H_L} + \frac{G_R^2}{H_R} + \cdots \frac{G^2}{H}\right]$;
The leaf weight is determined by $w^* = -\frac{G}{H}$;
Then base learner is determined $b(x) = \sum_{j=1}^{T} wI$;
Add forest $f_k(x) = f_{k-1}(x) + b(x)$;
End
Result: $f(x) = \sum_{k=0}^{M} f_k(x)$
```

Figure 7: Algorithm of XGBRF Hybrid ensemble method [Bhatele and Bhadauria, 2020]

- **CatBoost Model:** CatBoost is an Machine learning model which uses gradient boosting on decision trees. It uses a schema of estimating leaf values when choosing a tree structure, which helps to overcome the over-fitting problem. It has four principal merits, first one is creative model for computing the categorical features which means there is no need for processing features on your own - it is constructed out of the box. For a dataset having categorical features results like accuracy if greater than other algorithms [Li et al., 2020]. Implementation of direct boosting, a permutation - driven different to other classic boosting models. On small datasets gradient boosting causes over-fitting while there is special modification based on CatBoost for such cases. CatBoost makes it fast and easy use of GPU implementation training and at last it produces missing value great support visualisation. CatBoost tune key parameters based on bayesian optimization and grid search like learning rate, and types of boosting etc. In our work lerning rate is 0.1 and 199 iterations have been taken. Table 2 shows the important parameters to create our model while other parameters remains default.

Table 2: Catboost parameters table.

Parameter	**Value**
iterations	199
learning_rate	0.1
depth	12
12_leaf_reg	1
rsm	0.98
loss_function	Logless
random_seed	42
Boosting_type	gbdt
eval_metric	Confusion matrix, ROC-AUC, Coss validation accuracy

Hence, we can see that XGBRF and CatBoost model produces better results for Giloma Segmentation and online transaction detection [Bhatele and Bhadauria, 2020] [Li et al., 2020] so we will implement both of them on detection of PCOS in order to prove novelty of this research.

3.4 Evaluation Metrices

Confusion matrix is implemented to evaluate our models and calculate the evaluation matrices like Accuracy, Precision, Recall and F1-score. Apart from this ROC-AUC score and Cross validation accuracy were also calculated. The following evaluation matrices are as follows:

1. **Accuracy:** Accuracy is a union of precision and trueness. Better accuracy means better precision and truesness. It is reported and as uncertainty. Classification Accuracy and Error is calculated as (see Equation 11 and 12)

$$Accuracy = \frac{CorrectPredictions}{TotalPredictions} \tag{11}$$

$$Error = \frac{IncorrectPredictions}{TotalPredictions} \tag{12}$$

2. **Precision:** In classification of PCOS, precision is the fraction of instances allocated to positive class which belong to positive class(see Equation 13).

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

3. **Recall:** In classification of PCOS, recall summarizes how accurately the psotive class was predicted which means fraction of total amount of related instances which were actually retrieved(see Equation 14)

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

4. **F1-Score:** In classification of PCOS, F1-Score is the combination of both precision and recall which seeks to balance analysis. Further, at 1 it indicated the perfect precision and recall value while as at 0 it indicate worst precision value(see Equation 15)

$$F1 - Score = \frac{2Precision \times Recall}{Precision + Recall} \tag{15}$$

   Here, TP(True Positive) = means PCOS women are detected in PCOS dataset and correctly classified as PCOS women

   TN(True Negative) = means Normal women are detected in PCOS datset and correctly classified as Normal women

   FP(False Positive) = means PCOS women are detected in PCOS dataset and wrongly classified as Normal women

   FN(False Negative) = means Normal women are detected in PCOS dataset and wrongly classified as PCOS women

5. **ROC-AUC:** In classification of PCOS, ROC_AUC summarises the behaviour of the model. It is calculated as false positive rate and true positive rate for predictions by various thresholds. Every threshold is a point on plot and points are connected to produce a curve. Classifier which has no skill will develop a diagonal line from bottom left to top right. 0.5 is a score of no skill classifier and 1 is the score of perfect classifier(see Equation 16 and 17)

$$Truepositiverate = \frac{TruePositive}{TruePositive + FalseNegative} \tag{16}$$

$$Falsepositiverate = \frac{FalsePositive}{FalsePositive + TrueNegative} \tag{17}$$

6. **Cross validation accuracy:** This approach includes randomly dividing the set of instances into k groups, or folds with equal size. In our work we have take k = 10,20,30,40. and for better practice standard deviation is also calculated for skill score variance.

   The evaluation of implemented models are further discussed in next section.

4 Results and Discussions

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 models, 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 uni-variate 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 their is the ratio of Follicle-stimulating hormone (FSH) Luteinizing hormone (LH) represneted as FSH/LH. Follicle No.(R) is on number third respectively. AMU is the Anti-Mullerian hormone and BMI is Body Mass Index. Now we will explain the evaluation of our classifiers in next subsections.

4.1 Baseline Approach

4.1.1 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 HRFLR 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.

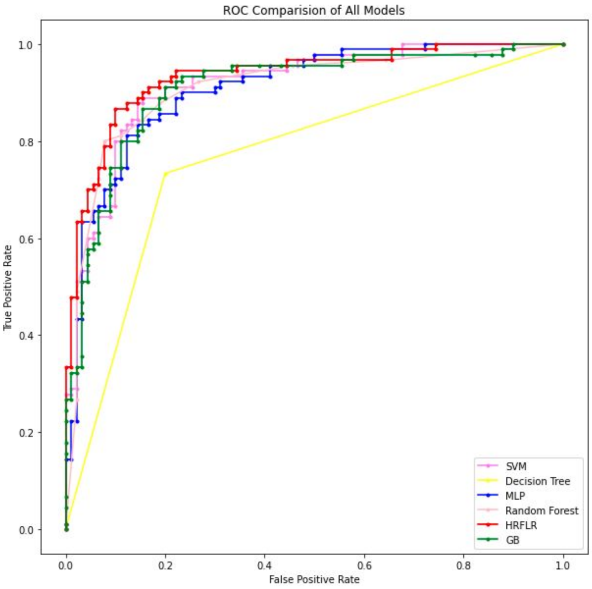


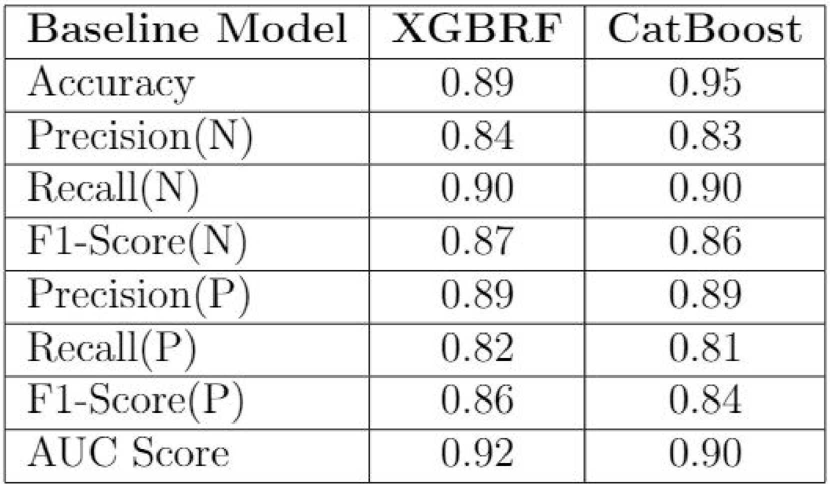
Figure 8: ROC comparison of all models

This ROC curve shows the trade off 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.

4.2 Novel Aprroach

4.3 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 whoc truely 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 are 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.

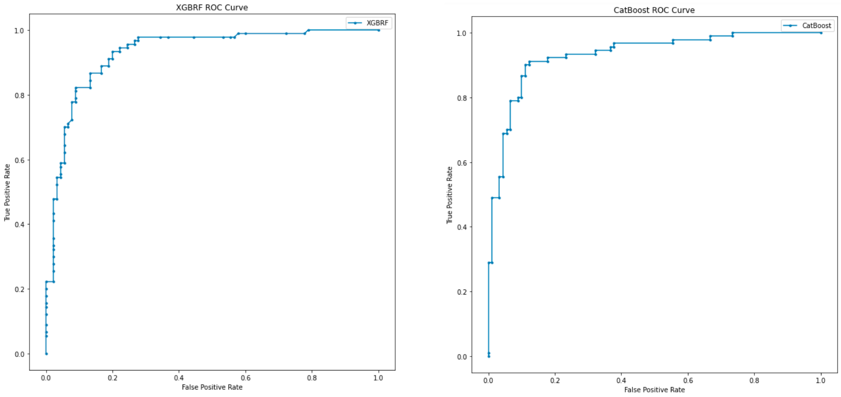


Figure 10: ROC of XGBRF and CatBoost

In Figure 10 it is quite evident from ROC curve plot of both classifiers are on the top left corner which means there 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.

4.4 Discussion

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 [Bhatele and Bhadauria, 2020] [Li et al., 2020].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 dont have PCOS disease and FN=8, that means they predicted no but they have the diesease. 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 a important improvement as they successfully detected the true PCOS. A 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 giloma segmentation and 0.98 for online transaction detection [Bhatele and Bhadauria, 2020] [Li et al., 2020], which justify our reseacch 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
XGBRF	10	0.83	0.4
	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**

5 Conclusion and Future Work

Detection PCOS at an early stage enhance the early treatment of the patient's. 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 thousands. Furthermore, a new hybrid algorithms 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.

Acknowledgements: I would like to thank Dr. Rashmi Gupta and National college of ireland for providing valuable support in terms of knowledge, guidance, and validate this research work.

References

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.

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.

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.

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.

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.

Chandan Chakraborty, Palak Mehrotra, Biswanath Ghoshdastidar, Sudarshan Ghoshdastidar, and Kakoli Ghoshdastidar. Automated ovarian follicle recognition for polycystic ovary syndrome. In *2011 International Conference on Image Information Processing*, pages 1–4, 2011. doi:10.1109/ICIIP.2011.6108968.

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.

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.

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.

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.

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 Computing and Communication Workshop and Conference (CCWC)*, pages 1046–1050, 2021. doi:10.1109/CCWC51732.2021.9375994.

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.

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.

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

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

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.

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.

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

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.

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

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

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.

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.

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

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.

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.

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

Pilar Vigil, Manuel E. Cortés, Ana Zúñiga, Jessica Riquelme, and Francisco Ceric. Scanning 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.

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.

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.