JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

COMPARATIVE ANALYSIS OF CLASSIFIERS FOR PREDICTING POLYCYSTIC OVARY SYNDROME USING DEEP LEARNING MODELS

[1]Renju K, [2]Pavithra B
[1]Assistant Professor, [2]Student
[1]Department of Computer Science of 1st Author,
[1]Mount Carmel College, Autonomous, Bengaluru, India

Abstract: Polycystic Ovary Syndrome (PCOS) is a condition diagnosed commonly in young women in their reproductive age. PCOS will cause prolonged or irregular menstrual periods because of the variations in androgen levels. Early detection of PCOS will reduce the risk of weight gain, mood swings, heart disease and other long-term complications. Researchers have proved that diagnosing PCOS at the early stage and undergoing treatment will reduce the risk factors in order to lead a normal life. The proposed research work focuses on diagnosing PCOS using deep learning techniques with optimal and minimal set of parameters. The paper also discusses about the preprocessing techniques used for dimensionality reduction and perform a comparative analysis of linear and non-linear methods. Artificial Neural Network model is used to predict the result based on different parameters and the result is evaluated with different layers in the network.

IndexTerms **- Polycystic ovary syndrome, Radial Basis Function, Multilayer Perceptron, Deep Learning models, Principal Component analysis, ISOMAP**

I. INTRODUCTION

Polycystic ovary syndrome is frequently encountered endocrinopathy in reproductive aged women. It is a cause for infertility and this lifelong health issue continues beyond the child bearing years. The reproductive organs of women called ovaries produce progesterone and estrogen-hormones that regulate the menstrual cycle, are affected. Ovaries also produce male hormone called androgen and women with this problem are associated with increased risk of hypertension, heart disease, obesity, type 2 diabetes and gynaecological cancer. Some of the common symptoms for PCOS are uncertain or irregular periods, abnormal hair growth, skin darkening, depression and mood swings, high blood pressure, sleep apnea, risk of cancer, obesity and many more [1] . A healthy lifestyle is the cornerstone of treatment for PCOS, which includes maintaining a healthy weight and deep sleep. Treatment for these symptoms is performed individually such as correcting abnormal bleeding in uterus, restoring fertility, improving androgen deficiency like acne, hair loss etc., prevention of heart diseases and diabetes. In this research, deep learning model is proposed that served as the early marker of the polycystic ovary syndrome [2]. Although research is advancing to diagnose PCOS using various machine learning algorithms, there is possibility for improvement in accuracy and precision based on clinical data. The dimension of the dataset is reduced to improve the performance of the model. Various deep learning models such as Multilayer Perceptron (MLP), Artificial Neural Network (ANN), and Radial Basis Function Network (RBFN) are applied and a comparative analysis was performed to find the best performing algorithm.

2. LITERATURE REVIEW

Many research was conducted to diagnose PCOS and a comparative analysis was performed on different classification algorithms that achieved highest accuracy.The research work proposed by Palak Mehrotra *et.al*., incorporates clinical and metabolic parameters to explain a method for detecting PCOS. The technique includes the creation of a feature vector based on clinical and metabolic variables, as well as the selection of statistically significant features for differentiating between normal and PCOS groups using a two-sample t-test. A comparison was made between Bayesian Classifier and Logistic Regression classifier models and it was observed that Bayesian classifier model performed well which gave the highest accuracy of 93.93% [3]. Likewise, another study proposes a strategy for detecting and predicting PCOS in its early stages. SPSS V 22.0 is used to select 8 potential features based on the significance of the 23 features from clinical data. The performance of different classification algorithms such as Logistic Regression, Linear Discriminant Analysis, K-nearest neighbour, Random Forest Classifier, Naïve Bayes Classifier and Support Vector Machine were compared. Random Forest Classifier model gave better performance with accuracy of 89.02% which was addressed by Amsy Denny *et.al*., [4].

Malik Mubasher Hassan *et.al.,* determines differences between 10 selected features and entire attributes are insignificant. They indicate selected features might be beneficial in building a better model for the PCOS dataset. They compare performance of three algorithms such as logistic regression, support vector machine, random forest. Out of this random forest gave best result [5]. Correspondingly research was done on detection of PCOS by using two analytical tools i.e, Python-based open-source Scikit learn version 0.21 and RapidMiner studio version 9.5. Classification algorithm like k-nearest neighbour, SVC, random forest, naïve bayesian, Multilayer perceptron, Bagging Classifier GBOOST was used for comparison. RapidMiner showed 93.12% accuracy by using RF, was addressed by Satish C. R Nandipati *et.al.*, [6]. The research work by Namrata Tanwani, made a comparison between k-nearest neighbour and logistic Regression. Among them logistic regression gave good accuracy of 93% compared to k-nearest neighbour [7]. The model suggested by Priyanka R. Lele stal *et.al.*, considers physical as well as hormonal symptoms as a feature set. Different machine learning models were used, among them K star algorithm performed well compared to other algorithms [8]. The research proposed by Vaidehi Thakre *et.al.*, presents a method that uses an optimum and minimum set of parameters to support in the early identification and prediction of PCOS treatment. Different machine learning models such as Radial Support vector machine, Linear Support vector machine, Random Forest, Logistic Regression, K-Nearest Neighbour and Naïve Bayes were compared. Although the highest accuracy of 90.9% was achieved by Random Forest classifier [9].

The study conducted by Pijush Dutta *et.al.*, focused on applying classification algorithm after handling class imbalance and applying dimensionality reduction technique on the dataset. Class imbalance is handled using Synthetic minority over sampling technique (SMOTE), dimensionality reduction technique is Principal component analysis (PCA). Algorithms used for comparison are logistic regression, Decision Tree, random forest, k-nearest neighbour, support vector machine. Out of them SMOTE based logistic regression outperformed with accuracy 95.05% compared to other ML algorithms [10]. Comparably research was made on handling data outlier issue and to solve class imbalance problem. Algorithm used for the comparison are support vector machine, random forest, k-nearest neighbour, XGBoost, AdaBoost, naïve Bayesian, multilayer perceptron and Class imbalance was handled using SMOTE & ENN (Edited Nearest Neighbour). Out of these XGBoost outperformed with 95.83% accuracy was addressed by Muhammad Sakib Khan Inan etal., [11]. Likewise, another research focused on using hybrid XGBRF and Catboost models also comparing the performance of different algorithms like Gradient Boosting, random forest, logistic regression, support vector machine, decision tree, multilayer perceptron, XGBRF, CatBoost. Although, CatBoost gave better performance with 95.00% which was addressed by Shakoor Ahmad Bhat [12].

The research conducted by Vikas B *et.al.*, proposed deep learning approaches such as convolutional neural networks (CNN), which can be used to diagnosis of PCOS. Transfer learning with fine-tuning and picture augmentation gave accuracy of 98 %, which is improved by 10% over the standard CNN model [13]. The study by R M Dewi *et.al.*, have developed a system to identify PCOS using feature extraction and Neural Networks. CNN was chosen because it is a mix of Hemming Net and The Max Net, allowing data categorization to be done based on the unique characteristics of ultrasound data. Neural Network obtained the maximum accuracy of 80.84 % [14].C. Gopalakrishnan *et.al.*, suggested various image processing techniques to evaluate ultrasound images of the ovary for detecting PCOS. The Canny edge detection technique is used to identify the follicular edges. Scale-Invariant feature transform is used to identify the presence of the condition. A Support Vector Machine is used for data training and categorization [15]. In the study proposed by M Sumathi *et.al.*, CNN-based image processing is used to classify cysts in the dataset. The algorithm can detect cysts in the dataset using segmentation and feature extraction methods. This approach takes certain input ultrasound pictures as train data and then classifies test data to determine if the ovary is damaged and which metrics, such as size, solidity, extension, and perimeter, are affected. The results obtained is 85 % [16].An automated PCOS diagnosis tool would assist to reduce the amount of time spent manually tracking follicles and assessing their geometric properties. The research proposed by Rachana B *et.al.*, uses k-nearest neighbour classifier, the suggested technique was able to obtain classification accuracy of 97 %. The classifier will shorten the time it takes to diagnose PCOS and enhance its accuracy [17].

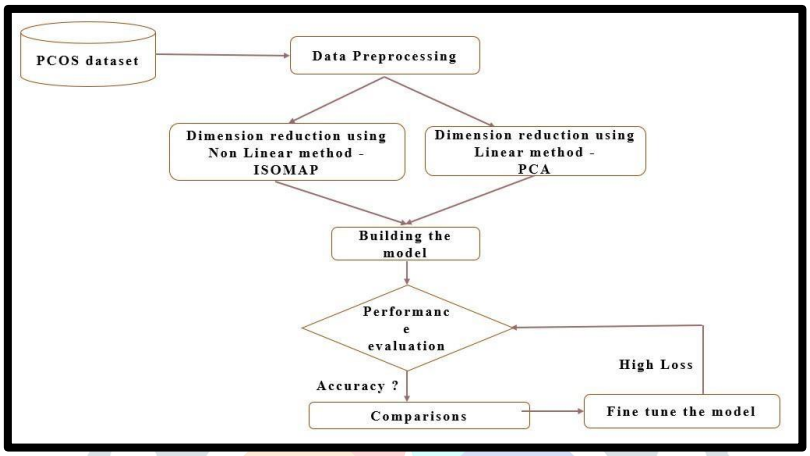
3. METHODOLOGY

The PCOS_infertility dataset is taken from Kaggle repository and contains all physical and clinical parameters to determine PCOS. The data is collected from 10 different hospital across Kerala, India. The dataset contains 514 records having 42 attributes and the patient file number attribute is not considered for data analysis. Eventually the dataset contains 40 attributes and the target variable has two values 0/1 hence it is a binary classification problem. The dataset is obtained and pre-processed using the capping approach to handle outliers in each attribute. Once the outliers are removed, dimension of the dataset is reduced using principal component analysis (PCA) and Isometric (ISOMAP) techniques. Furthermore, various deep learning algorithms are used, and their results were compared to find the best performing algorithm in terms of accuracy and precision.

3.1 WORK FLOW DIAGRAM

A total of 541 records were considered for processing, with 0.70 % in two classes used for the training set and 0.30 % in two classes used for the validation set.

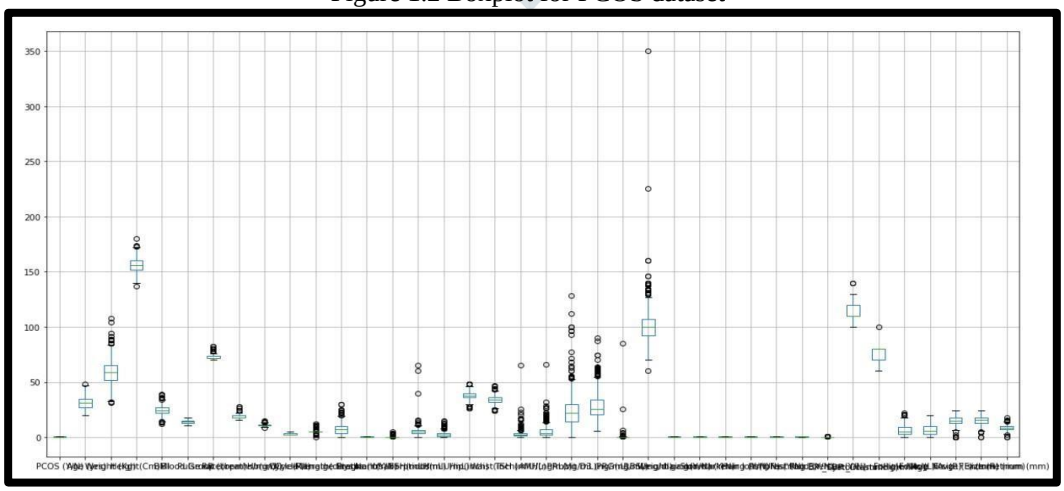
Figure 1.1 Work Flow Diagram



3.3 DATA PRE-PROCESSING

Handling outliers is one of the most crucial tasks in exploratory data analysis. An outlier is a value that has a significant discrepancy from the entire set. Deep learning models are based on the attribute's distribution or range of values. Outliers in the dataset may cause the training data to be misled, resulting in a decrease in algorithm performance. Outliers are less important to handle in big datasets since individual points carry less weight, but addressing outliers in small datasets is significant. Outliers are managed utilising the capping approach since the considered PCOS without infertility dataset has a lesser number of records [18]. As illustrated in fig 1.2, the outliers in the dataset are seen using a boxplot.

Figure 1.2 Boxplot for PCOS dataset



3.3.1 DIMENSIONALITY REDUCTION TECHNIQUE

Principal Component Analysis is a linear method used to decrease the dimension of feature sets in a dataset. It works by detecting patterns in datasets, and establishing correlations between feature. The associated data are then removed by deleting such attributes directly by decreasing the variances. It extracts strong patterns from the input dataset. It usually tries to project the high-dimensional data onto a lower-dimensional surface. If the data is complex, expressing it in a nonlinear way can help us to preserve more information. Principal Component Analysis may not perform efficiently in this instance if the data is not represented in a linear direction. ISOMAP is a manifold learning method that attempts to keep the geodesic distance between samples while lowering dimension.

3.4 MODEL SELECTION

The ANN is the core structure of deep learning models since it identifies the patterns on its own. we have used sklearn's multilayer perceptron classifier to classify the problem. The parameters considered are discussed in table I. The deep learning model is built using the Keras package with a TensorFlow backend. 'Yes': 0, 'No: 1' are the classification classes used. Table I shows the parameters that were taken into account. To determine the best-performing model, we looked at different deep learning algorithms which includes ANN and RBFN. The best performing model is determined by comparing ANN with different layers. The models are also analysed using the confusion matrix. Among the models RBFN was the most successful.

3.5 PARAMETER TUNING

The hyper parameters used in each of our prosed model's classifiers are presented. The hyper parameters are listed in Table I. We used MLP, RBFN, ANN, with PCA and ISOMAP to test our data during the modelling phase. Through a random search, we found the optimal hyperparameter.

Table I: Hyperparameters Used to Build the Network

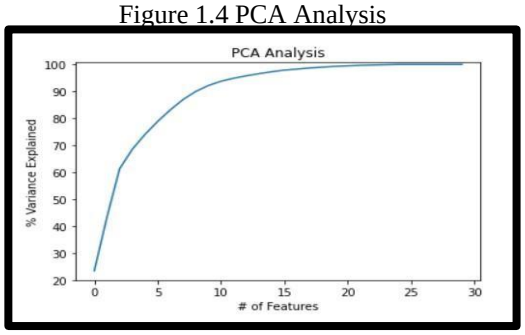
Classifier	Hyper Parameters
ANN- 2Layers	random_state=13,hidden_Layer1(units=20), hidden_layer2=(unit=1),activation=sigmoid, learning_rate=0.0001,loss='binary_crossentropy, optimizer=rmsprop
ANN- 4Layers	random_state=13,hidden_Layer1=(unit=64), hidden_layer2=(unit=32),hidden_layer3=(unit=16), hidden_layer4=(unit=1),activation=sigmoid,learning_rate=0.0001,loss='binary_crossentropy',optimizer=rmsprop
ANN- 6Layers	random_state=13,hidden_Layer1=(unit=64), hidden_layer2=(unit=64),hidden_layer3=(unit=32), hidden_layer4=(unit=16),hidden_layer5=(unit=16),hidden_layer6=(unit=1),activation=sigmoid,learning_rate=0.0001, loss='binary_crossentropy,optimizer=rmsprop
ANN- 8Layers	random_state=13, hidden_Layer1=(unit=32), hidden_layer2=(unit=32), hidden_layer3=(unit=32), hidden_layer4=(unit=32),hidden_layer5=(unit=32),hidden_layer6=(unit=16),hidden_layer7=(unit=16),hidden_layer8=(unit=16),activation=sigmoid,learning_rate=0.0001,loss='binary_crossentropy,optimizer=rmsprop
MLP	random_state=13, hidden_layer_sizes=15, activation='relu', solver='sgd', verbose=5, max_iter=80
RBFN	random_state=13,loss='binary_crossentropy' hidden_layer1=(unit=10), activation='relu', optimizer='Adam'RBFLayer=(Layer=10,dropout=0.3),Outputlayer=(unit=1,activation='sigmoid').

4. EXPERIMENTAL RESULTS

There are 541 instances in the dataset, 364 of women are normal and 177 of women suffer from PCOS. Since the data is highly dimensioned it is difficult to visualize how the data looks like. Hence the dataset consisting of 41 attributes is reduced to 10 attributes using PCA and ISOMAP. The figure 1.4 depicts that 1st principal component explains about 23% of information is

preserved in the data whereas 10 Principal component explains 92% of the information is preserved in the data. Hence 10 principal components is considered in the work to achieve the maximum accuracy.

Figure 1.4 PCA Analysis



The performance of each classification algorithm is based on accuracy. The RBFN with PCA shows the performance with the highest accuracy of 92.571%, followed by ANN (4 Layers) with ISOMAP (90.857%). In comparison of 3 models along with PCA and ISOMAP, the RBFN with PCA shows good precision (91%) and recall (91%) respectively (Refer Fig 1.5 & 1.6).

Fig 1.5 Comparison of Accuracy of ANN

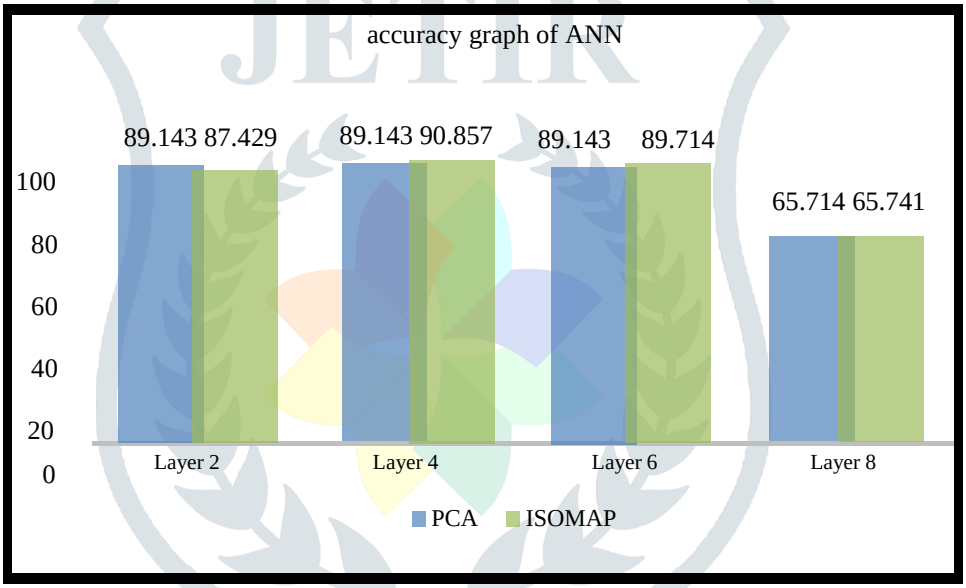
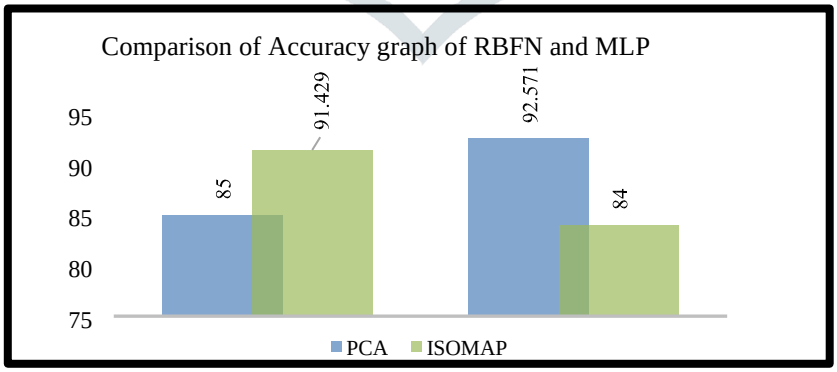


Fig 1.6 Comparison of Accuracy of RBFN and ML



Hence by comparing the different models it is interpreted that RBFN with PCA is the best performing algorithm, resulted in following values for the respective measures are shown in Table II.

Table II: Confusion matrix of RBFN

PCOS (Y/N)	PREDICTED	
Actual	TP = 109	FN = 6
	FP = 9	TN = 51

5. FUTURE ENHANCEMENT

The dataset utilized to create the model is quite limited. As a result, a larger number of records should be analyzed. The dataset with a greater number of records should be deployed with the proposed approach, and then DL models should be compared. Additionally, for enhanced performance, we would like to do more detailed hyper-parameter tuning of DL algorithms as well as improved feature engineering. IOT technology may be integrated with the suggested method to capture health data for the development of an integrated medical system.

6. CONCLUSION

PCOS is a condition caused by a hormonal imbalance in the body of young women. It is a highly common problem that affects a large number of people throughout the world. Infertility and anovulation may occur from this. The illness can be treated, if it is detected early on. This method can help doctors diagnose diseases more quickly, allowing patients to receive treatment sooner. As a result, based on the symptoms presented, we were able to correctly establish the appropriate classification model applying deep learning methods and techniques to diagnose PCOS. On the clinical data, we implemented PCA and ISOMAP with MLP, RBFN, and ANN with various layers. The RBFN with PCA outperforms the other DL models in terms of accuracy (92.571 percent). In comparison to other algorithms, the ANN with 4 Layers has the lowest Root Mean Squared Error (0.2646). Other classifiers should be used to increase the accuracy of this algorithm. By increasing the training and validation datasets accuracy, recall and precision might be improved. Building a model with varied parameters, such as optimizers, activation functions, and the amount of training cycles, may be refined to accurately classify data.

REFERENCES

[1]. F. Stein, M.L. Leventhal, “Amenorrhea associated with bilateral polycystic ovaries,” American Journal. Obstetrics and. Gynecology, 1935; 29; 181-191.

[2]. UCLA Health. “Polycystic Ovary Syndrome (PCOS)”. Polycystic Ovary Syndrome (PCOS). https://www.uclahealth.org/obgyn/pcos (accessed April. 25, 2022).

[3]. Palak Mehrotra, JyotirmoyChatterjee, ChandanChakraborty, BiswanathGhoshdastidar, Sudarshan Ghoshdastidar. “Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques”. IEEE, 2011. DOI: 10.1109/INDCON.2011.6139331

[4]. Amsy Denny, Anita Raj, Ashi Ashok, Remya George, Maneesh Ram C. “i-HOPE: Detection and Prediction System for Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques” IEEE, 2019. DOI: 10.1109/TENCON.2019.8929674

[5]. Malik Mubasher Hassan, Tabasum Mirza “Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome”. ResearchGate, International Journal of Computer Applications, 2020. DOI: 10.5120/ijca2020920688

[6]. Satish C. R Nandipati, Chew XinYing and Khaw Khai Wah. “Polycystic Ovarian Syndrome (PCOS) Classification and Feature Selection by Machine Learning Techniques” University Malaysia, Applied Mathematics and Computational Intelligence, Volume 9, Dec 2021.

[7]. Namrata Tanwani. “Detecting PCOS using Machine Learning”. IJMTES, International Journal of Modern Trends in Engineering and Science, jan 2021. DOI: 10.13140/RG.2.2.10265.24169

[8]. Priyanka R. Lele, Anuradha D. Thakare. "Comparative Analysis of Classifiers for Polycystic Ovary Syndrome Detection using Various Statistical Measures". International Journal of Engineering Research & Technology (IJERT), Vol. 9 Issue 03, March 2020.

[9]. Vaidehi Thakre1, Shreyas Vedpathak, Kalpana Thakre and Shilpa Sonawani. . "PCOcare: PCOS Detection and Prediction using Machine Learning Algorithms". Biosc.Biotech.Res.Comm, Vol 13 No 14, Pp-240-244, 2020. Doi:http://dx.doi.org/10.21786/bbrc/13.14/56

[10]. Pijush Dutta, Shobhandeb Paul, Madhurima Majumder. "An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOS". Research Square. Nov 2021. DOI: https://doi.org/10.21203/rs.3.rs-1043852/v1

[11]. Muhammad Sakib Khan Inan, Rubaiath E Ulfath, Fahim Irfan Alam, Fateha Khanam Bappee, Rizwan Hasan. "Improved Sampling and Feature Selection to Support Extreme Gradient Boosting for PCOS Diagnosis". IEEE Xplore. 2021. DOI: 10.1109/CCWC51732.2021.9375994

[12]. Shakoor Ahmad Bhat. "Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms". School of Computing, National College of Ireland. 2021

[13]. Vikas B, Radhika Y, Vineesha K. "Detection of Polycystic Ovarian Syndrome using Convolutional Neural Networks". International Journal of Current Research and Review, Vol 13, Issue 06. 2021. DOI: http://dx.doi.org/10.31782/IJCRR.2021.13630

[14]. R M Dewi et al. "Classification of polycystic ovary based on ultrasound images using competitive neural network". Journal of Physics: Conference Series, Ser. 971 01200. 2018. doi :10.1088/1742-6596/971/1/012005

[15]. C. Gopalakrishnan, M. Iyapparaja. "Detection of Polycystic Ovary Syndrome from Ultrasound Images Using SIFT Descriptors". Bonfring International Journal of Software Engineering and Soft Computing, Vol. 9, No. 2, April 2019. doi: https://doi.org/10.1016/j.gltp.2021.08.010

[16]. M Sumathi, P Chitra, R Sakthi Prabha and Srilatha K. "Study and detection of PCOS related diseases using CNN". IOP Conf. Ser.: Mater. Sci. Eng. 1070 012062, 2021. doi:10.1088/1757-899X/1070/1/012062

[17]. Rachana B, Priyanka T, Sahana K N, Supritha T R, Parameshachari B D, Sunitha R. "Detection of Polycystic Ovarian Syndrome Using Follicle Recognition Technique". Global Transitions Proceedings 2021, doi: https://doi.org/10.1016/j.gltp.2021.08.010

[18]. Data Science Foundation. "Knowing all about Outliers in Machine Learning". Outliers: To drop or not to Drop. https://datascience.foundation/sciencewhitepaper/knowing-all-about-outliers-in-machine-learning (accessed April. 30, 2022).