International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Detection of Polycystic Syndrome in Ovary Using Machine Learning Algorithm

Manjunathan Alagarsamy[1], Nithyadevi Shanmugam[2], Dinesh Paramathi Mani[3], Meenal Thayumanavan[4], K.Karpoora Sundari[5], Kannadhasan Suriyan[6]

Submitted: 26/10/2022 **Revised:** 22/12/2022 **Accepted:** 02/01/2023

Abstract: PCOS is the reproductive metabolic condition in which the ovary produces the number of follicles that is unusually high. The number of follicles, size, and location of the ovary are observed using the data set of the ovary. Because of the varied sizes of follicles and the fact that it is strongly linked to veins and tissues, radiologists have traditionally had a tough time diagnosing PCOS. To predict the PCO syndrome fertility and infertility for the data collected the from KAGGLE repository, preprocessing techniques are being used to extract useful information for analysis. Heat Map is the preprocessing technique used for identifying correlated features. Then the extracted data are considered for training and testing to classify the occurrence and nonappearance of PCO syndrome. For data training and classification, Support Vector Machine, KNN, Naive Bayes, and Hybrid Algorithm are used. The proposed approach outperforms other current methods and has been proven to be effective.

Keywords: PCOS – SVM – KNN – Naïve Bayes – Ensemble – Evaluation Metrics

1. Introduction

Polycystic Ovarian Syndrome (PCOS) is a difficult endocrine condition that has a significant impact on women's health. The ovary develops multiple follicular pimples, which represent perplexity. PCOS disrupts the follicular balance for motivating hormone, which is required for optimal egg maturation inside the ovaries. [1]. PCOS is caused by the accumulation of incompletely formed follicles, which results in infertility, irregular or absent menstrual periods, undesirable growth of hair, face, body, weight, stoutness, skin inflammation, and hyperinsulinemia [2]. CT, MRI, and Ultrasound are the restorative imaging techniques included to examine the human body in order to diagnose and cure illnesses.Because of its non-invasive nature, ultrasound is commonly used to image the liver, kidneys, and uterus. It is also adaptable, transportable, diverse, has amazing worldly aims, and is relatively inexpensive.

An ultrasound examination of the ovaries should be performed as a result of these symptoms. Since the image quality of ultrasound is poor, it's possible that PCOS is a mistake. The number of follicles in polycystic ovaries isn't specified; in some circumstances, the usual ovary also has a growing number of follicles [3]. The European Medical Council's standard criterion for distinguishing polycystic ovary from many follicles in a regular ovary. According to the prescription, polycystic ovaries have at least 12 follicles with a width of 2–9 mm. During the picture securing a stage, the disturbance is visible for an obtained image in a variety of ways. The air hole among the human stiff and the transducer test will cause turbulence in the image which affects the determination. [4]. When compared to CT and MRI modalities, the ultrasound image contains a lot of noise, especially Speckle Noise. Because it weakens the image's edge and the finest detectable detail, dot clamor is cumulative in nature, making analysis difficult. The test's repetition range and transducers have a major impact on the picture's goals and nature. Higher-end ultrasound machines now come with great tests and are employed to lessen air prattle between the test and the human body during the procurement season to boost ultrasound image quality. To improve the division outcome and for precise finding, ultrasonography images of the ovary are sifted using different image de-noising algorithms after procurement. It's critical to reduce spot clamor in the ultrasound ovary image to accurately determine the number of follicles. A great number of the cutting-edge separation

1Department of Electronics and Engineering, K.Ramakrishnan College of Technology, Trichy - 621112, Tamil Nadu, India, manjunathankrct@gmail.com
2Department of Electronics and Communication Engineering, Dr.N.G.P. Institute of Technology, Coimbatore - 641048, Tamil Nadu, India, ndsnithya88@gmail.com
3Department of Electronics and Communication Engineering, Sona College of Technology, Salem, Tamil Nadu, India, pmdineshece@live.com
4Department of Electronics and Communication Engineering, Kongunadu College of Engineering and Technology, Trichy - 621215, Tamil Nadu, India, meenushammu@gmail.com
5Department of Electronics and Communication Engineering, K.Ramakrishnan College of Technology, Trichy - 621112, Tamil Nadu, India, karpoorasundarik.ece@krct.ac.in
6Department of Electronics and Communication Engineering, Study World College of Engineering, Coimbatore, Tamil Nadu, India, kannadhasan.ece@gmail.com

methods available today have a number of control parameters that must be adjusted in order to achieve optimal results. The radiologist examines the ovary with typical ultrasound equipment, turning the test dynamically to filter ovaries while first perceiving individual follicles assessing their size and measurement. This arrangement is arduous, sustained, and unpleasant for the enduring. A robust computational paradigm for the detection of follicles is suggested to address these challenges.

2. Literature

e Survey

In Literature, several works to extract useful information from the obtained dataset using various preprocessing techniques and utilization of this data set for the classification of normal and abnormality of the disease. Various machine learning techniques used for the classification of syndrome and non-syndrome with better accuracy and precision have been discussed. First Theresa Princy, R. et al. [5] offered an overview of the many classification systems used in predicting heart illness. To perform their survey, Nave Bayes, KNN, Decision Tree Algorithm, and Neural Network are the classification techniques used by them. Although KNN and ID3 were shown to be more effective when the danger of cardiac syndrome was recognized, they were not found to be as effective once it was revealed that they were. Radhimeenakshi [6] offered a manuscript for employing ANN and SVM Classifiers to forecast cardiac disease, and it was accepted for publication. SVM has been demonstrated to be more accurate. Yao Xiao et al [7] used RF Miner to conduct a study on precautionary vascular health, in which eight base classification techniques were utilized, one of the best was Naive Bayes and Random Forest, and the worst of which was Naive Bayes and Random Forest. The rewards of both Naive Bayes and Random Forest are combined into a single classification technique using a cascaded classifier.

In their study, Messan Komi et al [8] suggested a Sugar estimate arrangement combining five data mining groupings demonstrating procedures. The Gaussian mixture model (GMM), the Support vector machine (SVM), the Logistic regression, the Extreme learning machines (ELM), and artificial neural networks (ANN). When examined using the MATHLAB Tool, ANN was found to be more effective than ELM and GMM, with an accuracy of 89 percent. The egg does not mature adequately due to a lack of hormones, resulting in ovum freedom failure. In diagnosis, categorizing follicles from an ultrasound image is crucial. Over-segmentation occurs when improved watershed methods are employed to retrieve follicles from the ovary. When the result of the watershed method is fed into the clustering algorithm, the recognition rate improves [9]. Broken edges are the result of using an edge-based segmentation method. To solve these issues, morphological operations are applied. Images that are shrunk may lose important features. For effective diagnosis in medical imaging, accurate results are required [10]. Growing strategies to efficiently segment follicular regions were introduced. These procedures necessitate the selection of a seed point by a physician, which falls under the category of manual follicle detection. The use of region-expanding methods reduces noise sensitivity and improves recognition rates [11]. A binary image is frequently used by morphological operators. Top hat transform was utilized to remove the upbeat features from the ultrasound image background utilizing morphological opening and closing processes. For segmentation, the improved picture is fed into the Scan line thresholding. When compared to manual detection, this technique has a lower error rate [12]. Histogram equalization is used to improve the contrast of the de-speckled image. In addition, because the suggested segmentation approach operates on high-intensity valued items, the histogram equalized image is subjected to a negative modification. The energetic outline devoid of boundaries is employed in the segmentation stage. After applying the energetic outline approach, the final image contains segmented sections [13]. The researchers M. Alotaibi and A. Alsinan et al, [14] showed a study on PCOS in Gulf countries. They demonstrated the system through the use of mobile healthcare devices. The statistical software SPSS was used to conduct the investigation. A method for automated cyst detection and categorization in the ovary was proposed by Sandy Rihana et al. [15] combining multiscale morphological and SVM analysis. S. Rethinavalli et al. [16] published a study that used the type of menstruation as a predictor for PCOS. The authors started with a 32-attribute dataset, which was subsequently reduced to 7 through feature opting during the pre-processing step. To compare patients with PCO syndrome to those who do not have PCO syndrome,) and the ANN algorithm is utilized.

Uma Ojha and colleagues [17] studied the prediction of breast cancer using eight data mining algorithms. To create their predictions, they used K means, EM, PAM, and Fuzzy c-means, as well as SVM, C5.0, KNN, and Naive Bayes. In the prediction of sickness occurrence, the classification techniques C5.0 and SVM, among others, have been displayed accuracy of 81%.The suggested strategy for Polycystic Ovary Syndrome Detection Using Advanced Techniques is now being implemented as a result of the preceding survey. The journal is prepared in the described flow: The supplies and procedures followed to treat polycystic ovarian syndrome are described in Section I. The survey is being done in section to select the proper classifiers in portion II. Preprocessing of the data using heat map techniques in section III. SVM, KNN, Nave Bayes, and Ensemble are among the classification approaches covered in Portion IV. Finally, Portion V says, the best among all the four machine learning techniques is being justified based on the accuracy and error plot.

3. Materials and Methods

Polycystic ovary syndrome is a hormonal imbalance that is characterized by irregular or prolonged menstrual cycles, as well as increased levels of the male hormone testosterone in the blood. Follicles, which are small collections of fluid formed by the ovarian follicles, may struggle to start releasing eggs on a routine basis if the ovaries must not develop enough of them. The Polycystic ovarian syndrome data set is collected from the KAGGLE repository. It consists of a total of 541 data among which 364 are Polycystic ovarian infertility data and 177 are Polycystic ovarian fertility data's that have been taken for analysis. From this overall database, both training and testing data

are separated in a certain ratio. Thus, the collected database is raw data's some preprocessing techniques are carried out for further analysis. Figure 1 below shows the flow for analyzing polycystic ovarian syndrome.

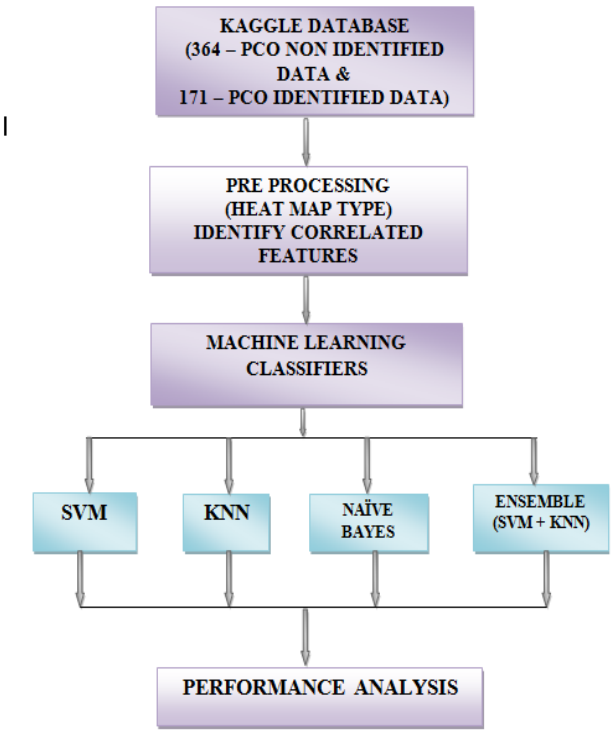

Fig.1. Flow Chart for PCO Analysis

4. Preprocessing

Preprocessing is a procedure that should be used immediately after obtaining the raw database to be studied. However, the database retrieved may contain unwelcomed information for analysis, such as name, age, weight, and height. This strategy is the sole way to get the essential data for analysis. Similarly, there were 541 PCO databases received in total. It contains about 42 different parameters, including S.No, Patient Name, Weight, Height, Pressure, Follicle No, and so on. From this data set, the undesired data set of about 17 parameters that is not necessary is removed using the proper channel. The Heat map method, which identifies the most linked features for subsequent analysis, is a major contribution to this technique [18].

A heat map is a two-dimensional data visualization tool that uses color to represent the magnitude of a phenomenon. The change in color could be in hue or intensity providing the reader with an easy-to-understand graphic display of how the incidence is grouped or changes over time The clump heat map and the spatial scatter diagram are two different types of heat maps that really are fundamentally different from one another in their design. Magnitudes are placed in a matrix with defined cell size, whose rows and columns represent discrete phenomena and categories, and the ordering of rows and columns is intentional and somewhat random, with the goal of discovering or displaying clusters as established by statistical study. The cell size isn't set in stone, but it must be large enough to be visible. On the other hand, the position of a magnitude on a spatial heat map is determined by its location in that space, with no idea of cells; the phenomenon is believed to vary continuously.

5. Classification Techniques

The classification algorithm is used to distinguish between PCO Syndrome data and normal ovarian data with non-syndrome acquire key information using attribute origin techniques. The classification methods used in this study are outlined below.

5.1 Support Vector Machine

The support vector machine (SVM) remains the most consistent and competent approach for machine learning. It is the goal of this classifier to develop the best classifier model to distinguish between two classes in the training data in a binary-class situation. To estimate the separating hyperplane between glaucoma and normal retina, a linear classifier model function must be used in a linearly separable sample. SVM suggests optimum purpose obtained by maximizing the variance between classes because of a huge number of hyperplanes [19]. There is a noticeable difference in the margin between the two groups. Data sets that are the closest to a hyperplane point are defined as their margin. Despite the unlimited number of hyperplanes, only a trickle can be deliberated by SVM solutions. When it comes to classification, SVMs are the broadest, by the way not only transport the extreme outcomes, also offer a proportion of flexibility when it comes to categorizing new data [22]. The SVM model is included in the package is as,

$$\frac{1}{2}(\omega^m)^T\omega^m + C\sum_{i=1}^{l}\xi_i^m \quad \textbf{[1]}$$

$$(\omega^m)^T\phi(z_i) + b^m \geq 1 - \xi_i^m, \;\; if\; y_i = m$$

[2]

$$(\omega^m)^T\phi(z_i) + b^m \leq 1 - \xi_i^m, \;\; if\; y_i \neq m$$

[3]

$$\xi_i^m \geq 0, i = 1, \ldots\ldots, l$$

[4]

Where,

z_i - Attribute data

ϕ, C - Penalty parameter

The data scattered, which is laborious and makes it impossible to separate linearly, is one of the issues in the classification task. The kernel function, which is added by SVM, converts the real data into a novel space. The renovation practice with a dot product is used in this process. The idea is to easily separate data which have been converted into an advanced level before. The hyperplane function can be written in the form of an equation.

$$K(z_n, z_i) = \emptyset(z_n)\emptyset(z_i) \quad [5]$$

The RBF kernel is utilized in this classification; hence the hyperplane formula is used.

$$K(z_n, z_i) = \exp\left(-\gamma\|z_n - z_i\|^2 + c\right)$$

[6]

Where

c and γ → Parameter for Optimization

z_n → Data in support vector

The categorization of SVM is written as follows

$$f(z_i) = \sum_{n=1}^{N} \propto_n y_n K(z_n, z_i) + b \quad [7]$$

Where

y_n= Data in labeling
$\propto_n$= Multiplier with Lange range

5.2 Naïve Bayes

The objective is to come up with a way to classify retinal fundus images into a specific group. This innovation will help researchers figure out if the new retinal scans show signs of glaucoma. Because it won't require a sophisticated recurrent parameterization mechanism, Naive Bayes is straightforward to implement [20-21]. Say, with large datasets, Naive Bayes are easily produced and understood by non-experts. Even if it isn't the best predictor for a certain function, it can be trusted and executed well. The fundamental nave Bayesian theorem is presented in the equation.

$$P\left(\frac{G}{N}\right) = \frac{P(G)P\left(\frac{N}{G}\right)}{P(N)} \quad [8]$$

N → Attributes
G → Class
$P\left(\frac{G}{N}\right)$ → Nth probability of G
$P\left(\frac{N}{G}\right)$ → Gth probability of N
$P(N)$ → N th probability
$P(G)$ → G th probability

Z indicates in the equation is,

$$N = (n_1, n_2, n_3, \dots \dots, n_n) \quad [9]$$

Substitute Z in the equation

$$P\left(\frac{G}{n_1, n_2, n_3, \dots\dots, n_n}\right) = P(G)P\left(\frac{n_1, n_2, n_3, \dots\dots, n_n}{G}\right)$$

[10]

$$= P(G)P\left(\frac{n_1}{G}\right)\left(\frac{n_1, n_2, n_3, \dots\dots, n_n}{G, n_1}\right)$$

[11]

$$= P(G)P\left(\frac{n_1}{G}\right)P\left(\frac{n_2}{G, z_1}\right)\left(\frac{n_1, n_2, n_3, \dots\dots, n_n}{Y, n_1, n_2}\right)$$

[12]

$$= P(G)P\left(\frac{n_1}{G}\right)P\left(\frac{n_2}{Y, z_1}\right)P\left(\frac{n_3}{G, n_1, n_2}\right) \dots . P\left(\frac{n_n}{G, n_1, n_2, n_2, \dots\dots, n_{n-1}}\right) \quad [13]$$

Because it has significantly highly complex components, the exceeding equation is much more difficult to comprehend. Consider a case in which each of the equation's components is unrelated to the others.

$$P\left(\frac{G}{n_1, n_2, n_2, \dots\dots, n_n}\right) = P(G)\prod_{i=1}^{n} P\left(\frac{n_i}{G}\right) \quad [14]$$

In the Nave Bayes Predictor, Hypothesis Maximum a Posteriori (HMAP) is utilized to maximize the probability inside each category.

$$H_{MAP} = argmax\ P\left(\frac{G}{n_1, n_2, n_2, \dots\dots, n_n}\right) \quad [15]$$

$$= argmax\ P(G)\prod_{i=1}^{n} P\left(\frac{n_i}{G}\right) \quad [16]$$

The PCO fertility and infertility data is classified using the aforementioned equation to determine if the patient has PCO Syndrome or not.

5.3 K-Nearest Neighbor

The K-NN classifier is widely used in biological research. Byassociating a certain training and testing set, training is carried out. The training data is determined using n features. N-dimensional design training samples are often preserved because each set reflects a feature of an n-dimensional field. When applied to an unknown dataset, k-NN looks for the k training examples that are the most similar to the unknown set in its field of application [22]. These k training samples represent the k "NN" of the unrevealed collections. The "likeness" of the collections was determined using a distance metric like Euclidean. k-NN is the second most common class in the k-NN categorization of these unknown collections [23]. Training samples are assigned to unnamed collection n samples distant when k is equal to 1. As the training dataset size grows, so does the k value.

The implementation technique for the KNN algorithm is as follows

Estimate the Euclidean distance D.

$$D = \sqrt{\sum_{i=1}^{n}(x_n - y_n)^2}$$

[17]

i) Organize the n distances calculated in ascending order.
ii) Choose k as a proper key number and exploit the prior agreement to determine the first distances of k.
iii) Guess the k places by distances of k.
iv) X_n corresponds to the PCO Syndrome grouping for Y_n.
v) If the criteria $Y_n > X_n$ is met, the data is classified as normal.

objects that are 'near' each other will also have a similar
neighbor." k-NN is an improvisation over the nearest

5.4 Ensemble (SVM + KNN)

The ensemble algorithm under consideration is a hybridization of the Support Vector Machine and the KNN algorithm, as illustrated in the following figure 2. This improves the efficiency, accuracy, and precision for evaluating and identifying the PCO syndrome by cascading the performance of the SVM technique with the performance of the KNN technique. This is the innovative technology that has been concentrated in this study for further investigation and evaluation.

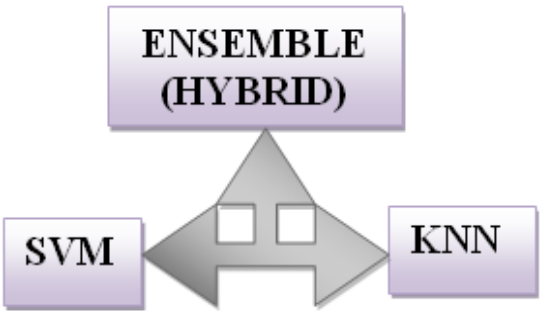

Fig.2. Ensemble SVM & KNN

6. Result and Discussion

There are 541 PCO fertility and infertility records in the KAGGLE database. The total amount of PCO infertility data for the study is 364, whereas normal fertility data is 177. The dataset is separated into two sections: one for educating the model for classifier and the further for testing it. The performance measures are used to evaluate the developed classification model. accurateness, True Positive Rate, True Negative Rate, false-positive rate, false-negative rate, precision, and remind are the performance metrics used in this study. The performance metrics are calculated using the confusion matrix items, which are discussed below.

True Positive (TP): PCO syndrome is the actual and expected output, implying that the ovary is afflicted by PCO syndrome in real life and predicts the same.

True Negative (TN): Both actual and expected output is normal, indicating that the ovary is in good health (no PCO Syndrome).

False Positive (FP): The actual result is Normal, while the anticipated output is PCO syndrome, indicating that the ovary is healthy in reality but is impacted by PCO syndrome according to the model.

False Negative (FN): The actual output is PCO infertility, while the anticipated output is normal, indicating that the ovary is affected by PCO syndrome in real life but that the ovary is healthy according to the model.Accuracy: The ratio of the total corrected PCO Syndrome and normal prediction to the total number of samples obtained for testing is used to calculate the accuracy, which is then expressed in the equation.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100 \quad [18]$$

True Positive Rate: The True Positive Rate is intended by isolating the number of correct PCO syndrome predictions by the total number of correct PCO Syndrome predictions and false normal predictions.

$$TPR = \frac{TP}{FN+TP} * 100 \quad [19]$$

True Negative Rate: The True Negative Rate is designed by separating the number of right standard predictions by the total number of precise normal and wrong PCO syndrome predictions.

$$TNR = \frac{TN}{TN+FP} * 100 \quad [20]$$

False Positive Rate: The False Positive Rate is formulated by dividing the number of incorrect PCO syndrome predictions by the total number of correct normal and incorrect glaucoma predictions.

$$FPR = \frac{FP}{TN+FP} * 100$$

[21]

False Negative Rate: The False Negative Rate is calculated by dividing the number of incorrect normal predictions by the total number of correct PCO syndrome predictions and incorrect normal predictions.

$$FNR = \frac{FP}{FN+TP} * 100$$

[22]

Precision: The precision is proportional to the exact detection of PCO Syndrome follicles to the total number of correctly predicted PCO Syndrome follicles as expressed.

$$Precision = \frac{TP}{FP+TP} * 100 \quad [23]$$

F-Measure: The F-Measure is a statistic for evaluating a model's accuracy on a given dataset. It's used to put binary classification algorithms to the test, which divide retrieved instances into "positive" and "negative" categories. In the PCO analysis, the F-score is a prominent statistic for evaluating information retrieval systems such as search engines, as well as a variety of machine learning models, especially in natural language processing.

$$F = \frac{TP}{TP + \frac{1}{2}(FP+FN)} \quad (24)$$

Error Rate: The error approach is defined as the difference between the projected output values and the actual values. If the target values are categorical, the mistake is reported as a proportion of the overall number of targets. This represents the proportion of instances in which the prediction is incorrect.

ERR = 1 – Accuracy (25)

Balance Accuracy: When analyzing the performance of a binary classifier, it is possible to utilize a metric called balanced accuracy. Particularly effective in cases where the classes are unequally distributed; for example, when one of the two classes appears significantly more frequently than the other This occurs frequently in a variety of circumstances, including anomaly detection and the presence of a disease.

$$Balanced\ Accuracy = \frac{1}{2}(\frac{TP}{P} + \frac{TN}{N})$$

(26)

The presentation criteria aid in identifying the most effective classification model for diagnosing PCO Syndrome and a healthy ovary. The examination of classifiers like as K-Nearest Neighbour, Support Vector Machine, Nave Bayes, and Ensemble (SVM + KNN) is shown in Table 1. It shows that the level of correctness is very high, equalling 97 percent for Ensemble (SVM + KNN) analysis. The accuracy level is then modest, with KNN, Nave Bayes, and SVM Analysis accuracy levels of 90 percent, 92 percent, and 95 percent, respectively. Ensemble (SVM + KNN) and SVM have high and low precision of 98 percent and 96 percent, respectively, whilst other analyses such as nave Bayes and KNN have 93 percent and 93 percent, respectively. The error rate for the Ensemble algorithm is about 0.02, for KNN error rate is about 0.09, and for Nave Bayes and SVM error rates are 0.07 and 0.02 respectively. For a better understanding, a heat map with multiple classifiers for different features including accuracy, TNR, TPR, FNR, FPR, Precision, F-Measures, Error rate, and Balance was used to analyze the performance of Polycystic Syndrome in the ovary.

Figure 3 depicts the accuracy rate of all four different classifiers used for the analysis of PCO syndrome, whereas Figure 4 depicts the range of error rate variation for the various classifiers that have been introduced for the PCO syndrome study. This also helps to demonstrate that the suggested Ensemble (SVM + KNN) approach is superior to other classifiers in terms of analyzing and detecting PCO Syndrome. In addition, it is justified strongly that the Ensemble (SVM + KNN) is the best machine learning technique by viewing the evaluation metrics comparison chart shown in figure 5. This also shows the comparison of all the eight parameters extracted using four classifiers.

Table 1. Analysis of Polycystic Ovarian Syndrome (PCOS) using Heat Map Preprocessing Technique with the special classifiers

Classifier	**Accuracy**	**TNR**	**TPR**	**FNR**	**FPR**	**Precision**	**F-Measure**	**Error Rate**	**Balance Accuracy**
SVM	95.37037	90.625	97.3684	2.63157	9.375	96.103896	96.73203	0.046296	93.99671
Naïve Bayes	92.59259	84.8484	96	4	15.1515	93.506494	94.73684	0.074074	90.42424
KNN	90.74074	81.8181	94.6666	5.33333	18.1818	92.207792	93.42105	0.092593	88.24242
Ensemble (SVM + KNN)	97.22222	96.6666	97.4359	2.56410	3.33333	98.701299	98.06452	0.027778	97.05128

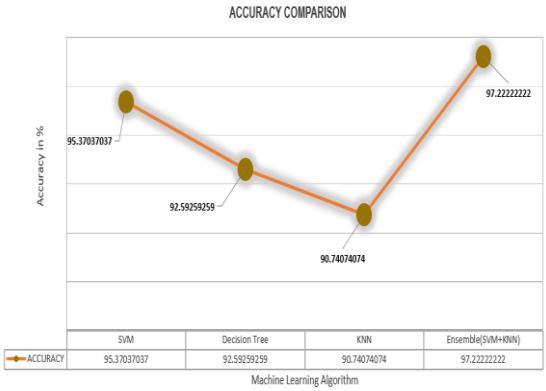

Fig. 3. Graphical analysis of Accuracy Comparision with the different classifiers

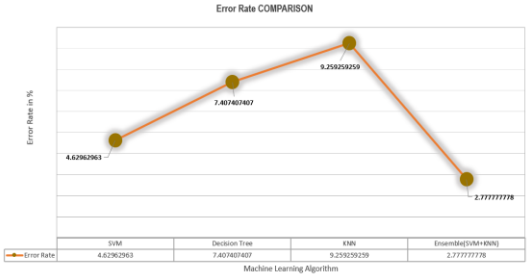

Fig.4. Error Rate Analysis for PCO Syndrome with Different Classifiers

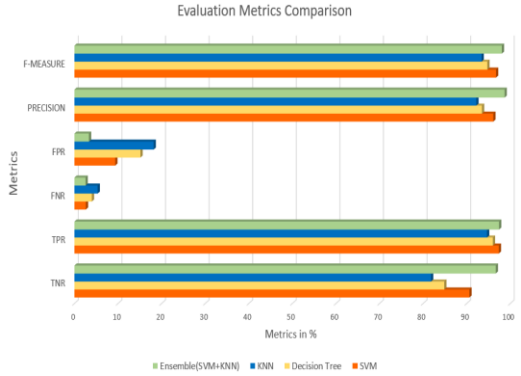

Fig. 5. Graphical Analysis of PCO Syndrome using Heat Map with the different classifiers

7. Conclusion

The number of people suffering from PCO Syndrome is increasing on a daily basis. Detecting and treating diseases at an early stage is critical for successful treatment; otherwise, the disease can cause significant distress to the patient. In order to compile the PCO fertility and infertility syndrome dataset, it was obtained from the KAGGLE repository. The data has been divided into symmetric halves, one for guidance purposes and the rest for trying purposes. It is necessary to apply pre-processing to raw data in order to make them suitable for categorization. The heat map techniques are functional to the pre-processed statistics in regulating to identify and extract the most important information from the data set that has been gathered. Last but not least, the retrieved feature data is fed into a variety of classifier models for the detection of PCO syndrome follicles. Performance measures, as well as feature extraction approaches, are utilized to evaluate the model and its effectiveness. Based on the outcome of the data analysis, the Ensemble (SVM + KNN) classifier has a high accuracy of 97 percent and is the most accurate classifier. The SVM then provides a second-high accuracy level of 95 percent, which is impressive. The Naive Bayes classifier achieves the third-highest level of accuracy, which is 93 percent, with its classifications. This evaluation aids in the identification of the most effective permutation of element withdrawal and a machine learning classifier model for distinguishing between a vigorous and a PCO Syndrome-affected ovary in the future.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] A. Pellicer, P. Gaitán, F. Neuspiller, G. Ardiles, C. Albert, J. Remohíand C. Simón, “Ovarian Follicular Dynamics: From Basic Science to Clinical Practice”, Journal of Reproductive Immunology, Vol. 39, No. 12, Pp.29-61,2021.

[2] Affiliates of Medifocus.com, “Medifocus Guidebook: Polycystic OvarySyndrome”, Medifocus.com, Inc, 2007.

[3] T.W. Kelsey and Wallace W.H.B. “Ovarian Volume Correlates Strongly with Number of Non-Growing Follicles in the Human Ovary”, Obstetrics and Gynecology International, Vol. 2012, Pp. 1-5, 2012.

[4] C. Battaglia, P.G. Artini, A.D. Genazzani, R. Gremigni, M.R.Slavatoriand M.R. Sgherzi, “Color Doppler Analysis in Oligo and Amenorrheic Women with Polycystic Ovary Syndrome”, GynecologicalEndocrinology, Vol. 11, No. 2, Pp. 105-110, 2020.

[5] Theresa Prince. R, J. Thomas,” Human Heart Disease, Prediction System using Data Mining Techniques, Vol. 12, No. 3, Pp. 100-105 2016.

[6] M.C. Nicolae, “Comparative Approach for Speckle Reduction in medical Ultrasound Images”, Romanian Journal of Bio-Physics, Vol.20, No. 1, Pp. 13-21, 2010.

[7] Yao Xiao, Ruogu Fang,” RFMiner: Risk Factors Discovery and Mining for Preventive Cardiovascular Health”, IEEE/ACM, (CHASE), Pp- 278-279, 2017

[8] Messan Komi, Jun Li, Y Yongxin Zhai, Xianguo Zhang,” Application of Data Mining Methods in Diabetes Prediction”, 2nd International Conference on Image, Vision, and Computing. Vol. 53, no. 1, Pp. 64-78, 2017.

[9] Y. Deng, Y. Wang, and P. Chen, “Automated Detection of PolycysticOvary Syndrome from Ultrasound Image”, 30th Annual International IEEE Engineering in Medicine and Biology Society ConferenceVancouver, British Columbia, Canada, pp. 20-24, 2008.

[10] P.S. Hiremath and Jyothi R. Tegnoor, “Automatic Detection of Follicles in Ultrasound Images of Ovaries using Edge Based Method”, International Journal of Computer Applications Special Issue on RecentTrends

in Image Processing and Pattern Recognition Pp. 15-16, 2010.

[11] Y. Deng, Y. Wang, and Y. Shen, "An Automated Diagnostic System polycystic Ovary Syndrome based on Object Growing", Journal of

Artificial Intelligence in Medicine, Elsevier Science Publishers Ltd.Essex, UK, Vol. 51, No. 3, Pp. 199-209, 2011.

[12] P. Mehrotra and C. Chakraborty, "Automated Ovarian FollicleRecognition for Polycystic Ovary Syndrome", International Conference on Image Information Processing, pp. 1-4, 2011.

[13] P.S. Hiremath and R. JyothiTegnoor, "Automatic Detection of Follicles in Ultrasound Images of Ovaries using Active Contours Method," International Journal of Service Computing and Computational Intelligence Vol. 1, No. 1, Pp. 26-30, 2011.

[14] M. Alotaibi, A. Alsinan." Mobile Polycystic Ovarian Syndrome Management and Awareness System for GulfCountries ", IEEE, SAI Computing Conference, Vol. 1, No. 1, Pp. 26-30, 1315,2016.

[15] Sandy RIHANA, Hares Moussallem, Chiraz Skaf, Charles Yaacoub," Automated Algorithm for Ovarian Cysts Detection in Ultrasonogram", IEEE, 2nd International Conference on Advances in Biomedical Engineering, 2013.

[16] S. Rethinavalli, Dr. M. Manimekalai," A Hypothesis Analysis on the Proposed Methodology for Prediction of Polycystic Ovarian Syndrome", S., November 2016 | Vol 6, Issue 11, Pp396-400

[17] Uma Ojha, Dr. Savita Goel," A study on prediction of Breast cancer Recurrence using Data mining techniques", 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence, Pp-527-530, 2017.

[18] M. Iyapparaja, P. Sivakumar, "Metrics Based Evaluation for DiseaseAffection in Distinct Cities", Research J. Pharm. and Tech., Vol. 10, No.8, pp. 2487-2491, 2017.

[19] D. Selvathi, N.B. Prakash, V. Gomathi, G.R. Hemalakshmi, "Fundus image classification using wavelet-based features in detection of Glaucoma", *Biomed. Pharmacol. J*. pp. 795–805. 2018

[20] D.C. Shubhangi, N. Parveen, "A dynamic ROI based Glaucoma detection and region estimation technique", *Int. J. Computer Sci. Mobile Comput.* pp. 82–86. 2019.

[21] A Manjunathan, ED Kanmani Ruby, W Edwin Santhkumar, A Vanathi, P Jenopaul, S Kannadhasan, "Wireless HART stack using multiprocessor technique with laxity algorithm", Bulletin of Electrical Engineering and Informatics, 10, 6, 3297-3302 (2021)

[22] C Jeyalakshmi, A Manjunathan, A Karthikram, T Dineshkumar, W Edwin Santhkumar, S Kannadhasan, "Automatic Wireless Health Instructor for Schools and Colleges", Bulletin of Electrical Engineering and Informatics, 11, 1 (2022)

[23] S Karthikeyani, C Ganesh Babu, M Ramkumar, R Sarath Kumar, G Priyanka, A Manjunathan, "Determination of Cardiac Output based on Minimally Invasive Impedance Plethysmography in Various Healthy Subjects", Int. J. Aquat. Sci, 12, 2021, 1078-1086