

Vector Conversion Based PCOS Detection In Data Segmentation Using Multi Task Learning By Dynamic Deep Learning Architecture

Sathiyapriya Jagadeesan and Praveena Marannan*

Department of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science,
Coimbatore, Tamilnadu, India.

<http://dx.doi.org/10.13005/bbra/3355>

(Received: 14 October 2024; accepted: 09 January 2025)

Polycystic ovarian syndrome (PCOS), the most prevalent endocrine abnormality in women who are fertile, interferes with hormone secretion over time, leading to a large number of ovarian cysts and other serious health problems. However, the doctor's experience plays a significant role in the accuracy of interpretations, which makes the practical clinical diagnostic approach for PCOS essential. Therefore, a PCOS prediction model powered by artificial intelligence might be a workable supplement to the labor-intensive and prone to error diagnosis technique. This research proposes a novel technique in clinical data-based PCOS detection by dimensionality reduction with segmentation using a deep learning model. Here the input PCOS clinical data has been collected and processed for removing missing values and dimensionality reduction based on vector conversion with Kernel Principal Component Analysis. Then the data quality is enhanced with annotation coverage for PCOS detection using a dynamic Bayesian hidden Markov v model. The experimental analysis is performed for the PCOS clinical dataset in terms of prediction accuracy, validation accuracy, RMSE, precision, and F-1 score. The proposed method obtained an overall 97% F-1 score and prediction accuracy of 98%, validation accuracy of 98%, RMSE of 1%, and precision of 99%.

Keywords: Artificial Intelligence; Deep Learning; Dimensionality Reduction; PCOS detection; Segmentation; Vector conversion.

Girls and women of reproductive age are more susceptible for a medical disorder referred to as polycystic ovarian syndrome (PCOS), which presents with a range of signs. Changes in menstrual cycle, abnormal face and body hair development, acne, obesity, the existence of ovarian cysts, and infertility are a few of these signs. It is a complicated, incurable disorder caused by a hormonal imbalance that typically has an impact on how the ovaries function. Future development of various health issues such as type 2 diabetes,

gestational diabetes, high cholesterol, high blood pressure, sleep apnea, and stroke is quite likely in women with PCOS. According to studies¹, 3–10% of people in an undefined population have PCOS². PCOS and endometriosis are the two reproductive illnesses that affect women the most frequently. PCOS is characterized by polycystic ovaries, hyperandrogenism, and ovulatory dysfunction. In conjunction with other syndrome symptoms, hyper androgenism is acknowledged as a critical diagnostic factor (1). Different diagnostic methods

*Corresponding author E-mail: praveenamarannan@gmail.com



are used in different parts of the world. National Institutes of Health (NIH) states that the presence of both hyperandrogenism and oligo/amenorrhea can be used to diagnose PCOS (2). Depending on diagnostic standards, PCOS incidence varies³. About 7% of reproductive-aged women have hyper androgenic chronic an ovulation. One polycystic ovary is adequate to make diagnosis of PCOS according to Rotterdam criteria (3), which stipulate that the polycystic ovary must be ultrasonically visible. The prevalence of PCOS increases from 55% to 91% in women with norm gonadotropic an ovulation when using the NIH criteria (5). Secondary causes should be ruled out before primary causes, according to all diagnostic strategies. Machine learning (ML) techniques for PCOS identification have promise⁴. Using image data, a ML method applied to ovarian ultrasound scans could accurately diagnose PCOS (6). Although the usage of intelligent programmes in medicine and healthcare is growing, a strong and focused strategy is still required. Targeting patients or medical professionals has demonstrated enormous potential for machine learning⁵. Convolutional neural networks (CNN) and recurrent neural networks (RNN), which are used in machine learning techniques such as DL, the diagnosis can be made more accurately while the system can be made simpler. With the aid of annotated training data, these techniques can be widely utilised to model or simulate an intelligent system or process⁶.

Related Works

Deep learning algorithms like CNNs and RNNs have been used extensively in research to identify PCOS. An expert system using an artificial neural network⁷ was built to categorise PCOS cases using a dataset of 1000 samples divided into two groups: healthy individuals and subjects with diseases. For instance, author⁸ distinguished between PCOS as well as non-PCOS criteria using Adaptive Boosting (AdaBoost), Bagging Ensemble with Random Forest, and Multi-Layer Perceptron (MLP) classification methods. An improved chi-squared (CS-PCOS) strategy⁹ for feature selection is presented for PCOS detection. Another study in this sector used the Chi-square approach to choose top 30 features from data¹⁰, and utilizing this condensed feature vector, RF, SVM,

Logistic Regression, Gaussian Naive Bayes, and K Neighbours were utilized to predict the underlying condition of PCOS. Additionally, the diagnostic model that¹¹ suggested using MATLAB to identify PCOS included seven many types of classifiers, and results revealed that Linear Discriminant classifier performed the best. In a different study, suggested by¹², a sample of follicular fluid from 100 women was taken, resulting data set was pre-process reutilizing Raman spectra as well as efficient feature selection methods to be used for ML classification. RF, MLP, boosting algorithm, classification and regression tree methods were used to identify PCOS. An artificial neural network (ANN)¹³ method for detecting PCOS early on after conducting a performance comparison analysis utilizing various NN optimizers to investigate best performing ANN model for classifying dataset into 2 categories: PCOS and Non-PCOS. Another related piece of research in this area was done by¹⁴, who looked at four different ML classifiers to categorise PCOS and non-PCOS records, compared their performance, and found that RF classifier outperformed others. Using RF classifier, DT was then created to determine main factors causing PCOS. To categorise PCOS patients based on scleral images, a DL method using U-net, ResNet 18, Multi Instance Learning design¹⁵ was proposed to improve accuracy. In total, the items in their possession included 721 images, 388 of them included people with PCOS. Their suggested method achieved an average AUC of 98% and accuracy of 93%. A CAD model¹⁶ is developed a method to identify PCOS in ultrasound images using image segmentation and convolutional neural network (CNN). The K-Nearest Neighbour algorithm was also used to categorise the images. Using their own dataset¹⁷ it is fine-tuned 16 Layered VGG-16 methods to categorise PCOS from ultrasound images. Their model had a 92.11% accuracy rate. Image binarization was used by¹⁸ on ultrasound B-mode pictures. Image binarization^{19,20} was used as an image pre-processing technique to convert ultrasonic images of ovarian cysts to grey scale images. The post-processing which includes labeling and connecting of various portions in PCOS, produced the extraction of geometrical features as well as the classification of cysts has been discussed. 90% accuracy is achieved using

SVM as the classifier. Some of the additional deep learning techniques were utilized for improvisation of accuracy and boosting data optimization^{21,22}.

MATERIALS AND METHODS

The following subsection discusses a new hybrid approach for vector conversion-based PCOS detection utilizing deep learning techniques by dimensionality reduction and segmentation. Here to collect input PCOS data and processed for noise removal with vector conversion. Dynamic Bayesian hidden Markov-v-model is employed to improve the quality and coverage of the PCOS image annotations. The data dimensionality reduction is carried out using kernel principal component analysis (KPCA). The proposed architecture for PCOS detection is exhibited in Figure 1.

After completing image collection, the process for obtaining clinical data was begun to investigate how clinical data affected the diagnosis of PCOS. This procedure seeks to obtain clinical data for patients whose ultrasound images were gathered in the previous phase from the hospital system. The features were chosen with the assistance of professional judgments and with consideration of previous studies that identified the significance of those attributes on PCOS diagnosis in some manner^{11,22}. There are many clinical data points that are lacking and need further validation, as we discovered while collecting clinical data for 391 patients whose ultrasound photos were already accessible. Ultimately, a dataset has 285 samples and 22 characteristics. The clinical dataset is composed of 156 non-PCOS individuals and 129 PCOS cases. Diagnoses for patients in the dataset

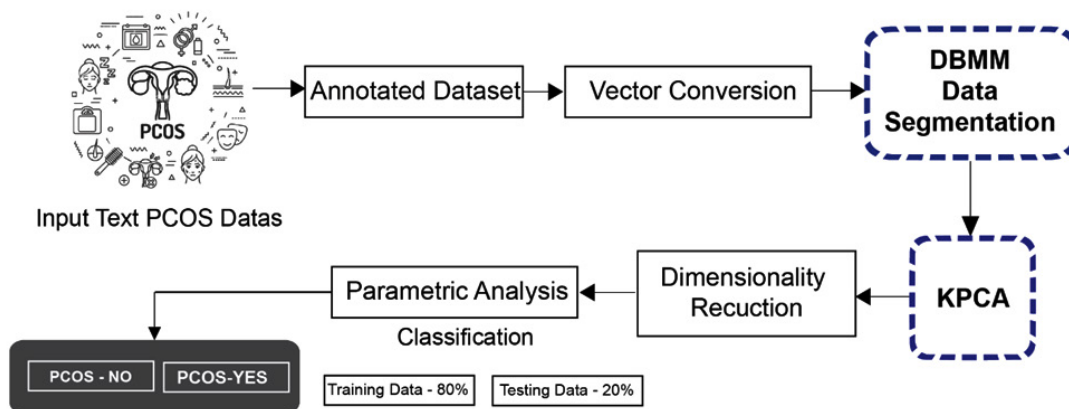


Fig. 1. Proposed PCOS Detection Architecture

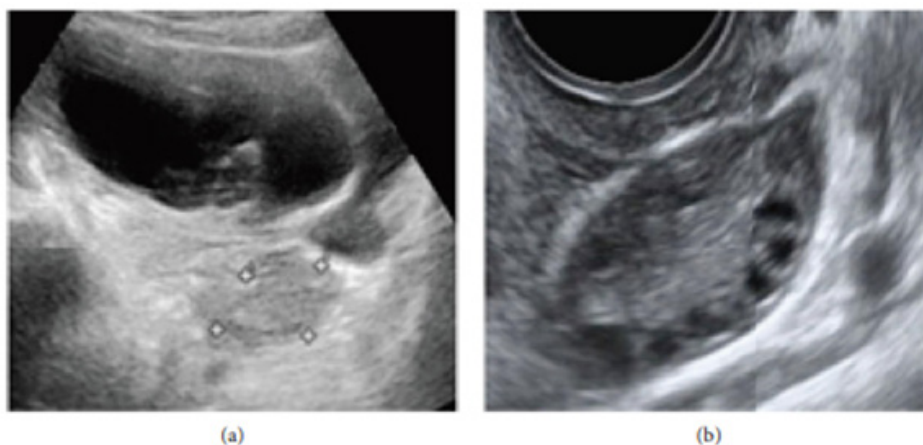


Fig. 2. Clinical view of ovary ultrasound image: (a) ovary with PCOM and (b) normal ovary

are made using laboratory results, medical records, and radiologist analysis of ultrasound pictures to distinguish between PCOS and non-PCOS. Characteristics are divided into four categories, including demographics, vital signs, lab results, and doctors' comments, such as cycle regularity. The dataset has two data types, four nominal/categorical features, and 17 numeric features. Additionally, only photos that clearly show the ovary were chosen for the investigation. Radiologists divided the selected images into two categories for categorization: the sonographically visible PCOM and those with a normal morphology (non-PCOM). As seen in Figure 2, PCOM is represented by an ovary with several, uniformly sized, peripherally positioned follicles that are smaller than one centimetre. The dataset of photographs consists of 391 images which has clinical record result of 127 PCOM, and 264 normal ovaries that are not PCOM

Four categories can be used to group the characteristics, including demographics, vital signs, laboratory tests, and medical notes. The dataset has two data types, four nominal/categorical features, and 17 numeric features. Additionally, only photos that clearly show the ovary were chosen for the investigation. Radiologists divided the selected images into two categories for categorization: those with a non-PCOM and those displaying the sonographic PCOM. Figure 2 illustrates an ovary with numerous, uniformly sized follicles that are positioned peripherally and have a diameter of less than 1 cm. This is the definition of the polycystic ovarian morphology. 391 pictures, 127 PCOM, and 264 normal ovaries that are not PCOM make up the image dataset.

Kernel Principal Component Analysis Based Dimensionality Reduction

A variety of techniques may be used as part of the dimensionality reduction pre-processing

Table 1. Proposed analysis based on various PCOS text dataset

Dataset	Prediction Accuracy	Validation Accuracy	Root Mean Squared Error	Precision	F1-Score
Data 1	97	97	2	96	96
Data 2	96	95	4	96	93
Data 3	98	98	1	99	97

Proposed PCOS Text Dataset Analysis

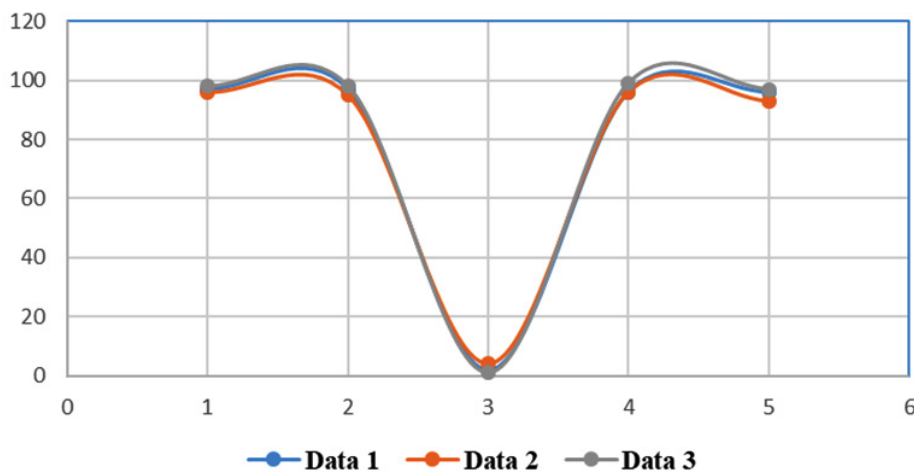


Fig. 3. Proposed analysis for various PCOS dataset (a) Data 1, (b) Data 2, (c) Data 3

techniques. The following advantages that the DR offers are taken into consideration:

- As the number of dimensions rises, the amount of memory that can be used to store data gets smaller.
- Reducing the number of dimensions requires shorter training or computation times.
- Most feature extraction techniques struggle to handle many-dimensional data.
- Multi-collinearity between various data features is successfully handled using DR techniques, and redundant features are removed.
- Reduced data dimensions also facilitate visualization.

The PCA changes the attributes in a dataset into new functions referred to as primary components. The crucial element is a linear combination of the starting variables of the dataset. Primary elements are arranged in a way that the first primary element indicates the greatest discrepancy in the outcomes. However, there is no correlation between them and the first main data variable. Singular Value Decomposition (SVD) splits the

original matrix into its component pieces, which are then used to find redundant features. SVD is based on the concepts of eigenvalues and eigenvectors. The main head variable is discussed in a way that demonstrates the greatest possible variation. The numerical concept of PCA is as follows: A given information focal x_t with $t = 1, 2, \dots, x_t \in \mathbb{R}^n$ and $\sum_{t=1}^p C x_t = 0$, usually $n < p$, so that x is $n \times p$, such that x is $n \times p$ input grid. PCA resolves the conceded Eigen esteems issue by converting this vector space into new vectors.

$$\lambda_i \mu_i = c \mu_i \quad i = 1, 2, \dots, n \quad \dots(1)$$

$c = \left(\frac{1}{p} \left(\sum_{t=1}^p C x_t x_t^T \right) \right)$, and μ_i s Eigenvector whose corresponding to Eigen value λ_i ;

$$y_t(i) = \mu_{1i} x_{1i} + \mu_{2i} x_{2i} + \dots + \mu_{ni} x_{ni} = \mu_i^T x_i \quad i = 1, 2, \dots, n \quad \dots(2)$$

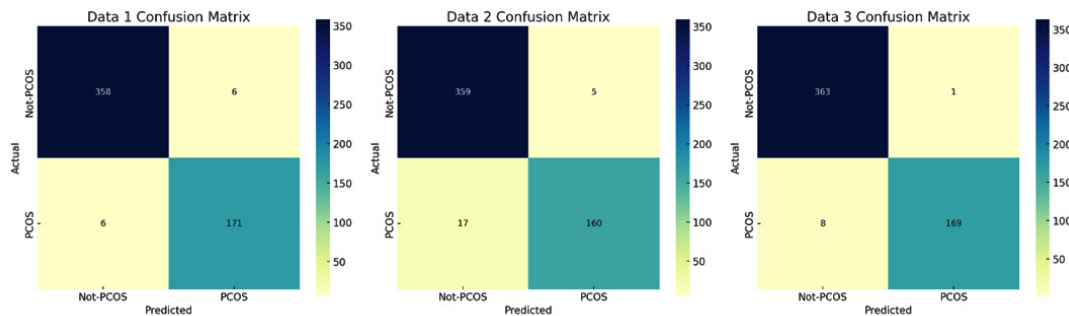


Fig. 4. Confusion Matrix Analysis of Various PCOS Datasets

Table 2. Proposed Analysis for PCOS Dataset

Dataset	Techniques	Prediction	Validation Accuracy	RMSE Accuracy	Precision	F1-Score
Data 1	Reinforcement_PSO	85	88	45	83	78
	SVM_ICA	88	89	46	86	80
	DBMM_KPCA[Proposed]	97	97	2	96	96
Data 2	Reinforcement_PSO	92	86	57	84	85
	SVM_ICA	94	94	58	89	89
	DBMM_KPCA[Proposed]	96	95	4	96	93
Data 3	Reinforcement_PSO	88	91	59	87	84
	SVM_ICA	94	95	60	90	87
	DBMM_KPCA[Proposed]	98	98	1	99	97

Primary $y_i(i)$ variable with highest Eigen value corresponds to data set's largest variance vector. The first few Eigenvectors are only used when the initial vector space PCs are organized in decreasing order by their corresponding Eigen values. Only then can the initial vector space PCs be transformed into new PC vectors, $k < n$.

The Hilbert space H is utilized to implement the self-adjoint operator for the trace class G eqn (3).

$$\text{tr}(T) := \langle \psi_i, T\psi_i \rangle c_1 \|f\|^2 \leq \sum_{\alpha \in A} |\langle h_\alpha, f \rangle|^2 \leq c_2 \|f\|^2 \text{ for all } f \in \mathcal{H} \quad \dots(3)$$

Let $(h_\alpha)_{\alpha \in A}$ be a frame in H . Set $L: H \rightarrow l$ by eqn (4)

$$L: f \mapsto (\langle h_\alpha, f \rangle)_{\alpha \in A} \quad \dots(4)$$

Comparative Analysis - PCOS Dataset 1

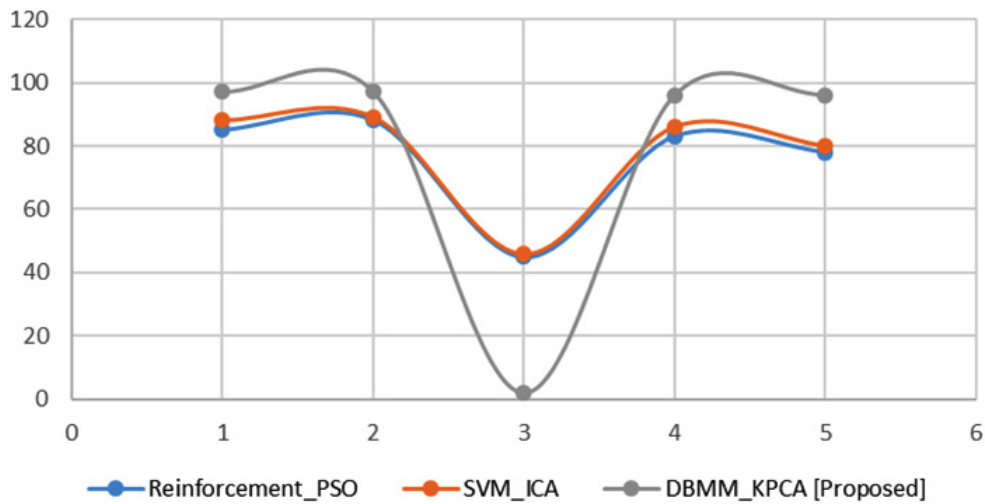


Fig. 5. Comparative Analysis based on PCOS Dataset - 1

Comparative Analysis - PCOS Dataset 2

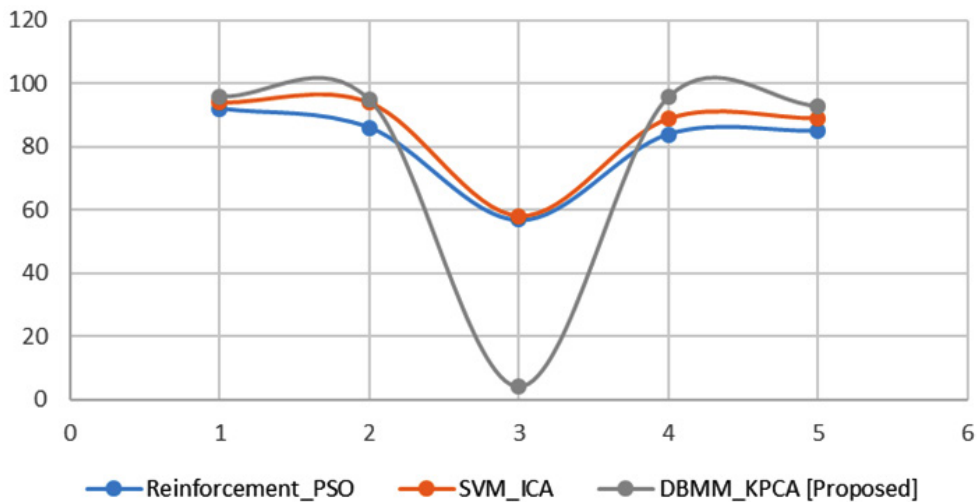


Fig. 6. Comparative Analysis based on PCOS Dataset - 2

Then $L^*: l^2 \rightarrow \mathcal{H}$ represented by eqn (5),

$$L^*((c_\alpha)) = \sum_{\alpha \in A} c_\alpha h_\alpha \quad \dots(5)$$

where $(c_\alpha) \in l^2$; and by eqn (6)

$$L^*L = \sum_{\alpha \in A} |h_\alpha\rangle\langle h_\alpha|$$

$$\sum_{i,j=1}^N \bar{c}_i c_j K(v_i, v_j) \geq 0 \quad \dots(6)$$

for all $\{x_i\}_{i=1}^N \subset S, \{c_i\}_{i=1}^N \subset \mathbb{C}$, and $N \in \mathbb{N}$.

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}(K)} \quad \dots(7)$$

Additionally, replicating characteristic that follows is true by eqn (8):

$$f(x) = \langle K_x, f \rangle_{\mathcal{H}(K)}$$

$$\text{span}\{K_x := K(\cdot, x)\} \quad \dots(8)$$

$\mathcal{H}(K)$ -inner product by eqn (9)

$$\left\langle \sum c_i K_{x_i}, \sum d_j K_{x_j} \right\rangle_{\mathcal{H}(K)} := \sum \bar{c}_i d_j K(x_i, x_j) \quad \dots(9)$$

Since each patient may experience a unique set of symptoms or combinations of symptoms, the most significant features may be what determines the diagnosis. Therefore, feature selection methods might be extremely important in order to choose the most effective features and lower the dataset's dimension to aid doctors in the diagnosis process.

Dynamic Bayesian Hidden Markov Model

It is defined by the HMM model's transition probability, emission probability, and prior distribution. Our model presupposes that each topic 't' and each authors 'a' are connected by multinomial distributions 't' over words and 'a' over topics, respectively. More specifically, two matrices, the TA matrix of author-topic distributions and the WT matrix of topic-word distributions, provide the parameters by eqn (10).

$$D_\alpha(\theta_{\cdot a}) = \frac{\Gamma(T\alpha)}{(\Gamma(\alpha))^T} \prod_{t=1}^T \theta_{ta}^{\alpha-1} \quad \dots(10)$$

and for t, similarly. Assuming that A is corpus's set of authors and that document d has authors from A, it is clear that, under these presumptions, a document's likelihood is given by:

$$P(d | \theta, \Phi, \mathcal{A}) = \prod_{i=1}^{N_d} \frac{1}{A_d} \sum_a \sum_{t=1}^T \phi_{wit} \theta_{ta}$$

Which can be combined across the variable j and q and their respective Dirichlet

Comparative Analysis - PCOS Dataset 3

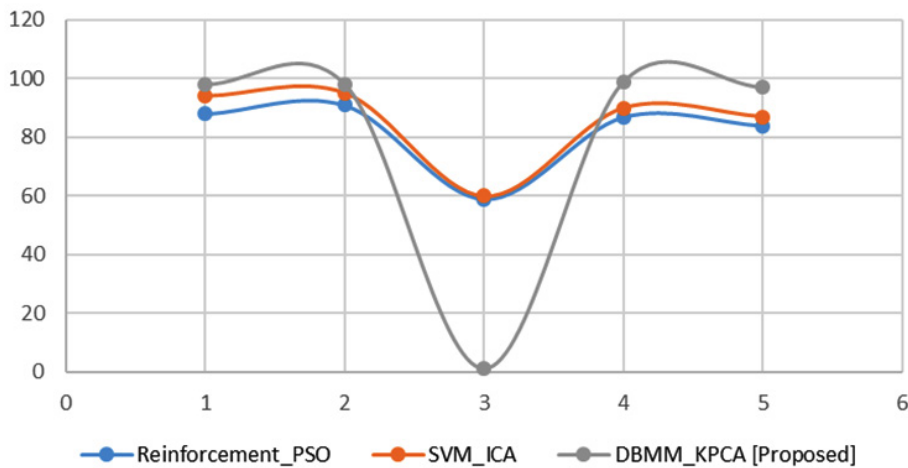


Fig. 7. Comparative Analysis based on PCOS Dataset - 3

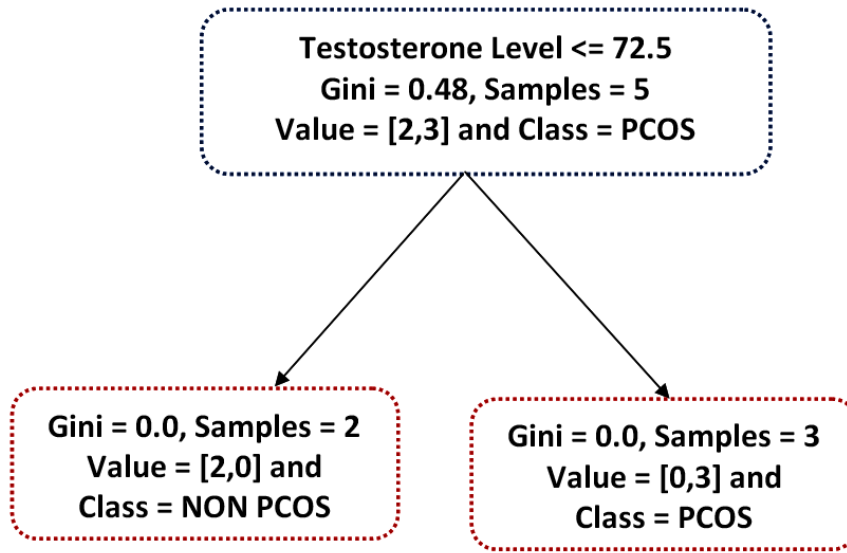


Fig. 8. Custom Decision Tree for PCOS detection on features like BMI, testosterone level, and menstrual regularity

distributions to obtain $P(d|\alpha, \beta, A)$. Markov chain Monte Carlo Methods can be used to efficiently sample the posterior and MAP or MPE methods, for example can be used to estimate Θ and Φ specification matrices. A repository R that stores evolution of a group of NF files serves as the main component of an SCM system by eqn (11):

$$R = \{F_i \mid i = 1, \dots, NF\} \quad \dots(11)$$

Each file F_i is kept as a collection of NVI versions in a repository by eqn (12):

$$F_i = \{V_{ij} \mid j = 1, \dots, NV_i\} \quad \dots(12)$$

Every version is a tuple with several attributes. The most common ones are a unique version ID, the committer's name, the commit time, a log message, and its contents by eqn (13):

$$V_{ij} = \langle id, author, time, message, content \rangle \quad \dots(13)$$

In the following, we shall eliminate the file index i when referring to a single file in order to simplify the notation. The properties message, time, author, and id are unstructured. Eqn (14) models the content using a collection of entities:

$$Content = \{e_i \mid i = 1, \dots, NE\}. \quad \dots(14)$$

Convergence is achieved when a path's edges all have pheromone levels of s-max and all other edges have pheromone levels of s-min. Following that, a rule matching the path with s-max is retrieved and added to the rule set. Ultimately, training data in the training set that satisfies this criterion is removed. until an early stopping requirement is met, repeat this procedure iteratively by eqn (15).

$$P_{ij}(t) = \frac{\left[\tau_{(v_{i-1,k}, r_{i,j})}(t)\right]^\alpha \cdot \left[\eta_{v_{i,j}}(t)\right]^\beta}{\sum_{l=1}^{p_i} \left[\tau_{(v_{i-1,k}, v_{i,j})}(t)\right]^\alpha \cdot \left[\eta_{v_{i,l}}(t)\right]^\beta}$$

$$\eta_{ij} = \frac{|T_{ij} \& CLASS = class_{ant}|}{|T_{ij}|}$$

$$\tau_{(v_{i-1,j}, r_{i,j})}^{(0)} = \tau_{\max}$$

$$\tau_{(v_{i-1,k}, t_{i,j})}(t+1) = \rho \cdot \tau_{(v_{i-1,k}, v_{i,j})}(t) + \frac{Q_{best}^+}{10} \quad \dots(15)$$

Suppose that we are trying to use eqn (16) to find the expected value of a random variable X.

$$E[g(X)] = \int g(x)f(x)dx \quad \dots(16)$$

To a certain g-function. If functions f, g is such that it is not possible to evaluate the integral in (1) directly, then the following can be done. Assuming we can draw a sample from the distribution of X, produce a random sample of size n by eqn (17):

$$x_1, x_2 \dots x_n \text{ and } a_n = \frac{1}{n} \sum_{i=1}^n g(x_i) \quad \dots(17)$$

Given a uniform (0,1) U_k and an i.i.d. the variable X_k with a common frequency h. Let B come from by eqn (18).

$$B = \{(x, u): u \leq f_1(x)/Mh(x)\}$$

$$\begin{aligned} P(\tau = m) &= P(Z_1 \notin B, \dots, Z_{m-1} \notin B, Z_m \in B) \\ &= P(Z_1 \notin B)^{m-1} P(Z_m \in B) \\ &= 1 - P(Z_1 \notin B)^{m-1} P(Z_m \in B) \end{aligned} \quad \dots(18)$$

and hence $P(t < \infty) = 1$, Now

$$\begin{aligned} P(Z_w \in A \mid \tau = m) &= P(Z_m \in A \mid Z_1 \notin B, \dots, Z_{w-1} \notin B, Z_w \in B) \\ &= P(Z_m \in A \mid Z_m \in B) \\ &= P(Z_1 \in A \mid Z_1 \in B), \end{aligned} \quad \dots(19)$$

hence

$$\begin{aligned} P(Z_r \in A) &= \sum_m P(Z_m \in A \mid \tau = m) P(\tau = m) \\ &= \sum_m P(Z_1 \in A \mid Z_1 \in B) P(r = m) \\ &= P(Z_1 \in A \mid Z_1 \in B). \end{aligned} \quad \dots(20)$$

taking $A = (-\infty, a] \times [0, 1]$ for a $\in R$, we have (using $\{Z_t \in A\} = \{W \leq a\}$),

$$\begin{aligned} P(W \leq a) &= P(X_1 \leq a \mid Z_1 \in B) \\ &= \frac{P(X_1 \leq a, Z_1 \in B)}{P(Z_1 \in B)} \\ &= \left\{ \int_{-\infty}^{\infty} \int_0^1 1_n(x_0 u) h(x) du dx \right\} / \left\{ \int_{-\infty}^{\infty} \int_0^1 1_n(x_2 a) h(x) du dx \right\} \\ &= \left\{ \int_{-\infty}^{\infty} \frac{f_1(x)}{Mh(x)} h(x) dx \right\} / \left\{ \int_{-\infty}^{\infty} \frac{f_1(x)}{Mh(x)} h(x) dx \right\} \\ &= \left\{ \int_{-\infty}^{\infty} f_1(x) dx \right\} / \left\{ \int_{-\infty}^{\infty} f_1(x) dx \right\} \\ &= \int_{-\infty}^{\infty} f(x) dx. \end{aligned} \quad \dots(21)$$

For all variables $X_1 \dots N$, the likelihood function of HMDBN has the following structure by eqn (22):

$$P(q, X | \text{HMDBN}) = P(q | \text{HMDBN}) P(X | q, \text{HMDBN}) \quad \dots(22)$$

It is assumed that the transition between hidden graphs follows the first order Markov chain. Additionally, it is assumed that the present observation is statistically independent of the preceding observations, according to the formula $P(x(1:T) \mid q(1:T)) = \prod_{t=1}^T P(x(t) \mid q(t))$

Consequently, it is simple to demonstrate that

$$P(q_i(1)) \prod_{i=1}^N \prod_{t=2}^T P(q_i(t) \mid q_i(t-1)) \prod_{i=1}^N \prod_{t=1}^T \quad \dots(23)$$

Furthermore, we presumptively consider that concealed state q_i for parameter X_i is statistically distinct from that of remaining variables. Concealed graph, which is also a DBN, adheres to conditional independence presumption. Consequently, above formula is by eqn (24)

$$\begin{aligned} &\prod_{i=1}^N P(q_i(1)) \prod_{i=1}^N \prod_{t=2}^T P(q_i(t) \mid q_i(t-1)) \prod_{i=1}^N \prod_{t=1}^T P(x_i(t) \mid q_i(t)) \\ &= \prod_{i=1}^N \left(\pi_i^{q_i(1)} \prod_{t=2}^T \alpha_i^{q_i(t-1)q_i(t)} \prod_{t=1}^T P(x_i(t) \mid \alpha_i^{q_i(t)}, \theta_i^{q_i(t)}) \right) \end{aligned} \quad \dots(24)$$

As a result, we broke down the probability function for each variable into a series of terms, where every term is solely dependent on parents chosen for that variable as well as any pertinent node-specific characteristics. This allows for a modular evaluation of all local adjustments to a possible non-stationary network. A sub-network in HMDBN_i is represented by a graph in each rectangle that only designs conditional relationships between the variable X_i and its parents G_i . The figures enclosed in parenthesis indicate duration. Observable states are represented by the darkened nodes. There is an observable state associated with each hidden graph. The time-points at which regulatory inputs to variables change are referred to as transition times for one variable. Because each non-stationary sub-network HMDBN_i might have a variable transition time and hidden graph, the HMDBN is node-specific.

Consequently, we use SEM algorithm 23.

It combines structural search for model selection with the conventional EM technique, which optimises parameters. BWBIC rating. This is how Q function is expressed by eqn (25):

$$Q(HMDBN, HMDBN') = \sum_q P(q | x, HMDBN') \log P(x, q | HMDBN') = \sum_{t=1}^N (\sum_{h=1}^{H_t} P(q_t(1) = h | x, HMDBN') \log \pi_t^h + \sum_{h=1}^{H_t} \sum_{u=1}^{H_t} \sum_{t=2}^T P(q_t(t-1) = h, q_t(t) = u | x, HMDBN') \log a_{t-1}^{hu} + \sum_{h=1}^{H_t} \sum_{t=1}^T P(q_t(t) = h | x, HMDBN') \log P(x_t(t) | G_t^h, \theta_t^h)) \quad \dots(25)$$

By solving with regard to and A, first two terms in parenthesis of Eq. (26) is maximised. Next, we can obtain their updated estimates:

$$\pi_t^{h'} = P(x, q_t(1) = h | HMDBN') / \sum_{h=1}^{H_t} P(x, q_t(1) = u | HMDBN') \\ a_{t-1}^{hu'} = \sum_{t=2}^T P(x, q_t(t-1) = h, q_t(t) = u | HMDBN') / \sum_{t=2}^T P(x, q_t(t-1) = h | HMDBN') \quad \dots(26)$$

After getting marginal likelihood for each hidden graph G_i^h , we may solve following optimisation issue to yield G_2^i in h , which optimises the third term by eqn (27).

$$G_i^{h'} = \arg \max_{G_i^h} \log \int \prod_{t=1}^T P(x_t(t) | G_i^h, \theta_i^h)^{P(q_i(t)=h|x, HMDBN')} P(\theta_i^h | G_i^h) d\theta_i^h \quad \dots(27)$$

A local maximum is still guaranteed to be reached by this generalised EM method. We used the Laplace approximation, a large-sample approximation method, to derive an efficient way to approximate this integral. As a result, the following formula comes close to expressing logarithmic marginal likelihood in formula (28):

$$l(\theta_i^h) - \frac{d_i^h}{2} \log m_i^h = \sum_{j=1, k=1}^{d_i^h} \sum_{t=1}^r P(q_i(t) = h | x, HMDBN') \chi(i, j, k; x(t)) \log \theta_{i,jk}^h \\ = \frac{\theta_i^h (r_i - 1)}{2} \log \sum_{t=1}^r P(q_i(t) = h | x, HMDBN')$$

where

$$\theta_{i,jk}^{h'} = \frac{\sum_{t=1}^T \chi(i, j, k; x(t)) P(q_i(t) = h | x, HMDBN')}{\sum_{k=1}^r \sum_{t=1}^T \chi(i, j, k; x(t)) P(q_i(t) = h | x, HMDBN')} \quad \dots(28)$$

Finding a female patient with PCOS requires a significant amount of time and work. The research work that has been provided thus suggests automating the implementation of disease diagnosis in order to further support it through manual evaluation by medical professionals and experts. This would make it easier to detect

the illness at the proper time and stop it from developing in later stages, enabling the treatment of ovarian disorders. Through a dataset that is given to the system, the implemented approach often predicts the presence of the disease.

Results of the Proposed Work

The experimental findings are presented and discussed in this section. The ML models were created using Scikit-learn. The trials were carried out using Google Colab. Additionally, other ML methods based on various feature selection techniques were compared to the stacking ML models. Two training and testing sets with ratios of 20:80 and 30:70 are used to track the models' performance.

Description of Set

The individuals who went to the Ghosh Dastidar. Establishment for Ripeness Exploration (GDIFR), located in Kolkata, between April 2010 and Walk 2010, were the focus of the review. The clinic provided a pre-made form for each patient to use in order to record their medical history and any additional physical exams. Of the 250 women that were remembered for this study, 150 had polycystic ovary disease (PCO), while the remaining 100 had normal ovarian function. The PCOS symptomatic criteria were developed by the American Society for Regenerative Medicine (ASRM) and the European Society for Human Reproduction and Embryology (ESHRE). The following three conditions must be met: oligo- and anovulation, polycysts in at least one ovary, and an ultrasound exam showing a clinical and/or biochemical evidence of increased male hormone production. PCO should be confirmed if the patient satisfies any two of the ASRM/ESHRE-recommended requirements. The following criteria were used to diagnose PCOS: 1) The length of the cycle; 2) the clinical and metabolic characteristic; and 3) polycystic ovaries, with hyperthyroidism and Cushing syndrome excluded. Patients who were either virgin or refused treatment, as well as those whose ovaries did not contain a follicle, were not included in the analysis. Exploring best subsets of features to ensure best possible data description is the process. In this review, dataset contains 40 credits subsequent to pre-processing, which might bring down exactness of classifier on off chance that the less critical ones are all considered. Accordingly, the elements in this setting have been

focused on and chosen thoroughly utilizing three kinds of component determination methods to figure out the ideal arrangement of highlights from the PCOS informational index.

Dataset Link

1. <https://www.kaggle.com/datasets/ayamoheddhine/pcos-dataset>
2. <http://biogps.org/dataset/tag/pcos/>
3. <https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos>

DISCUSSIONS

The aforementioned table-1 presents a suggested analysis based on multiple PCOS text datasets in terms of F-1 score, RMSE, precision, and prediction and validation accuracy. We investigate and compare various fusion algorithms to merge image characteristics with clinical features in order to determine whether fusion of image and clinical features can affect PCOS prediction. Joint fusion and late have been contrasted in the beginning. Due to the fact that all cases have been determined to be negative, the late fusion does not produce any hopeful results. Figure 3 portrays the proposed analysis based on various PCOS data for all performance metrics.

The estimations of the true positive rate and false positive rate serve as the foundation for the analysis. The second experiment looks at how clinical characteristics affect ultrasound image-based PCOS diagnosis. The ultrasound picture features are combined with clinical information in a fusion deep learning model to obtain a conclusive diagnosis of PCOS. The suggested method is trained by minimising loss of a “objective” function against dataset with the goal of minimising the error, and the selection of the loss function is a key hyperparameter for achieving the desired results. The suggested confusion matrix for several PCOS text datasets is projected in Figure 4

Using logical transformations like erosion, dilation, and others, the process of feature extraction uses comparisons based on different PCOS text datasets in terms of Prediction accuracy, Validation accuracy, RMSE, Precision, and F-1 score. These adjustments are straightforward image-dependent actions. Typically, these processes are carried out on a grayscale image. The limits of the foreground

are destroyed by the erosion process. With the kernel, it does a 2D convolution. If each pixel is below the kernel value of 1, as in a characteristic pixel image (original image), it erodes differently. As a result, the front body’s thickness is minimal. The process of dilation is the exact opposite of erosion. While narrowing the object, erosion narrows the white noises. The surface of the item increases when the noise is reduced. Therefore, dilatation undoes the lost area of the item caused by erosion. Therefore, dilation typically occurs after the erosion process.

From above Figure 5, 6 and 7, the comparison is shown for various PCOS dataset. Here proposed technique Prediction accuracy of 97%, validation accuracy of 97%, RMSE of 2%, precision of 96%, F-1 score of 96%; existing Reinforcement_PSO attained Prediction accuracy of 85%, validation accuracy of 88%, RMSE of 45%, precision of 83%, F-1 score of 78%, SVM_ICA attained Prediction accuracy of 88%, validation accuracy of 89%, RMSE of 46%, precision of 86%, F-1 score of 80% for dataset 1; for dataset 2 proposed technique Prediction accuracy of 96%, validation accuracy of 95%, RMSE of 4%, precision of 96%, F-1 score of 93%; existing Reinforcement_PSO attained Prediction accuracy of 92%, validation accuracy of 86%, RMSE of 57%, precision of 84%, F-1 score of 85%, SVM_ICA attained Prediction accuracy of 94%, validation accuracy of 94%, RMSE of 58%, precision of 89%, F-1 score of 89%; proposed technique Prediction accuracy of 98%, validation accuracy of 98%, RMSE of 1%, precision of 99%, F-1 score of 97%; existing Reinforcement_PSO attained Prediction accuracy of 88%, validation accuracy of 91%, RMSE of 59%, precision of 87%, F-1 score of 84%, SVM_ICA attained Prediction accuracy of 94%, validation accuracy of 95%, RMSE of 60%, precision of 90%, F-1 score of 87% for dataset 3.

It can be shown that among the different classifiers for CV 20, voting 1 (soft) offers the highest accuracy. There is a lot of literature on PCOS, however not all of the datasets are comparable. Direct comparison is therefore not possible.

A 70:30 train-to-test split ratio and an 80:20 split ratio were the two sets of frameworks that were created. This would aid us in analysing

the top-performing framework in the next steps. Feature scaling came after the separation. We used the common scalar approach to scale the features in the training dataset. After standardisation, the data is unaffected by outliers and has no bounding range. Only 177 out of 541 samples in the target PCOS (Y/N) class had PCOS, creating an imbalance. We balanced the data on the training dataset using the Borderline Synthetic Minority Oversampling Technique (SMOTE) to lessen the risk of improper model training. This technique creates fake data points from minority classes by oversampling. In contrast to conventional SMOTE, it addresses problem of incorrectly categorised outliers. We acquired 364 and 364 counts for both PCOS and Non-PCOS classes following Borderline-SMOTE balancing.

Model Analysis with Decision Tree Visualization

The proposed PCOS detection model leverages Kernel Principal Component Analysis (KPCA) for dimensionality reduction and a Dynamic Bayesian Hidden Markov Model (DBHMM) for segmentation and classification. To illustrate its classification process, a decision tree model is used to visualize feature-based decision points, reflecting how key features like hormonal levels, BMI, and menstrual irregularities influence PCOS prediction.

Advantages of the Proposed Methodology

- **High Accuracy and Precision:** The model achieves 98% accuracy, demonstrating its reliability in diagnosing PCOS.
- **Robust Data Processing:** KPCA effectively reduces data dimensions, mitigating overfitting risks.
- **Temporal Feature Integration:** DBHMM accounts for time-dependent changes, enhancing clinical relevance.
- **Interpretability:** The decision tree visualization aids in understanding critical diagnostic features.

Disadvantages of the Proposed Methodology

- **Computational Complexity:** Combining KPCA and DBHMM requires significant computational power.
- **Data Dependency:** The model's performance is highly dependent on dataset quality and size.
- **Limited Real-World Deployment:** Real-time PCOS detection may face delays due to processing complexity.

Future Work

Future efforts should focus on integrating ensemble techniques like Random Forest or Gradient Boosting to enhance robustness and accuracy. Furthermore, employing larger, diverse datasets and exploring multi-modal learning combining clinical and imaging features can improve real-world application potential. Additionally, incorporating explainable AI (XAI) tools can ensure trustworthiness in critical healthcare systems.

CONCLUSION

This research work suggests a novel technique in dimensionality reduction with segmentation of text analysis in PCOS detection using deep learning techniques. Here the input text PCOS clinical data is pre-processed and segmented using Kernel Principal Component Analysis and Dynamic Bayesian Hidden Markov Model. To develop a PCOS diagnosis as well as an analysis method for CAD systems is the goal of this research project. One dataset contains images of the ovaries taken using ultrasound technology, while the other contains clinical data such as vital sign data, lab test results, and other symptoms that can be used to diagnose PCOS. It is suggested to use a variety of DL methods to implement AI-based CAD. In order to lower the false positive rate as well as improve the performance of the proposed method, the first proposed model uses ovary ultrasound images to detect PCOS morphology using a deep learning model for automated PCOM diagnosis. Using a DL method to detect PCOS, this research activity has published a study to examine the effects of merging imaging as well as clinical information. The proposed method achieved an F-1 score of 97% and prediction accuracy of 98%, validation accuracy of 98%, RMSE of 1%, and precision of 99%.

ACKNOWLEDGEMENT

I, Sathiyapriya Jagadeesan would like to express my sincere gratitude to Praveena Marannan, my research supervisor and the research institution, Dr SNS Rajalakshmi College of Arts & Science, Coimbatore, Tamil Nadu for the support

and research facility provided throughout this research work.

Funding Sources

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of interest

The authors do not have any conflict of interest.

Data Availability Statement

This statement does not apply to this article.

Ethics Statement

This research did not involve human participants, animal subjects, or any material that requires ethical approval.

Informed Consent Statement

This study did not involve human participants, and therefore, informed consent was not required.

Clinical Trial Registration

This research does not involve any clinical trials.

Author Contribution

Sathiyapriya Jagadeesan: Collected the PCOS text data and initiate the research work and implemented in the application for proposed model and guided the analysis of the PCOS dataset. Praveena Marannan: Played a key role in refining the manuscript as a corresponding author and provided crucial insights for interpreting the results in simulator tool. Both authors collaborated in writing and revising the paper, ensuring intellectual rigor, and have given their approval for its publication.

REFERENCES

1. Kumari R, Singh J, Gosain A. SmS: SMOTE-stacked hybrid model for diagnosis of polycystic ovary syndrome using feature selection method. *Expert Systems With Applications*. 2023;225:120102.
2. Alamoudi, A., Khan, I. U., Aslam, N., Alqahtani, N., Alsaif, H. S., Al Dandan, O., & Al Bahrani, R. A Deep Learning Fusion Approach to Diagnosis the Polycystic Ovary Syndrome (PCOS). *Applied Computational Intelligence and Soft Computing*. 2023; 2023:1-15.
3. Elmannai, H., El-Rashidy, N., Mashal, I., Alohal, M. A., Farag, S., El-Sappagh, S., & Saleh, H. Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence. *Diagnostics*. 2023;13(8):1506.
4. Chitra P, Srilatha K, Sumathi M, Jayasudha FV, Bernatin T, Jagadeesh M. Classification of Ultrasound PCOS Image using Deep Learning based Hybrid Models. 2023 Second IEEE International Conference on Electronics and Renewable Systems (ICEARS). 2023.
5. Khanna VV, Chadaga K, Sampathila N, Prabhu S, Bhandage V, Hegde GK. A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. *Applied System Innovation*. 2023;6(2):32.
6. Nithyanandh S, Omprakash S, Megala D, Karthikeyan MP. Energy Aware Adaptive Sleep Scheduling and Secured Data Transmission Protocol to enhance QoS in IoT Networks using Improvised Firefly Bio-Inspired Algorithm (EAP-IFBA). *Indian Journal of Science and Technology*. 2023;16(34):2753-2766.
7. Arularasan R, Balaji D, Garugu S, Jallepalli VR, Nithyanandh S, Singaram G. Enhancing Sign Language Recognition for Hearing-Impaired Individuals Using Deep Learning. *IEEE International Conference on Data Science and Network Security (ICDSNS)*. 2024:1-6.
8. Lv, W., Song, Y., Fu, R., Lin, X., Su, Y., Jin, X., & Huang, G. Deep Learning Algorithm for Automated Detection of Polycystic Ovary Syndrome Using Scleral Images. *Frontiers in Endocrinology*. 2022;12.
9. Gupta A, Fatima H. A Systematic Review of Machine Learning for Ovarian Cyst Detection using Ultrasound Images. 2023 2nd IEEE International Conference on Applied Artificial Intelligence and Computing (ICAAIC). May 2023.
10. Mittapalli JS, Khanna K, Mutha JA, Nair S. A Cloud-Based Prediction and Self-Diagnosis System for PCOS Using Machine Learning Models. In: *Lecture Notes in Electrical Engineering*. ; 2023:477-484.
11. Selvam N, Joy J. K. Plant Leaf Disease Detection with Multivariable Feature Selection Using Deep Learning AEN and Mask R-CNN in PLANT-DOC Data. *Biotech Res Asia* 2024;21(4), 1649-1663.
12. Swamy SR, S NPK. Hybrid Machine Learning Model for Early Discovery and Prediction of Polycystic Ovary Syndrome. 2022 Second IEEE International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE). 2022; 9:1-8.
13. Kharb S, Joshi A. Multi-omics and machine

- learning for the prevention and management of female reproductive health. *Frontiers in Endocrinology*. 2023;14.
14. R S, R NB, Radhakrishnan R, P S. Computerized Diagnosis of Polycystic Ovary Syndrome Using Machine Learning and Swarm Intelligence Techniques. *Research Square (Research Square)*. January 2023.
15. Vora K, Shah A, Shah N, Verma P. Prediction of Polycystic Ovary Syndrome (PCOS) Using Optimized Machine Learning Classifiers. In: *Algorithms for Intelligent Systems*. ; 2023:1-12.
16. Chen, W., Yang, Q., Hu, L., Wang, M., Yang, Z., Zeng, X., & Sun, Y. Shared diagnostic genes and potential mechanism between PCOS and recurrent implantation failure revealed by integrated transcriptomic analysis and machine learning. *Frontiers in Immunology*. 2023;14.
17. Harish KP, Dhivyanchali MN, Devi KN, Krishnamoorthy N, Sree RD, Dharanidharan R. Smart Diagnostic System For Early Detection And Prediction Of Polycystic Ovary Syndrome. Conference: 2023 IEEE International Conference on Computer Communication and Informatics (ICCCI). 2023;1:1-6.
18. Dutta P, Paul S, Sadhu A, Jana GG, Bhattacharjee P. Performance of Automated Machine Learning Based Neural Network Estimators for the Classification of PCOS. In: *Intelligent Human Centered Computing*. ; 2023:65-73.
19. Ambad R, Agrawal A, Lahoti R, Muley P, Pande P. Role of artificial intelligence in PCOS detection. *Journal of Datta Meghe Institute of Medical Sciences University*. 2022;17(2):491.
20. Faris NN, Miftin FS. An intelligence model for detection of PCOS based on k means coupled with LS SVM. *Concurrency and Computation Practice and Experience*. 2022;34(21).
21. Devi PA, Megala D, Paviyasre N, Nithyanandh S. Robust AI Based Bio Inspired Protocol using GANs for Secure and Efficient Data Transmission in IoT to Minimize Data Loss. *Indian Journal of Science and Technology*. 2024;17(35):3609-3622.
22. Eldho KJ, Nithyanandh S. Lung Cancer Detection and Severity Analysis with a 3D Deep Learning CNN Model Using CT-DICOM Clinical Dataset. *Indian Journal of Science and Technology*. 2024;17(10):899-910.