

Deep learning model for diagnosing polycystic ovary syndrome using a comprehensive dataset from Kerala hospitals

Divya Rao¹, Riddhi Rajendra Dayma¹, Sanjeev Kushal Pendekanti¹, Aneesha Acharya K.²

¹Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

²Department of Instrumentation and Control Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

Article Info

Article history:

Received May 28, 2024

Revised Jul 11, 2024

Accepted Jul 17, 2024

Keywords:

Deep learning

Disease prediction

Genetic algorithm

Healthcare

Polycystic ovarian syndrome

ABSTRACT

Polycystic ovary syndrome (PCOS) requires early and precise diagnosis to manage and prevent long-term health consequences effectively. In this research, a large dataset of healthcare data gathered from various hospitals in Kerala, India, was evaluated using multiple machine learning (ML) and deep learning (DL) models to identify a highly reliable and accurate prediction of PCOS. The six algorithms used for comparison with the proposed DL model are support vector classification, random forest, logistic regression, k-nearest neighbors, and gaussian naive Bayes; they were selected due to their strengths in handling features in large datasets. The highly parameterized neural networks were tuned using efficient approaches like Optuna and genetic algorithms. The results indicated that the model implemented using our proposed combination of DL model and Optuna, outperformed the traditional models, achieving 93.55% reliability. This suggests the potential for using deep learning for decision-making in diagnosing PCOS. This method demonstrates the importance of integrating various data types with powerful analytic tools in medical diagnostics to support customized therapy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Aneesha Acharya K.

Department of Instrumentation and Control Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education

Manipal-576104, Karnataka, India

Email: ak.acharya@manipal.edu

1. INTRODUCTION

About 10% of women who are of reproductive age have polycystic ovary syndrome (PCOS), a relatively common endocrine condition. Menstrual irregularities as well as androgen excess and chronic anovulation are its features including androgen excess through ovarian or adrenal sources and/or androgen-exacerbated anovulation [1]. PCOS not only affects fertility but also poses a risk for a number of metabolic diseases such as type 2 diabetes hypertension and many more cardiovascular diseases [2]. Current diagnostic criteria for PCOS are irregular periods, elevated androgens, polycystic ovarian follicles and infertility [3], [4]. Significant associations between PCOS reproductive health, metabolic health and cardiovascular risk as well as mental health have been reported and therefore need to be managed as a whole [5]. The research studies regarding advances in deep learning (DL) and machine learning (ML) outline encouraging potentials for the enhancement of medical diagnostic methods. These technologies are capable of analyzing huge and complex datasets for possible pattern identification that may not be revealed by simple observation or traditional methods [6]. Developing a model for PCOS using ML and DL has the ability to combine different forms of data like hormone levels, metabolic states and genetic data to increase efficiency and predictive accuracy [7].

Challenges are that the disease is not uniform in its expression and the variable quality of the data collected. Good annotated data is an important factor for training models effectively but such data is difficult to come by in the context of PCOS.

Lim *et al.* [8] investigated the potential usage of radial pulse wave parameters that are utilized in traditional Chinese medicine using ML to classify and predict cases of PCOS. 459 individuals were allocated into two groups. It was found that long short-term memory (LSTM) and voting models were the best and both recorded the same accuracy rate of 72.0174 precision and AUC of 0.715 and 0.722, respectively. All these mean that radial pulse wave analysis is useful for the early detection and management of PCOS.

Yamini *et al.* [9] performed a PCOS prediction study where they considered different ML models by using clinical, hormonal, and biomarkers. The study was conducted using data which included attributes obtained from women with PCOS and from women who did not have the condition. It employed several ML models of logistic regression (LR), random forest (RF), support vector machine (SVM), naive Bayes (NB) classifier, k-nearest neighbors (KNN) and XGBoost to build predictive models. It is also crucial to state that among all the models that have been applied, the RF model emerged to be the most accurate with a rate of 90% accuracy.

Zad *et al.* [10] conducted a study where they aimed to predict PCOS using ML algorithms on electronic health records from a hospital. The study with a 30,601 dataset size evaluated the prediction of PCOS using LR, SVM, gradient boosted trees, and random forests. Hormone levels and obesity contributed the most to the prediction of the disease in this study. The model achieved AUC scores of 85%, 81%, 80%, and 82% for the different models. This research shows that ML can effectively predict PCOS by examining outpatient data which can help in early diagnosis and reducing long term health consequences.

Na *et al.* [11] used International Gene Expression Omnibus data for their research. The major aim of the research was to discover functional PCOS biomarkers and correlate them with immunological infiltration. They identified the correlation between two biomarkers and the infiltrating immune cell types was high, suggesting the two biomarkers might also be implicated in the pathophysiology of PCOS.

Poorani *et al.* [12] performed the investigation of classification of PCOS from the relevant ultrasound images of ovaries. The present research estimated PCOS in the absence of professional guidance from doctors using images of ultrasonic technology. The best CNN algorithm was found to be the algorithm that provided the highest accuracy in classifying ultrasound images of PCOS and non-PCOS images. This demonstrated the suitability of CNN for early screening of PCOS cases.

Kaur *et al.* [13] used a transfer based deep learning technique for detecting PCOS using ultrasound images. Their model used the InceptionV3 architecture which was trained before on a general medical image dataset. It aimed to classify ultrasound images as PCOS infected or not. The approach used a large dataset of ultrasound images which enhanced the training of the model with transfer learning techniques. The model exhibited a remarkable classification accuracy of 99.48%, demonstrating the usefulness of transfer learning in augmenting the precision of PCOS recognition in ultrasound images. This study shows how advanced deep learning techniques can significantly improve the precision of diagnosis and help medical professionals identify and treat PCOS early.

Kumar *et al.* [14] implemented MobileNet to predict PCOS from ultrasound images. This study is relevant because it provides a less time-consuming method to diagnose PCOS and a more accurate method of doing it. The use of higher-order neural networks is a good example of when ML is used to assist in the medical field, especially in terms of diagnosing complex conditions like PCOS.

Kapadia *et al.* [15] conducted a study using multinomial LR to test the prediction risk of PCOS from clinical and demographic factors. This study employed the use of data that was sourced from an online survey comprising a high number of participants who were unaware if they had PCOS or not. The LR model was trained for the classification of the tendency of PCOS occurrence with the current accuracy of 82%. The performance of the model also included the mean cross-validation score of 0.75%.

While there is substantial literature data about the use of ML and DL for the diagnosis and treatment of PCOS, there are several important areas of development that could help to improve existing diagnostic capabilities and optimize treatment processes. First of all, most of the existing studies investigating the applicability of various ML models in the prediction of PCOS are based on simulations, not the real-life use, which implies that while several ML-based approaches are confirmed to be rather efficient in terms of PCOS prediction, the utilization of these methods in the clinical practice is still quite limited. It is also concluded that more data with higher variability concerning the offered details referring to demographic and genetic characteristics require to be generated for enhancing the applicability of the models for the prediction. Lastly, the focus on individual ML/DL models overlooks the potential benefits of ensemble approaches or hybrid systems that could offer improved predictive accuracy and robustness. The contributions this paper makes are as follows: i) Curated and refined a comprehensive dataset for enhanced accuracy in PCOS prediction analysis; ii) Developed and optimized a deep learning model for high-precision diagnosis and management of PCOS; and

iii) Implemented a combination of ML and DL techniques to improve predictive performance and support personalized treatment strategies in PCOS care.

This paper begins with a review of the literature that provides a comprehensive overview of related existing studies, followed by the methodology including the dataset description and pre-processing. The methodology outlines the application of various ML and DL algorithms for predicting PCOS. The paper then presents results with visualizations of the outputs. Finally, it offers avenues for future research directions and a summary of the major discoveries made during the study.

2. METHOD

This section outlines detailed steps from data preprocessing to model optimization and comparative analysis of predictive performances are discussed to ensure the effectiveness of the diagnostic models. This data was gathered to perform the investigation of PCOS with the use of ML approaches. It contains records from 541 patients of PCOS and includes many important features that are used in the diagnosis of PCOS: hormones and metabolic profiles. This dataset description is in Table 1. It has been implemented and curated by Praseon Kottarathil in 2020 and hosted at Kaggle [16] for the use of medical researchers.

Key demographic details include age, weight, height, and blood group, with a derived metric of body mass index (BMI) partially available for 242 patients. Clinical parameters cover a range of reproductive and metabolic factors, including menstrual cycle characteristics, hormonal profiles (Follicle-stimulating hormone (FSH), luteinizing hormone (LH), thyroid-stimulating hormone (TSH), Anti-Müllerian hormone (AMH), Prolactin (PRL)), and glucose levels. Fertility-related measures include the number of follicles in the ovaries (both left and right) and endometrial thickness. Vital signs, including pulse rate and respiratory rate, are uniformly recorded across the cohort. Lifestyle factors that could influence PCOS symptoms or management, such as dietary habits (consumption of fast food), physical activity, and specific symptoms like gaining weight, hair growth/loss, darkening of the skin, and presence of pimples/acne, are included. Hormonal assays, particularly beta-HCG levels, are recorded in two instances (noted as “I beta-HCG” and “II beta-HCG”), providing insights into the hormonal milieu potentially affecting PCOS pathophysiology. The dataset also addresses cardiovascular health with systolic and diastolic blood pressure readings. FSH/LH ratio and waist: hip ratio, present significant amounts of missing data, available for only 9 patients each. This limitation underscores the necessity for careful handling of these variables during analysis.

Table 1. Data description of PCOS dataset used in research

Attribute	Part 1		Attribute	Part 2	
	Missing value	Data type		Missing value	Data type
SI. No	0	int64	Waist (inch)	0	int64
Patient file No.	0	int64	Waist: Hip ratio	532	float64
PCOS (Y/N)	0	int64	TSH (mIU/L)	0	float64
Age (yrs)	0	int64	AMH (ng/mL)	0	object
Weight (Kg)	0	float64	PRL (ng/mL)	0	float64
Height (Cm)	0	float64	Vit D3 (ng/mL)	0	float64
BMI	299	float64	PRG (ng/mL)	0	float64
Blood Group	0	int64	RBS (mg/dl)	0	float64
Pulse rate (bpm)	0	int64	Weight gain (Y/N)	0	int64
RR (breaths/min)	0	int64	Hair growth (Y/N)	0	int64
Hb (g/dl)	0	float64	Skin darkening (Y/N)	0	int64
Cycle (R/I)	0	int64	Hair loss (Y/N)	0	int64
Cycle length (days)	0	int64	Pimples (Y/N)	0	int64
Marriage status (Yrs)	1	float64	Fast food (Y/N)	1	float64
Pregnant (Y/N)	0	int64	Reg.Exercise (Y/N)	0	int64
No. of abortions	0	int64	BP Systolic (mmHg)	0	int64
I beta-HCG (mIU/mL)	0	float64	BP Diastolic (mmHg)	0	int64
II beta-HCG (mIU/mL)	0	object	Follicle No. (L)	0	int64
FSH (mIU/mL)	0	float64	Follicle No. (R)	0	int64
LH (mIU/mL)	0	float64	Avg. F size (L) (mm)	0	float64
FSH/LH	532	float64	Avg. F size (R) (mm)	0	float64
Hip (inch)	0	int64	Endometrium (mm)	0	float64
			Unnamed: 44	539	object

2.1. Data processing

Data preprocessing enhances the quality of data by ensuring that the data fed into the model is clean, accurate, and reliable, which is essential for obtaining meaningful results. Steps to preprocess data are shown in Figure 1. “Patient File No.”, “SI No”, and “Unnamed: 44”, are irrelevant columns and are removed. “AMH (ng/mL)” and “beta-HCG (mIU/mL)” are converted from object data types to numeric.

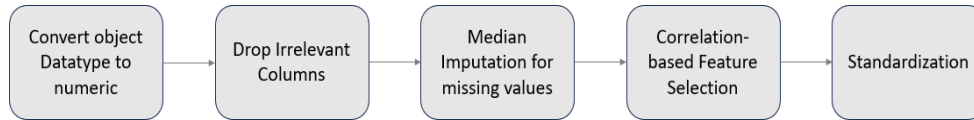


Figure 1. Preprocessing steps for the dataset

Missing values in the “BMI” column are imputed using the formula:

$$BMI = \frac{Weight (kg)}{(Height(m))^2} \quad (1)$$

For “FSH/LH”, the missing values are calculated by dividing the *FSH* column by *LH* column:

$$FSH/LH = \frac{FSH(mIU/mL)}{LH(mIU/mL)} \quad (2)$$

And for “Waist: Hip Ratio”, missing values are determined by dividing the “Waist (inch)” by “Hip (inch)”:

$$Waist: Hip Ratio = \frac{Waist(inch)}{Hip(inch)} \quad (3)$$

The columns “marriage status (Yrs)”, “II beta-HCG (mIU/mL)” and “AMH (ng/mL)” have their missing values filled with the median. Based on their association with PCOS, the target variable, a subset of attributes is chosen by measuring the Pearson correlation coefficient [17] values with the target variable. Features that show high correlation are selected for further analysis. The Pearson correlation coefficient is calculated using (4).

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

where r_{xy} is the Pearson correlation coefficient between two variables x and y , x_i and y_i are the values of the x and y variables for the i^{th} observation, \bar{x} and \bar{y} are the means of x and y variables, respectively, and n is the number of observations. Figure 2 represents the correlation values of all the attributes with the target variable PCOS and Figure 3 represents selected attributes and their correlation with the target variable, respectively. The correlation matrix for the attributes is represented in Figure 4. These features are chosen because they have a significant impact on the development of PCOS based on literature [18]–[20].

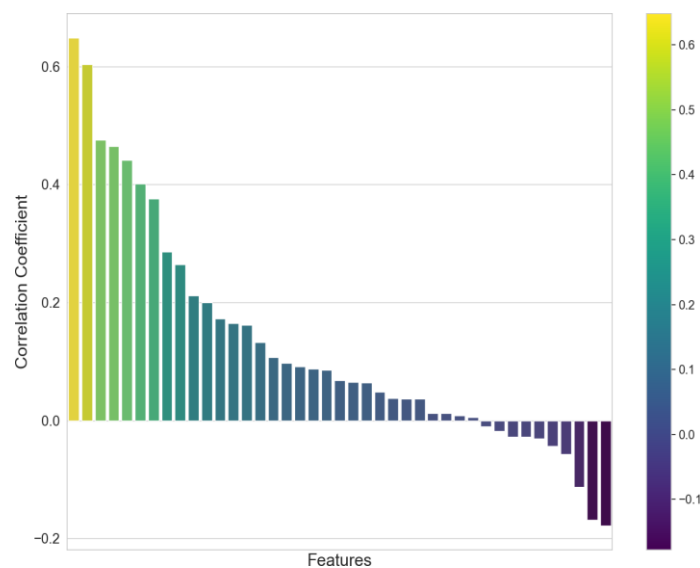


Figure 2. Features before correlation analysis

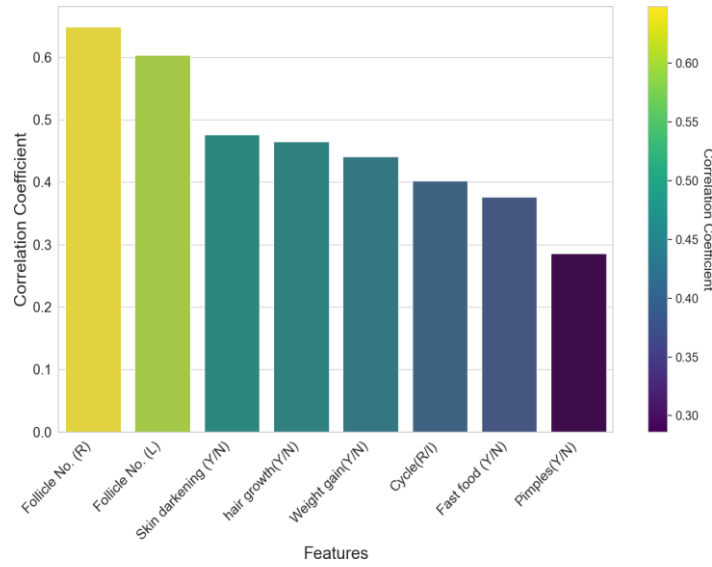


Figure 3. Selected features for model

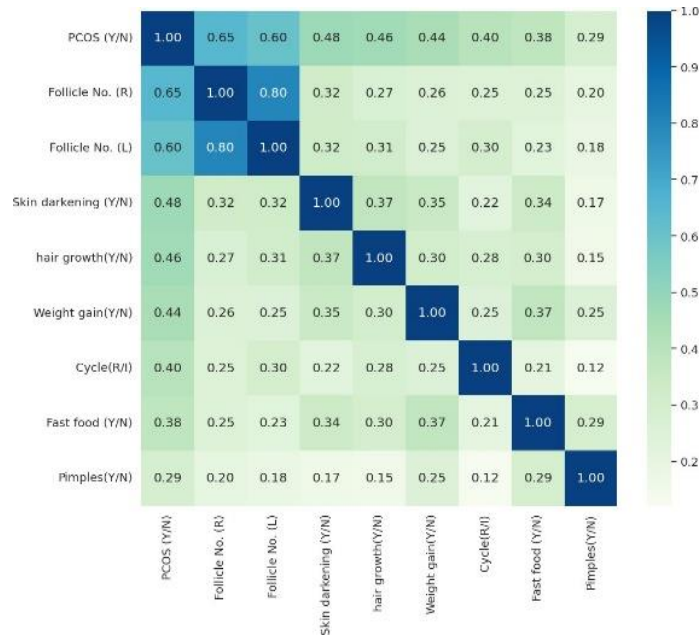


Figure 4. Correlation matrix after feature selection

Standardization ensures that all features have a similar scale which helps in reducing the impact of outliers and ensures that the model does not become biased towards features with larger scales. The StandardScaler from sklearn [21] transforms a feature x using the formula:

$$z = \frac{(x-\mu)}{\sigma} \tag{5}$$

where the standard deviation and mean of the feature values are denoted by σ and μ , respectively. The resulting standardized feature z has a mean of 0 and a standard deviation of 1.

2.2. Model architecture

The effective prediction of PCOS depends on the design of the suggested neural network. The proposed model is structured as a Sequential model within the TensorFlow framework, facilitating a linear stack of layers [22] as shown in Figure 5. To optimize the model, we employ two distinct yet complementary

approaches for hyperparameter tuning: Optuna and genetic algorithms (GA). Both methods are integrated to refine the architecture by pinpointing the optimal values of parameters. These methods of optimization have proved to be effective in hyperparameter tuning as seen in [23].

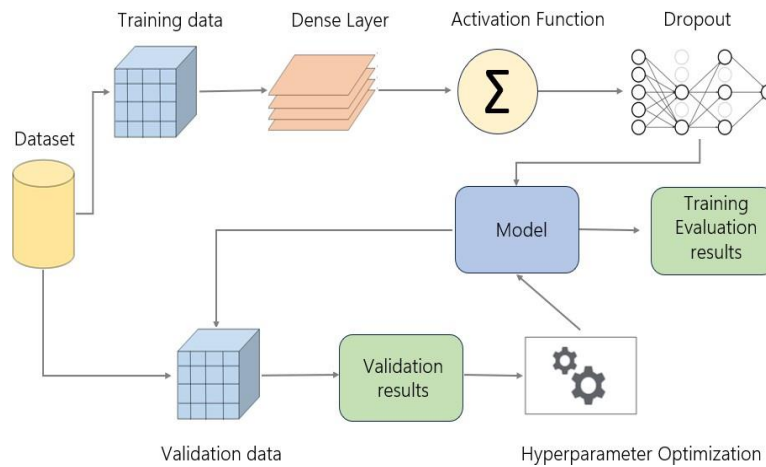


Figure 5. Methodology diagram for proposed deep learning model

2.2.1. Dense layers

Dense layers contain all input neurons which are linked to every neuron in the current layer. The typical function of development of these Dense layers is to acquire intricate structures of the high-dimensional data related to PCOS. Through hyperparameter tuning, the number of neurons in each Dense layer is defined, and it is set between 32 and 512. This helps the model to regulate how complex the model's structure will be according to the structure of the given data. Along with rectified linear unit (ReLU), hyperbolic tangent (Tanh), exponential linear unit (ELU) are used as activation function allowing the model to embody a lot of subtle dependencies of the data [24].

2.2.2. Dropout layers

Incorporated within the architecture are dropout layers, which serve as a regularization technique to reduce overfitting [25]. During training, the proportion of neurons is randomly dropped (i.e., their output is set to zero) according to a rate determined again by hyperparameter optimization, ranging from 0.0 to 0.5. This introduces sparsity in the neuron activation and forces the model to learn robust features that generalize well to unseen data.

2.2.3. Compiling the model

Adam optimizer is chosen for its adaptive learning rate capabilities, which aids in faster convergence. Given that the PCOS prediction job is binary in nature, the loss function of choice is binary cross-entropy. The performance metric selected is accuracy, which provides a direct measure of model success in classifying the PCOS condition.

2.2.4. Early stopping

To further mitigate the risk of overfitting, an EarlyStopping callback is employed to monitor the validation loss for a set number of epochs, specified with patience of five, and halts the training process if no improvement is observed, ensuring that it retains its predictive power on new, unseen data.

2.3. Dynamic parameter optimization

To fine-tune the predictive model for PCOS, two distinct and advanced hyperparameter optimization techniques were independently utilized: Optuna and genetic algorithms (GA). Each technique independently assesses the impact of the following parameters on model performance and optimizes the same set of hyperparameters. The activation functions for the dense layers, identifying which function leads to the highest validation accuracy. Activation functions are essential because they add non-linearities to the model, which helps it learn intricate patterns that are beyond the scope of linear models. In this study, we examine three distinct activation functions.

The ReLU activation function is defined as $f(x) = \max(0, x)$ preserves the linear behavior for positive inputs and sets negative inputs to zero. This feature streamlines computations and helps in overcoming the vanishing gradient problem. The ELU enhances ReLU by allowing negative values when inputs are less than zero. It is defined as $f(x) = x$ for $x > 0$ and $f(x) = \alpha(ex - 1)$ for $x \leq 0$. This process helps in mitigating the “dying ReLU problem” and helps in achieving balanced activation levels that lead to quicker convergence. The Tanh function is defined by $f(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. It outputs values between -1 and 1. This centers the output range around zero, which can enhance the learning efficiency by normalizing the data’s meaning close to zero. This normalization is useful for optimization of the backpropagation process. The performance of this method is compared with other strategies to determine the most effective optimization approach for the proposed model. This comparison helps in identifying the optimal method that could lead to improvements in model accuracy and efficiency.

2.3.1. Optuna optimization

Optuna is an advanced framework designed for hyperparameter optimization, renowned for its capability to efficiently navigate complex parameter spaces [26]. In contrast to traditional approaches like grid or random search, Optuna employs Bayesian optimization methods. These methods utilize historical evaluation data to inform and steer the search process towards the most promising areas of the parameter space, enhancing the likelihood of finding optimal settings more efficiently. Optuna is an advanced framework that has been developed for hyperparameter optimization and is well-known for being able to efficiently move through complicated parameter spaces [26].

2.3.2. Bayesian optimization

This model is less expensive to evaluate and is used to predict the performance of different hyperparameter configurations based on historical data. The Gaussian process is characterized by mean function ($\mu(x)$) that predicts the expected outcome for a given parameter set and covariance Function ($k(x, x')$) that describes the relationship between points in the input space. The predictive distribution at any new point x , given observed data D , is normally distributed as (6):

$$f(x|D) \sim N(\mu(x), \sigma^2(x)) \quad (6)$$

where,

$$\mu(x) = \mu_0(x) + K(X, x)^T [K(X, X) + \sigma^2 n I]^{-1} (y - \mu_0(X)) \quad (7)$$

$$\sigma^2(x) = k(x, x) - K(X, x)^T [K(X, X) + \sigma^2 n I]^{-1} K(X, x) \quad (8)$$

Here, μ_0 is the prior mean, K represents the kernel matrix, X are the observed points, y are the observed targets, and σ^2 is the noise term. The expected improvement (EI) criterion is employed to decide which new points to evaluate. EI measures the expected amount by which a proposed set of parameters is predicted to improve over the current best-known value f_{best} . It is defined as (9):

$$EI(x) = \mathbb{E}[\max(f(x) - f_{best}, 0)] \quad (9)$$

For a Gaussian process $GP(\mu, \sigma^2)$, the expected improvement can be expressed as (10):

$$EI(x) = (\mu(x) - f_{best} - \xi)\Phi(Z) + \sigma(x)\phi(Z) \quad (10)$$

where $Z = \frac{\mu(x) - f_{best} - \xi}{\sigma(x)}$, and ξ is a small positive number facilitating the trade-off between exploration and exploitation. Φ and ϕ are the cumulative distribution function and probability density function of the standard normal distribution, respectively. This mechanism balances exploration of untested areas and exploitation of known good areas, thereby facilitating rapid convergence to the optimal solution.

2.3.3. Genetic algorithm

The genetic algorithm (GA) is a hyperparameter optimization method [27] implemented to adjust the model’s architecture. This algorithm iterates over multiple generations to evolve an optimal set of hyperparameters based on a predefined fitness function. The primary constants of the GA used in this study are defined in Table 2.

The fitness of each individual, measured as the accuracy of the model on a validation set, directs the selection process toward more promising hyperparameter sets. Optuna and GA optimize the model’s architecture by experimenting with the same set of hyperparameters. Each method’s results are critically

evaluated, comparing the highest validation accuracy achieved to determine the optimal hyperparameter configuration for the PCOS prediction model.

Table 2. Genetic algorithm parameters

Primary Constant	Function	Value
Population size	Defines the number of individual solutions in the population. A larger population size allows for a more diverse genetic pool, while a smaller size ensures a more computationally efficient process.	10
Mutation rate	Governs the probability of random alterations in the hyperparameters. Mutation introduces genetic diversity and helps to avoid local optima by enabling explorative steps in the hyperparameter space.	0.2
Crossover rate	Determines how often two individuals in the population will be combined to produce offspring. It represents the balance between preserving successful characteristics and introducing new ones.	0.5
Number of generations	Total generations to run	10

2.4. Model training and evaluation

The dataset underwent a standard 70/15/15 split for training, testing and validation respectively. During training, a batch size of 32 was utilized, with a maximum of 50 epochs. Validation loss was monitored throughout the training process. Subsequently, binary cross-entropy loss and accuracy were applied to the test set as evaluation measures, offering a thorough analysis of the predictive power of the model. Following optimization, the optimal model architecture and hyperparameters were determined and used for training the final model. Evaluation of the final model's performance was based on its accuracy on the validation set.

2.5. Comparison of machine learning models

Logistic regression (LR), support vector classification (SVC), random forest classifier (RF), naive Bayes and k-nearest neighbors (KNN) are tested in addition to the DL model. They are trained and evaluated using the same training and test sets, and their predictive performance is assessed through a series of metrics. LR is selected for its foundational simplicity and high interpretability, particularly valuable in clinical settings where decision-making relies on transparent model insights [28]. SVC is renowned for its proficiency in navigating the challenges of high-dimensional spaces, crafting intricate decision boundaries where linear separability is not feasible [29]. RF contributes a layer of robustness to overfitting, leveraging an ensemble of decision trees to enhance predictive reliability and maintain interpretability [30]. Naive Bayes offers computational efficiency, especially advantageous when dealing with categorical variables that are prevalent in medical datasets [31]. Lastly, KNN algorithm demonstrates its merit in datasets that manifest complex, non-linear decision boundaries [32].

2.6. Performance analysis

For evaluating model performance, several metrics are employed. Accuracy measures the proportion of correct predictions among all predictions, offering a broad view of model effectiveness. The F1 Score strikes a balance between prediction precision and recall by acting as the harmonic mean of precision and recall. While recall evaluates the capacity to recognize every true positive, precision shows the accuracy of positive predictions. When these indicators are taken into account together, accuracy and the coverage of positive cases are guaranteed, offering a thorough evaluation of the model's performance. The equations for these metrics are given in (11)-(14). Here, TP represents true positives, TN denotes true negatives, FP signifies false positives, and FN indicates false negatives.

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (14)$$

The models' performance is also assessed using the receiver operating characteristic (ROC) curve. A reference line charts the sensitivity (true positive rate) against the 1-specificity (false positive rate) for various levels of threshold setting [33].

3. RESULTS AND DISCUSSION

In this section, we delve into the performance evaluation of our DL model for PCOS alongside a comparative analysis with various machine learning models. We highlight each model's advantages and disadvantages through this thorough examination, as well as provide insight into how well-suited each is for PCOS prediction tasks.

3.1. Assessment of the deep learning model's performance

Our research into the predictive modeling of PCOS has been marked by a strategic deployment of deep learning techniques, optimized to achieve high performance. The Optuna optimization framework guided us to a deep learning model configuration that exhibited an accuracy of 93.55%. The resultant configuration of the proposed model is presented in Table 3. This model is distinct in its architecture, featuring 224 input layer units and two hidden layers with 160 and 96 units respectively. Tanh activation function is employed by the first hidden layer, and ReLU activation function is used by the second hidden layer, a structure that enables nuanced handling of the non-linear intricacies of medical diagnosis data. A considered dropout rate of 9.417% contributes to the model's resilience against overfitting, optimizing its performance on unseen data.

Table 3. Configuration of the proposed deep learning model for prediction of PCOS using Optuna as the hyperparameter optimization technique

Parameter	Value
Input layer units	224
Input activation function	Exponential linear unit (ELU)
Number of hidden layers	2
Units in hidden layer 0	160
Activation function of hidden layer 0	Hyperbolic tangent (Tanh)
Units in hidden layer 1	96
Activation function of hidden layer 1	Rectified linear unit (RELU)
Dropout rate	9.417%
Learning rate	0.000372
Accuracy	93.55%

Comparisons using a GA for hyperparameter optimization, as shown in Table 4, yielded a slightly less accurate model with 91.4% accuracy. Despite its high complexity, the GA model's performance was marginally lower, highlighting Optuna's effectiveness in exploring the hyperparameter space to fine-tune deep learning models for PCOS prediction in particular.

3.2. Comparative evaluation with machine learning models

Examining the metrics in Table 4 alongside the ROC-AUC curves in Figure 6, we establish a nuanced understanding of each model's ability to identify PCOS cases effectively. Figure 7 shows the comparison between classification metrics and AUC scores for all the models. SVC stands out as a high-performing traditional machine learning model, with the highest accuracy, precision, recall, and F1-Score among non-deep learning algorithms. This superior performance is attributed to its proficiency in managing high-dimensional data and finding the most distinct decision boundary, as reflected in its high AUC of 0.97, which is on par with the deep learning model optimized with Optuna. This consistency between the table metrics and ROC curve indicates a strong capacity for correctly classifying both positive and negative cases. KNN and LR present as competent models with accuracies over 92%, supported by their balanced precision and recall. KNN benefits from detecting local data structures, while Logistic Regression's strengths lie in capturing linear correlations. However, these models fall slightly short compared to the deep learning models, suggesting that while they perform well, they may not fully grasp the intricate complexities of the PCOS dataset.

The Gaussian naive Bayes, despite its moderate performance and the lowest accuracy next to random forest, displays a high recall, suggesting it is adept at identifying true PCOS cases but prone to false positives, due to its assumption of feature independence. The model's AUC of 0.91, while reasonable, highlights how limited its assumption is when dealing with intricate medical data. The random forest classifier lags in performance with the lowest accuracy and F1-Score, and an AUC of 0.86, implicating possible overfitting or a failure to capture essential feature interactions. Its lower position in both Table 5 and the ROC curve analysis points towards its less effective handling of the PCOS prediction task within this specific dataset. Our proposed deep learning models optimized with both GA and Optuna demonstrate high accuracy, precision, recall, and F1-Scores, reflecting their effectiveness for this application. The Optuna-

optimized model slightly edges out the GA-optimized model, implying Optuna’s optimization strategy generalizes better and shows strong discrimination capabilities.

Table 4. Configuration of the proposed deep learning model for prediction of PCOS using genetic algorithm as the hyperparameter optimization technique

Parameter	Value
Input layer units	187
Input activation function	Exponential linear unit (ELU)
Number of hidden layers	1
Units in hidden layer 0	76
Activation function of hidden layer	Rectified linear unit (RELU)
Dropout rate	20%
Learning rate	0.002649
Accuracy	91.4%

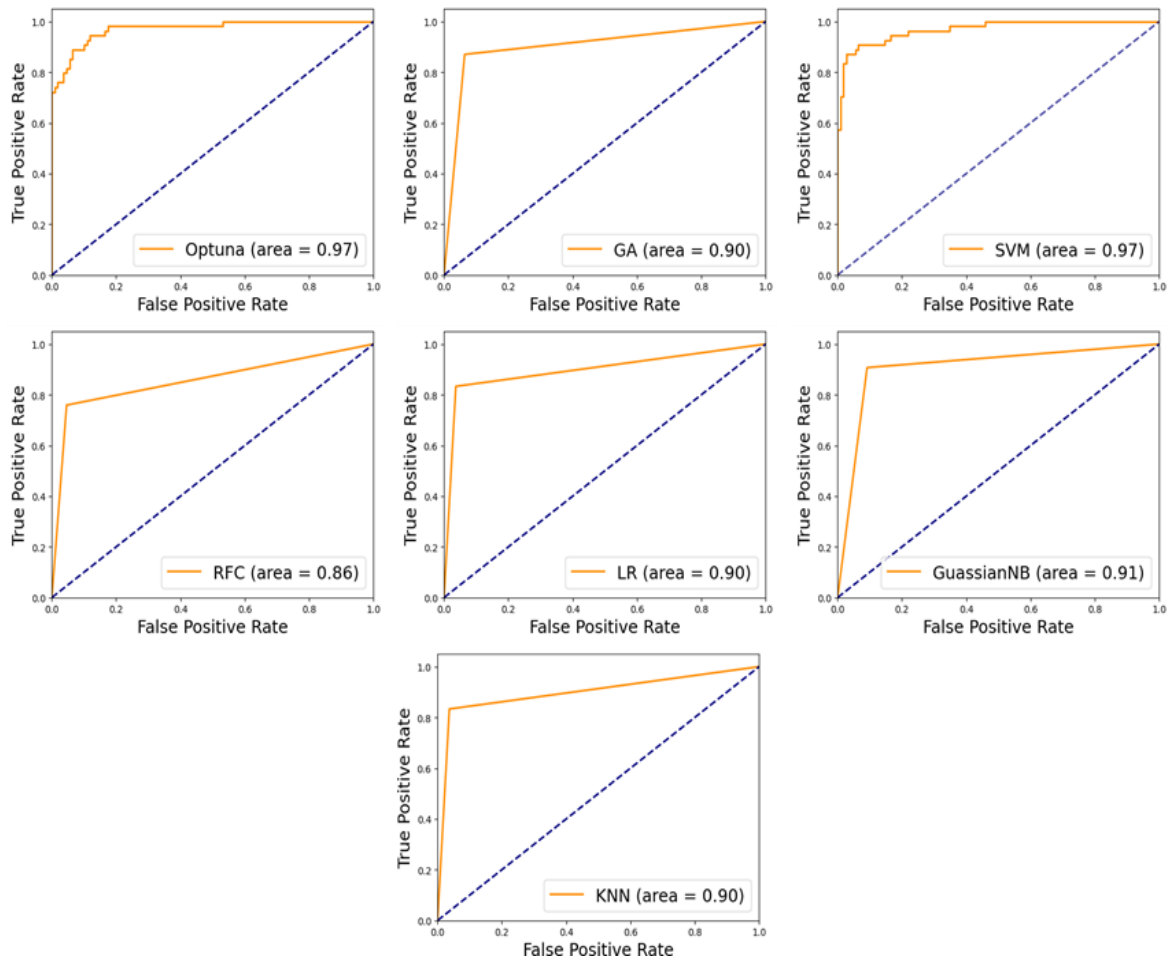


Figure 6. ROC curves for comparative analysis of model performances

Table 5. Model performance of algorithms for PCOS prediction

Model Name	Accuracy	Precision	Recall	F1-Score
SVC	93.25%	0.93	0.91	0.92
KNeighbors	92.02%	0.92	0.90	0.91
Gaussian NB	90.79%	0.89	0.91	0.90
Random forest	87.73%	0.87	0.84	0.86
Logistic regression	92.02%	0.92	0.90	0.91
Proposed model (Using GA)	91.4%	0.90	0.90	0.90
Proposed model (Using Optuna)	93.55%	0.91	0.90	0.90

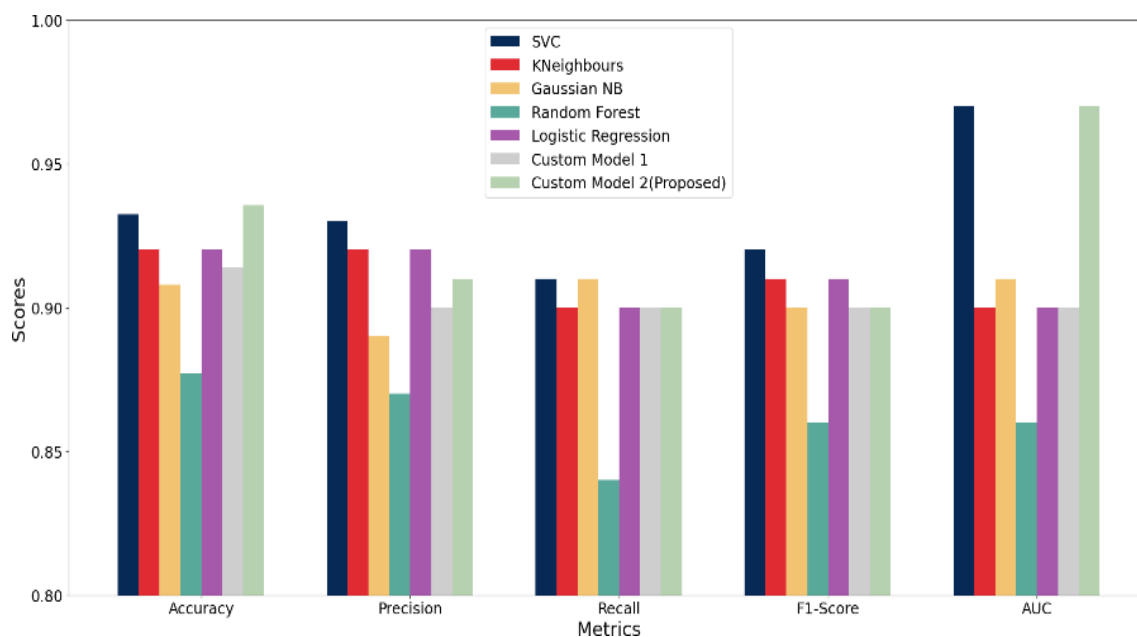


Figure 7. Comparative analysis of model performances based on classification metrics and AUC

4. CONCLUSION

In this study, we systematically developed a model and identified the most effective approach for predicting PCOS on a PCOS dataset from Kerala hospitals. Our exploration began with the meticulous preparation of the data by extensive preprocessing and feature engineering, followed by the application of both conventional machine learning models and advanced deep learning techniques that were carefully optimized for accuracy. SVC and the deep learning model developed by the Optuna were among the features that were distinguished by high performance among those that were tested. The high-performance deep learning model was also built using the Optuna framework and achieved 93.55 accuracy. This value shows that the model is accurate in diagnosing PCOS patients since it has a robust architecture, and hyper parameters are more powerful. The use of the publicly available dataset for this study was of significant benefit in that it increased the replicability and hence the transparency of this research.

Implementing the model in a clinical setting to validate its effectiveness in real-world scenarios would prove to be a vital step towards practical application of the model. Longitudinal studies could be utilized to assess the models' ability to track PCOS over time. This can aid in the development of personalized treatment plans for patients.





REFERENCES

- [1] M. T. Sheehan, "Polycystic ovarian syndrome: diagnosis and management," *Clinical Medicine and Research*, vol. 2, no. 1, pp. 13–27, Feb. 2004, doi: 10.3121/cmr.2.1.13.
- [2] A. A. Choudhury and V. D. Rajeswari, "Polycystic ovary syndrome (PCOS) increases the risk of subsequent gestational diabetes mellitus (GDM): a novel therapeutic perspective," *Life Sciences*, vol. 310, Dec. 2022, doi: 10.1016/j.lfs.2022.121069.
- [3] J. P. Christ and M. I. Cedars, "Current guidelines for diagnosing PCOS," *Diagnostics*, vol. 13, no. 6, p. 1113, Mar. 2023, doi: 10.3390/diagnostics13061113.
- [4] M. Gibson-Helm, H. Teede, A. Dunaif, and A. Dokras, "Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome," *Journal of Clinical Endocrinology and Metabolism*, vol. 102, no. 2, pp. 604–612, Dec. 2017, doi: 10.1210/je.2016-2963.
- [5] S. Palomba, S. Santagni, A. Falbo, and G. B. La Sala, "Complications and challenges associated with polycystic ovary syndrome: current perspectives," *International Journal of Women's Health*, vol. 7, pp. 745–763, Jul. 2015, doi: 10.2147/IJWH.S70314.
- [6] A. Rahman *et al.*, "Machine learning and deep learning-based approach in smart healthcare: recent advances, applications, challenges and opportunities," *AIMS Public Health*, vol. 11, no. 1, pp. 58–109, 2024, doi: 10.3934/publichealth.2024004.
- [7] F. J. Barrera *et al.*, "Application of machine learning and artificial intelligence in the diagnosis and classification of polycystic ovarian syndrome: a systematic review," *Frontiers in Endocrinology*, vol. 14, Sep. 2023, doi: 10.3389/fendo.2023.1106625.
- [8] J. Lim *et al.*, "Machine learning classification of polycystic ovary syndrome based on radial pulse wave analysis," *BMC Complementary Medicine and Therapies*, vol. 23, no. 1, Nov. 2023, doi: 10.1186/s12906-023-04249-5.
- [9] B. Yamini, V. R. Kaneti, P. Prema, C. Ambhika, M. Nalini, and R. Siva Subramanian, "Machine learning-driven PCOS prediction for early detection and tailored interventions," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 10, no. 9, pp. 61–75, Sep. 2023, doi: 10.14445/23488379/IJEEEE-V10I9P106.
- [10] Z. Zad *et al.*, "Predicting polycystic ovary syndrome with machine learning algorithms from electronic health records," *Frontiers in Endocrinology*, vol. 15, Jan. 2024, doi: 10.3389/fendo.2024.1298628.




- [11] Z. Na, W. Guo, J. Song, D. Feng, Y. Fang, and D. Li, "Identification of novel candidate biomarkers and immune infiltration in polycystic ovary syndrome," *Journal of Ovarian Research*, vol. 15, no. 1, Jul. 2022, doi: 10.1186/s13048-022-01013-0.
- [12] B. Poorani and R. Khilar, "Classification of PCOS using machine learning algorithms based on ultrasound images of ovaries," in *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Apr. 2023, pp. 1–7, doi: 10.1109/ICONSTEM56934.2023.10142359.
- [13] N. Kaur, G. Gupta, and P. Kaur, "Transfer-based deep learning technique for PCOS detection using ultrasound images," in *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, Sep. 2023, pp. 1–6, doi: 10.1109/NMITCON58196.2023.10276245.
- [14] A. S. Kumar, S. Annamalai, M. Kumaresan, P. Manikandan, R. Sekaran, and H. A. Pai, "CNN-based analysis of ultrasound images for PCOS diagnosis," in *Proceedings - International Conference on Technological Advancements in Computational Sciences, ICTACS 2023*, Nov. 2023, pp. 347–350, doi: 10.1109/ICTACS59847.2023.10390451.
- [15] D. Kapadia and R. Jain, "Pcos Prediction: advancements in medical informatics using artificial intelligence," in *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, Nov. 2023, pp. 1–5, doi: 10.1109/INCOFT60753.2023.10425743.
- [16] P. Kottarathil, "Polycystic ovary syndrome (PCOS)," *Kaggle*, 2020.
- [17] W. Kirch, "Pearson's correlation coefficient," in *Encyclopedia of Public Health*, Dordrecht: Springer Netherlands, 2008, pp. 1090–1091.
- [18] J. P. Christ *et al.*, "Follicle number, not assessments of the ovarian stroma, represents the best ultrasonographic marker of polycystic ovary syndrome," *Fertility and Sterility*, vol. 101, no. 1, pp. 280–287, Jan. 2014, doi: 10.1016/j.fertnstert.2013.10.001.
- [19] S. Islam, N. Nabi, S. A. Khushbu, N. J. Ria, and A. K. M. Masum, "A process of finding common symptoms and diagnosis age among PCOS patients through a survey," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2021, pp. 1–7, doi: 10.1109/ICCCNT51525.2021.9580114.
- [20] A. Radwan *et al.*, "The association of polycystic ovarian syndrome among reproductive-aged women with consumption of junk food in Jeddah, Saudi Arabia," *Cureus*, Nov. 2023, doi: 10.7759/cureus.48299.
- [21] A. Abraham *et al.*, "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, vol. 8, 2014, doi: 10.3389/fninf.2014.00014.
- [22] M. Abadi *et al.*, "TensorFlow: large-scale machine learning on heterogeneous distributed systems," *arXiv:1603.04467*, Mar. 2016.
- [23] S. K. Kamath, S. K. Pendekanti, and D. Rao, "LivMarX: an optimized low-cost predictive model using biomarkers for interpretable liver cirrhosis stage classification," *IEEE Access*, vol. 12, pp. 92506–92522, 2024, doi: 10.1109/ACCESS.2024.3422451.
- [24] A. M. Javid, S. Das, M. Skoglund, and S. Chatterjee, "A ReLU dense layer to improve the performance of neural networks," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 2810–2814, doi: 10.1109/ICASSP39728.2021.9414269.
- [25] B. Jabir and N. Falih, "Dropout, a basic and effective regularization method for a deep learning model: A case study," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 2, pp. 1009–1016, Nov. 2021, doi: 10.11591/ijeecs.v24.i2.pp1009-1016.
- [26] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: a next-generation hyperparameter optimization framework," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Jul. 2019, pp. 2623–2631, doi: 10.1145/3292500.3330701.
- [27] F. A. Fortin, F. M. De Rainville, M. A. Gardner, M. Parizeau, and C. Gagné, "DEAP: evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, 2012.
- [28] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, Jul. 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.
- [29] M. Pal and P. M. Mather, "Support vector machines for classification in remote sensing," *International Journal of Remote Sensing*, vol. 26, no. 5, pp. 1007–1011, Mar. 2005, doi: 10.1080/01431160512331314083.
- [30] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, pp. 272–278, 2012.
- [31] F. J. Yang, "An implementation of naive bayes classifier," *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, pp. 301–306, 2018, doi: 10.1109/CSCI46756.2018.00065.
- [32] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient kNN classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.
- [33] R. Kumar and A. Indrayan, "Receiver operating characteristic (ROC) curve for medical researchers," *Indian Pediatrics*, vol. 48, no. 4, pp. 277–287, 2011, doi: 10.1007/s13312-011-0055-4.

BIOGRAPHIES OF AUTHORS






Divya Rao     received the Ph.D. degree in artificial intelligence applied to oncology. Currently, she is an assistant professor senior scale at the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. Her research interests include machine learning, artificial intelligence and healthcare informatics. She can be contacted at email: divya.r@manipal.edu.






Riddhi Rajendra Dayma    is currently pursuing a Bachelor of Technology (B.Tech.) degree in information technology from the Department of Information and Communication Technology at Manipal Institute of Technology. She has actively participated in research projects focusing on the practical applications of deep learning methodologies, particularly in healthcare and human-computer interaction. Her research interests lie within the broader domain of artificial intelligence and its potential impact on society. She can be contacted at email: riddhi.dayma@learner.manipal.edu.



Sanjeev Kushal Pendekanti    is currently pursuing the bachelor's degree in information technology with the Information and Communication Technology Department, Manipal Institute of Technology. He is deeply interested in the field of data mining and machine learning and their applications in the healthcare domain. He aspires to contribute to impactful research in this domain and use technology to make significant improvements to the existing medical sector. He can be contacted at email kushalpendekanti2002@gmail.com.



Aneesha Acharya K.    received the M.Tech. degree (2011) in biomedical engineering and Ph.D. in Department of Electronics and Communication Engineering at Manipal Institute of Technology, Manipal (2022). He has eight publications in peer reviewed journals. His area of interest is biomedical instrumentation and electronics for biomedical applications. He can be contacted at email: ak.acharya@manipal.edu.