

Received 23 August 2022, accepted 6 September 2022, date of publication 12 September 2022, date of current version 20 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3205587



RESEARCH ARTICLE

A Novel Approach for Polycystic Ovary Syndrome Prediction Using Machine Learning in Bioinformatics

used the
Kaggle
data!!

SHAZIA NASIM¹, MUBARAK ALMUTAIRI², KASHIF MUNIR³,
ALI RAZA¹, AND FAIZAN YOUNAS¹

¹Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan 64200, Pakistan

²College of Computer Science and Engineering, University of Hafr Al Batin, Hafr Alabtin 31991, Saudi Arabia

³Faculty of Computer Science and IT, Khawaja Fareed University of Engineering & IT, Rahim Yar Khan 64200, Pakistan

Corresponding authors: Mubarak Almutairi (mutairims@gmail.com), Kashif Munir (kashif.munir@kfueit.edu.pk), and Ali Raza (ali.raza.scholarly@gmail.com)

This work was supported by the University of Hafr Albatin, Saudi Arabia.

ABSTRACT Polycystic ovary syndrome (PCOS) is a critical disorder in women during their reproduction phase. The PCOS disorder is commonly caused by excess male hormone and androgen levels. The follicles are the collections of fluid developed by ovaries and may fail to release eggs regularly. The PCOS results in miscarriage, infertility issues, and complications during pregnancy. According to a recent report, PCOS is diagnosed in 31.3% of women from Asia. Studies show that 69% to 70% of women did not avail of a detecting cure for PCOS. A research study is needed to save women from critical complications by identifying PCOS early. The main aim of our research is to predict PCOS using advanced machine learning techniques. The dataset based on clinical and physical parameters of women is utilized for building study models. A novel feature selection approach is proposed based on the optimized chi-squared (CS-PCOS) mechanism. The ten hyper-parametrized machine learning models are applied in comparison. Using the novel CS-PCOS approach, the gaussian naive bayes (GNB) outperformed machine learning models and state-of-the-art studies. The GNB achieved 100% accuracy, precision, recall, and f1-scores with minimal time computations of 0.002 seconds. The k-fold cross-validation of GNB achieved a 100% accuracy score. The proposed GNB model achieved accurate results for critical PCOS prediction. Our study reveals that the dataset features prolactin (PRL), blood pressure systolic, blood pressure diastolic, thyroid stimulating hormone (TSH), relative risk (RR-breaths), and pregnancy are the prominent factors having high involvement in PCOS prediction. Our research study helps the medical community overcome the miscarriage rate and provide a cure to women through the early detection of PCOS.

INDEX TERMS Bioinformatics, data analysis, infertility, machine learning, pregnancy complications, polycystic ovary syndrome, PCOS prediction, syndrome classification.

I. INTRODUCTION

PCOS is a medical ailment [1] which is the main reason for hormonal disorder in women during their reproduction phase. The PCOS arises due to a disorder in hormones [2]. The hormone disorder results in the ovaries growing small amounts of fluid called follicles (cysts). The ovaries are

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang¹.

unable to produce eggs due to PCOS, which is the prominent problem women with PCOS have critical complications in pregnancy [3]. PCOS disease is usually inherited and is an unexpected critical situation. The time and cost of countless medical tests are a burden for patients and doctors. A machine learning-based platform must be built for efficient and early prediction of PCOS.

The common indications of PCOS are a more ratio of androgen level (heightened male hormones) [4], an

unbalanced menstrual cycle, polycystic ovaries, and metabolism problems. Early detection of PCOS-related symptoms helps to adopt essential lifestyle changes. During pregnancy, the chances of miscarriage in women with PCOS are more than three times that of women without PCOS. Women with PCOS undergo infertility, resulting in gynaecological cancer [5]. Early detection of PCOS results in saving miscarriage.

PCOS affects many women at an early age. However, they are not diagnosed. Numerous studies show that 69% to 70% of women did not avail of a detecting cure [6]. According to a recent report, PCOS is diagnosed in 4.8% of women of white Americans, 8% of African Americans, 6.8% of women in Spain and 31.3% of women in Asia [7]. Due to these complications and statics, early diagnosis of PCOS is crucial.

The PCOS treatment [8] consists of modification in lifestyle, weight reduction, and an appropriate healthy diet plan. The women's everyday workout results in minimized free androgen indexed and reduced biochemical hyperandrogenism [9], [10], [11], [12], [13]. Studies show that with the increase in age, the PCOS symptoms become less extreme, and women get menopause [14], [15], [16].

Machine learning (ML) is the core area of computer science. Nowadays, ML allows computers to learn without going from their environment. The ML performs an essential role in the healthcare department [17]. The ML deals with obscure enormous datasets. The ML analyses the data, transform it into a useable form for clinical procedures and assists in identifying the nature of different diseases. The three main types of machine learning are used in the medical field [18]. Medical Image Processing, NLP in medical documentation, and statistical material about genetics are significant applications. Our primary research contributions are as follows:

- A novel CS-PCOS feature selection approach is proposed based on the optimized chi-squared mechanism. The twenty dataset features with a high importance value are selected using the CS-PCOS approach for building machine learning models. By using the CS-PCOS approach, our proposed model outperformed machine learning techniques and past proposed state-of-the-art studies;
- The PCOS exploratory data analysis (PEDA) is conducted to find the data patterns that are the primary cause of PCOS disease. The PEDa is based on graphs, charts, and statistical data analysis;
- The ten advanced machine learning models are applied in comparison to predict PCOS. The applied machine learning techniques are stochastic gradient descent (SGD), linear regression (LIR), random forest (RF), bayesian ridge (BR), support vector machine (SVM), k-neighbors classifier (KNC), multi-layer perceptron (MLP), logistic regression (LOR), gaussian naive bayes (GNB), and gradient boosting classifier (GBC). The GNB model is our proposed model;
- The k-fold cross-validation is applied to validate overfitting in our applied machine learning models. The ten

folds of research data are used during the k-fold analysis. The machine learning models are generalized and give accurate performance scores for unseen test data.

The remainder of the research study is as follows: Section II is based on the related literature analysis of PCOS. Our research methodology analysis is conducted in Section III. The employed machine learning models for PCOS prediction are examined in Section IV. The scientific results validation and evaluations of our research approaches are analyzed in Section V. The research study concluding remarks are described in Section VI.

II. RELATED WORK

The related literature to our proposed research study is examined in this section. The past applied state-of-the-art study for PCOS prediction is analyzed. The related research findings and proposed techniques are examined.

One of the most common health problems [19] caught in early age women is PCOS disease. PCOS disease is a complicated health dilemma distressing women of childbearing age, which can be identified based on different medical indicators and signs. Accurate identification and detection of PCOS is the essential baseline for appropriate treatment. For this purpose, researchers applied different machine learning approaches such as SVM, random forest, CART, logistic regression and naive bayes classification to identify PCOS patients. After comparing the results, the Random Forest algorithm gave a high performance with 96% accuracy in PCOS diagnostics on a given dataset [20].

Machine learning algorithms were implemented on a dataset of 541 patients, from which 177 have PCOS disease. The dataset consists of 43 features. As all features did not have equal importance, researchers used a feature selection model to rank them according to their value, called the univariate feature selection model. This model is implemented to get ten high-ranked features that can be used to predict the PCOS disease. After splitting the dataset into the train and test portion, different algorithms were implemented to get a result. These models include gradient boosting classifiers [21], logistic regression classifiers, random forest classifiers, RFLR abbreviation of random forest and logistic regression. As a result, the proposed RFLR algorithm achieved a 90.01% accuracy score in classifying the PCOS patients with ten highly ranked features [22].

A new technique was proposed for the early detection and identification of PCOS disease in 2021. The proposed model was based on XGBRF and catBoost. After preprocessing the data, the top 10 attributes were selected by the univariate feature selection method. The classifiers implemented to compare the accuracy results are MLP, decision tree, SVM, HRFLR, random forest, logistic regression, and gradient boosting. Results showed that XGBRF performed with an 89% accuracy score while catBoost outperformed with a 95% accuracy score. The accuracy scores of other classifiers lay between 76% and 85%. The catBoost technique was the best model for the early detection of PCOS disease [23].

Researchers have demonstrated that PCOS identification depends on morphological, biological, clinical processes [24] and methods [25]. Due to advanced technology such as ultrasonography, the surplus follicle has become a critical indicator of polycystic ovarian morphology (PCOM). Since 2003, most researchers have used the inception of twelve follicles (having 2-9 mm measurement in diameter) per complete ovary. However, that now appears to be outdated [26]. The variations in the amount of ovarian volume or having space may also be acknowledged as accurate indicators of PCOS morphology. However, their effectiveness compared with overweight and extra follicles remains mystified.

For the first time, researchers analyzed attributes and characteristics of woman's genes involved in PCOS with a specific pattern and order. The 233 patients with PCOS participated in the prediction process. Researchers used machine learning algorithms such as decision trees and SVM with various kernel features (linear, polynomial, RBF) and k-nearest neighbor (KNN) to predict PCOS by identifying new genes. From these classifiers, SVM (linear) gave the best accuracy performance as it was 80%, and the KNN accuracy score was between 57% to 79% [3].

According to a stat, every 3 to 4 women from 10 are presently distressed from PCOS. To detect and predict PCOS in the first phase, the authors proposed an automated system which can detect and predict PCOS disease for medical treatment. The authors applied five machine learning models: gaussian naïve Bayes, SVM, k-neighbours, random forest, and logistic regression. They used applied models on a dataset with 41 attributes. The top 30 features were selected by a statistical method. After comparing the results of all five models, it was observed that the accuracy of the random forest model is 90%, while the results of the other models were between 86% and 89%. The random forest model was the proposed approach to detect and predict the PCOS patient [27].

The gene expression classification in bioinformatics using a hybrid machine learning framework was proposed [28]. The proposed genetic model is based on a cuckoo search algorithm using an artificial bee colony (ABC). The six benchmark gene expression dataset was utilized for building a naive bayes classifier. The study contributes to high accuracy performance compared to the previously published feature selection techniques. The classification of cancer-based on gene expression using a novel framework was proposed [29]. The ABC-based modified metaheuristics optimization technique was applied for the classification task.

The identification of PCOS using a novel immune infiltration and candidate biomarker was proposed in this study [30]. The proposed approach was the machine learning-based logistic regression and support vector machine models. The five datasets were utilized for training and testing the models. The proposed model achieved a 91% accuracy score for PCOS identification. The study contributes to presenting a novel framework for analysis. The mutational

landscape screening-based modified PCOS-related genes analysis was proposed in this study [31]. The PCOS-related gene data of nsSNPs of the 27 were selected for analysis.

III. METHODOLOGY

Our research study uses the PCOS-related clinical and physical features dataset for machine learning model building. The dataset feature engineering is done by using the novel proposed CS-PCOS approach. The PCOS exploratory data analysis (PEDA) is applied to figure out the data patterns and factors that are the primary cause of PCOS disease. The dataset is fully preprocessed during feature engineering. The preprocessed dataset is split into two portions train and test. The split ratio used is 80% for training and 20% for the model's evaluations on unseen test data. The hyper-parametrized model is completely trained and tested. The proposed model is ready to predict the POCS disease in deployment. The research methodology working flow is examined in Figure 1.

A. POLYCYSTIC OVARY SYNDROME DATASET

The PCOS dataset [32] is utilized in our research study. The clinical and physical parameters of 541 patients are used to create the dataset. The PCOS dataset features are analyzed in Table 1. The dataset contains a total of 41 features. We have filled the null values in our dataset with zero to preprocess the dataset. We have dropped the dataset columns 'Sl. No' and 'Patient File No.' due to containing unnecessary information. The dataset was collected from ten different hospitals across Kerala in India. The memory usage size of the dataset is 177.6 KB.

B. NOVEL CS-PCOS FEATURE ENGINEERING TECHNIQUE

The feature engineering techniques are applied to transform the dataset features into the best fit for a predictive model with high accuracy. A novel CS-PCOS feature selection approach is proposed based on the optimized chi-squared mechanism. The operational flow of feature selection by the CS-PCOS approach is visualized in Figure 2. The proposed CS-PCOS technique checks the independence by comparing the observed frequencies (categorically data) with the expected frequencies (target data). The proposed CS-PCOS technique extracts the vital value statistics based on goodness of fit. The 39 features are input to our proposed feature selection technique and determine the importance values for each feature.

The feature importance values analysis is demonstrated in Table 2. The essential features have the highest value near one. Furthermore, the element which has zero value is non-vital. The Waist Hip Ratio is the most important in the segment. The feature having zero importance values is dropped. The drop features are Age (yrs), Weight (Kg), BMI, Cycle(R/I), Cycle length(days), Marraige Status (Yrs), FSH(mIU/mL), LH(mIU/mL), FSH/LH, AMH(ng/mL), Vit D3 (ng/mL), PRG(ng/mL), Weight

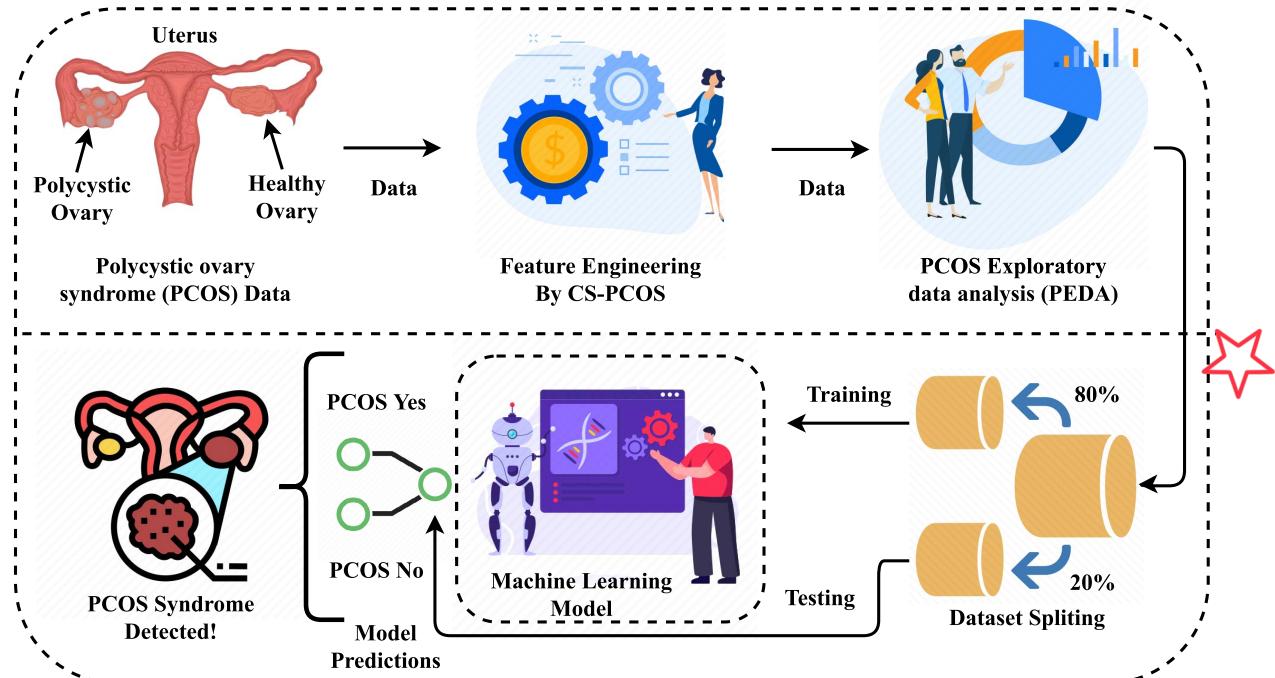


FIGURE 1. The methodological architecture analysis of the proposed research study in predicting the PCOS syndrome.

TABLE 1. The PCOS dataset descriptive feature analysis.

Sr no.	Feature	Non-Null Count	Data Type	Sr no.	Feature	Non-Null Count	Data Type
1	Sl. No	541	int64	22	TSH (mIU/L)	541	float64
2	Patient File No.	541	int64	23	AMH(ng/mL)	540	float64
3	PCOS (Y/N)	541	int64	24	PRL(ng/mL)	541	float64
4	Age (yrs)	541	int64	25	Vit D3 (ng/mL)	541	float64
5	Weight (Kg)	541	float64	26	PRG(ng/mL)	541	float64
6	Height(Cm)	541	float64	27	RBS(mg/dl)	541	float64
7	BMI	541	float64	28	Weight gain(Y/N)	541	int64
8	Blood Group	541	int64	29	hair growth(Y/N)	541	int64
9	Pulse rate(bpm)	541	int64	30	Skin darkening (Y/N)	541	int64
10	RR (breaths/min)	541	int64	31	Hair loss(Y/N)	541	int64
11	Cycle(R/I)	541	int64	32	Pimples(Y/N)	541	int64
12	Cycle length(days)	541	int64	33	Fast food (Y/N)	540	float64
13	Marraige Status (Yrs)	540	float64	34	Reg. Exercise(Y/N)	541	int64
14	Pregnant(Y/N)	541	int64	35	BP_Systolic (mmHg)	541	int64
15	No. of absorptions	541	int64	36	BP_Diastolic (mmHg)	541	int64
16	FSH(mIU/mL)	541	float64	37	Follicle No. (L)	541	int64
17	LH(mIU/mL)	541	float64	38	Follicle No. (R)	541	int64
18	FSH/LH	541	float64	39	Avg. F size (L) (mm)	541	float64
19	Hip(inch)	541	int64	40	Avg. F size (R) (mm)	541	float64
20	Waist(inch)	541	int64	41	Endometrium (mm)	541	float64
21	Waist: Hip Ratio	541	float64				

gain(Y/N), hair growth(Y/N), Skin darkening (Y/N), Hair loss(Y/N), Pimples(Y/N), Fast food (Y/N), Follicle No. (L), and Follicle No. (R). The twenty most prominent features are selected by our proposed technique and used for PCOS prediction in our research study. The selected feature correlation analysis is conducted in Figure 3. The correlation analysis demonstrates that all selected features have a positive correlation.

C. PCOS EXPLORATORY DATA ANALYSIS (PEDA)

This section analyses the PCOS data and the dataset's different patterns to understand the cause of PCOS. The analysis focus on 20 features with a significant value selected by the proposed CS-PCOS technique that is used to train the machine learning models. These features are analyzed from other angles using different graphs. The seaborn, pandas and matplotlib libraries of Python are used to visualize the chart.

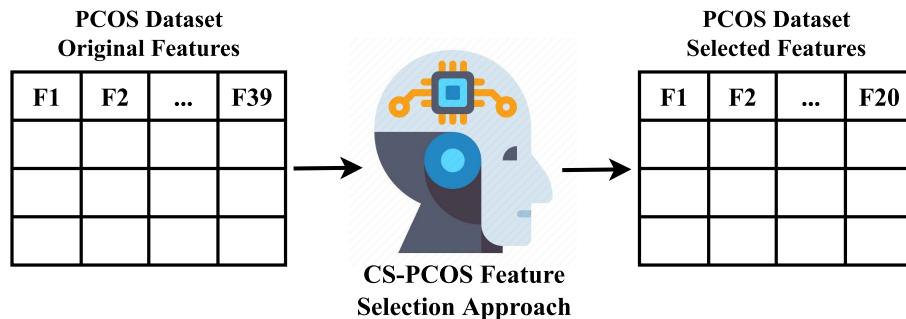


FIGURE 2. The CS-PCOS approach operational flow of feature selection from the original dataset.

TABLE 2. The CS-PCOS approach features importance values analysis.

Sr no.	Feature	Importance value	Sr no.	Feature	Importance value
1	Age (yrs)	0.00	21	AMH(ng/mL)	0.00
2	Weight (Kg)	0.00	22	PRL(ng/mL)	0.72
3	Height(Cm)	0.44	23	Vit D3 (ng/mL)	0.00
4	BMI	0.00	24	PRG(ng/mL)	0.00
5	Blood Group	0.67	25	RBS(mg/dl)	0.03
6	Pulse rate(bpm)	0.27	26	Weight gain(Y/N)	0.00
7	RR (breaths/min)	0.74	27	hair growth(Y/N)	0.00
8	Hb(g/dl)	0.60	28	Skin darkening (Y/N)	0.00
9	Cycle(R/I)	0.00	29	Hair loss(Y/N)	0.00
10	Cycle length(days)	0.01	30	Pimples(Y/N)	0.00
11	Marraige Status (Yrs)	0.00	31	Fast food (Y/N)	0.00
12	Pregnant(Y/N)	0.61	32	Reg. Exercise(Y/N)	0.19
13	No. of absorptions	0.09	33	BP_Systolic (mmHg)	0.90
14	FSH(mIU/mL)	0.00	34	BP_Diastolic (mmHg)	0.57
15	LH(mIU/mL)	0.00	35	Follicle No. (L)	0.00
16	FSH/LH	0.00	36	Follicle No. (R)	0.00
17	Hip(inch)	0.02	37	Avg. F size (L) (mm)	0.00
18	Waist(inch)	0.02	38	Avg. F size (R) (mm)	0.06
19	Waist: Hip Ratio	0.99	39	Endometrium (mm)	0.07
20	TSH (mIU/L)	0.61			

The count plots are drawn to see the number of instances of both classes in the PCOS dataset. In Figure 4(a), the count plot shows the number of instances of both categories. The no category has 364 instances, and the yes category has 177 instances in the dataset. The dataset is binary class. The 0 indicates No PCOS, and 1 represents Yes PCOS. In Figure 4(b), the pie chart shows the percentage of each class in the dataset. 67.3% of data belong to the PCOS No class, and 32.7% of data belongs to the Yes class.

The 3d scatter plot is to visualize and analyze the most critical feature data point in 3D. It plots data points on three axes to show the relationship between three features. When the value of PRL(ng/mL) is more than 40, and waist Hip Ratio is less than 0.90, PCOS happens, as shown in Figure 5(a). No PCOS occur when the TSH(mmHg) is less than 50 and Bp_Systolic is above 80. Figure 5(b) demonstrates that, When the value of TSH(mmHg) is above 50 and the Bp_Systolic value less than 80, then PCOS happen.

The lmplot is dragged on the dataset's high-value features to represent the PCOS regression described in Figure 6. The lmplot is a two-dimensional plot that combines regplot and FacetGrid. The FacetGrid class helps visualize the

distribution of one variable and the relationship between multiple variables separately within subsets of your dataset using numerous panels. The lmplot is more computationally intensive and is intended as a convenient interface to fit regression models across conditional subsets of a dataset.

In Figure 6(A), a lmplot is drawn between the Hip(inch) and Waist(inch) to visualize the PCOS Regression. As the waist and Hip size increase, the Chance of PCOS increases. In Figure 6(B), the Waist: Hip Ratio and Hb(g/dl) subset is used to analyze the PCOS regression. When the value of Hb(g/dl) is greater than 14 and less than 9, there is more chance of PCOS. Figure 6(C) plots the lmplot between Pregnant(Y/N) and BP_Systolic. This plot shows that if the value of BP_Systolic(mmHg) is 140 and the patient is Pregnant or not, the PCOS does not occur. In Figure 6(D), Blood Group and RR(breath/min) features are taken from the dataset to visualize the Regression plot. When the value of RR(breath/min) is more significant than 25, no PCOS happens. In Figure 6(E), the lmplot is plotted between TSH(mlU/L) and BP_Diastolic(mmHg) feature. When the value of TSH(mlU/L) is between 0 to 20 and BP_Diastolic(mmHg) is 80, there is more chance of PCOS.

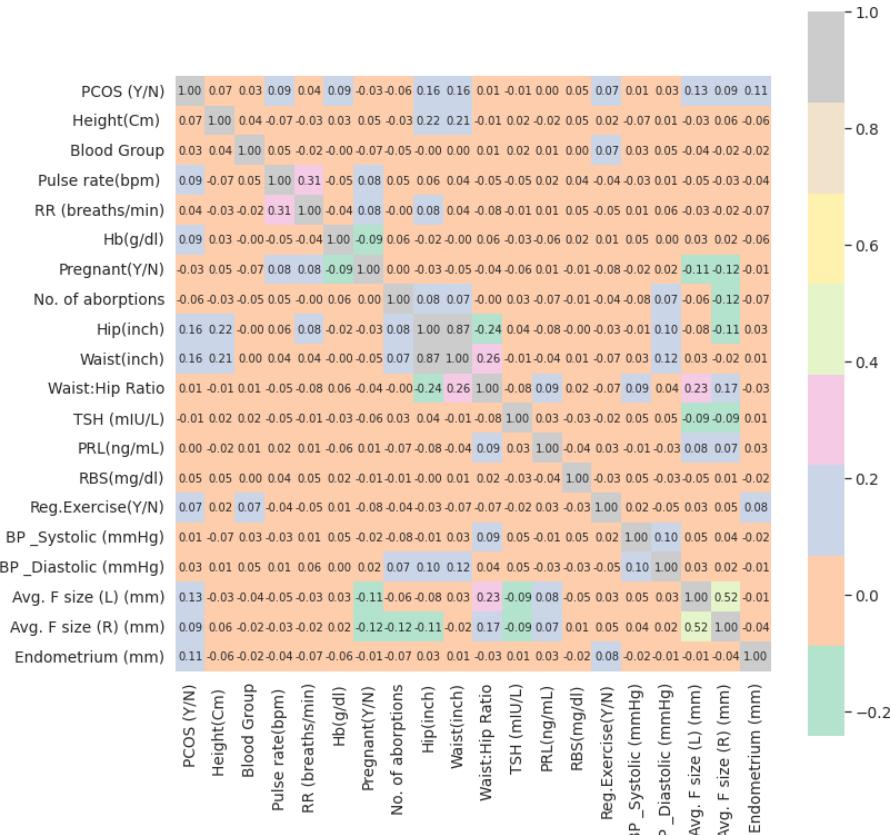
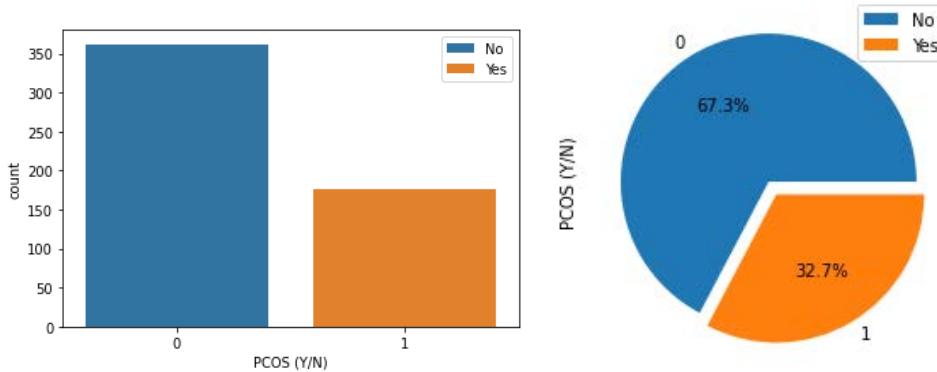


FIGURE 3. The correlation analysis of selected dataset features by the proposed CS-PCOS techniques.



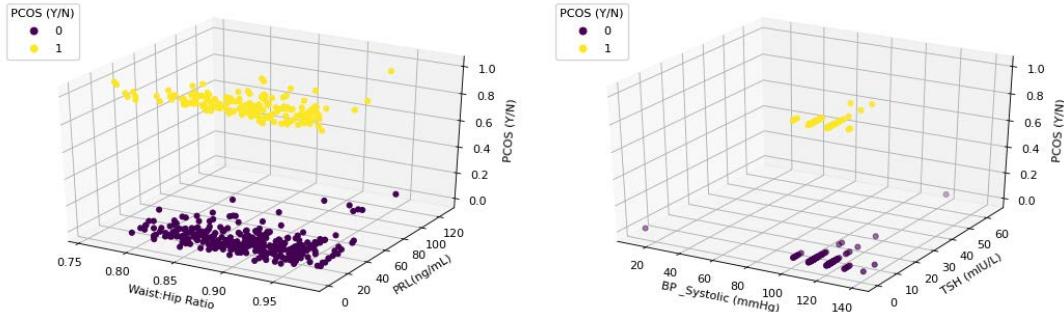
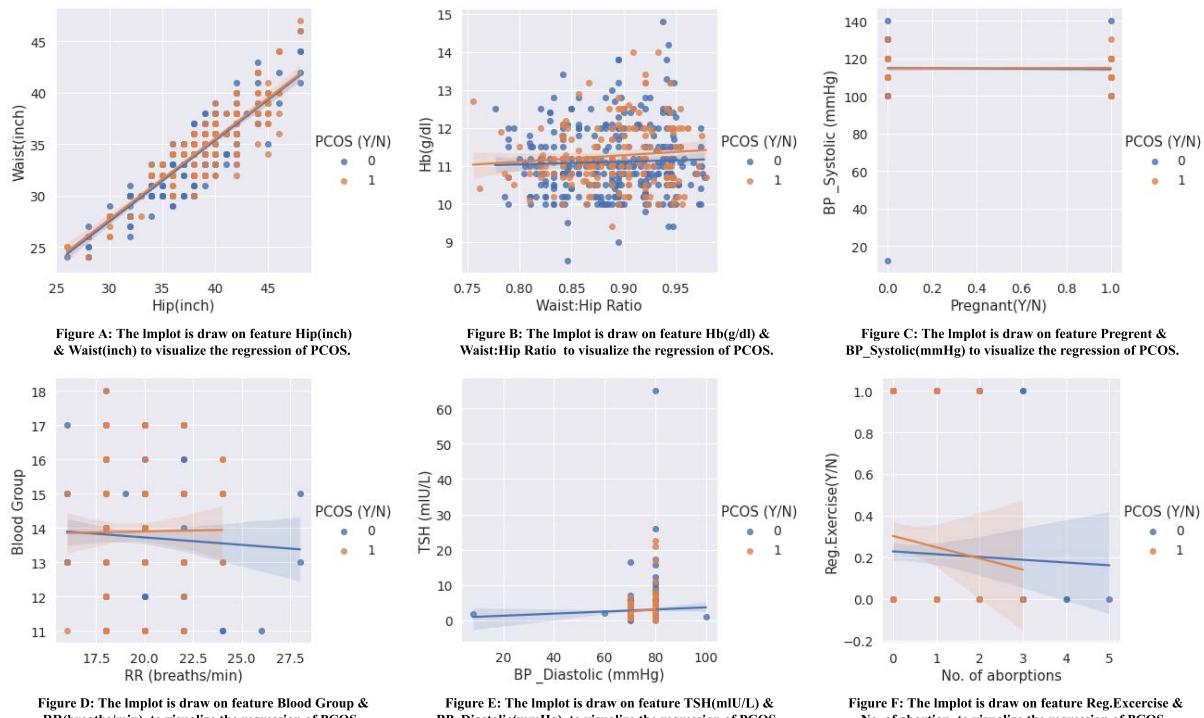
- (a) The count plot shows the number of instances of both classes in the dataset
 (b) The Pie chart shows the distribution of PCOS class in percentage

FIGURE 4. The PCOS patient's data distribution analysis by class.

In Figure 6(F), No of Abortion and Regular Exercise(Y/N) are taken to visualize the lmplot. When the number of abortions is above three and not doing regular exercise, PCOS does not occur.

The histogram is plotted to analyze the frequency distribution of PCOS Yes or No on imported features in Figure 7. Figure 7(A) plots the Hip(inch) to identify the frequency

distribution. The frequency of both classes is highest between 35 and 40. Figure 7(B) plots the histogram on Hb(g/dl). The PCOS yes has the highest count of 60 at HB(g/dl) 11. Furthermore, PCOS class No has a maximum count of 140 before the value of 11. In figure 7(C), the pregnant(Y/N) feature is used to plot the histogram. This graph presents the highest value of both classes at no pregnancy. In Figure 7(D), the

**FIGURE 5.** In the 3D analysis of features distribution analysis by class.**FIGURE 6.** The Implot regression graph analysis of values features with the PCOS class.

BP_Diastolic(mmHg) is taken to plot the histogram. The highest frequency of class 0 is 250 at 80 BP_Diastolic.

In Figure 7(E), the maximum frequency of RR(breath/min) is at a value of 10, which is above 175 for No PCOS and 75 for yes PCOS. In Figure 7(F), the feature TSH(mmU/L) has a frequency between zero and ten. For Yes, PCOS has the highest frequency, 90 at 0 and approximately 340 for No PCOS at a value of 5. In Figure 7(G), BP_systolic (mmHg) feature is taken to analyze the frequency distribution. The BP_systolic (mmHg) the highest frequency is 175 at 100 for No PCOS. In Figure 7(H), PRL (ng/ml) has the highest frequency at 20, gradually decreasing. In Figure 8(I), the frequency of Waist Hip Ratio is from 0.75 to 0.95. The highest frequency for yes No PCOS is 0.95.

D. DATASET SPITING

The data splitting is applied to prevent model overfitting and evaluate the trained model on the unseen test portion of the dataset. The PCOS dataset is split into two portions for the training and testing employing machine learning models. The 80:20 ratio is used for dataset splitting. The 80% portion of the dataset is used for model training, and a 20% portion of the dataset is used for employed model's results evaluations on unseen data. Our research models are trained and evaluated with high accuracy results.

IV. EMPLOYED MACHINE LEARNING TECHNIQUES

The employed machine learning techniques are examined for PCOS prediction in this section. The working mechanism

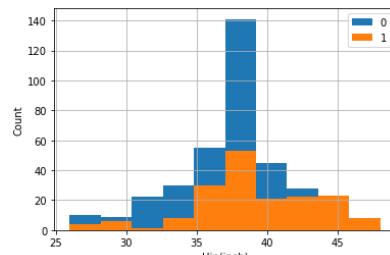


Figure A: The Histogram is plotted on Hip(inch) feature of dataset to examine the cause of PCOS

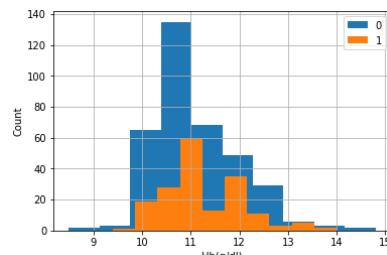


Figure B: The Histogram is plotted on Hb(g/dl) feature of dataset to examine the cause of PCOS

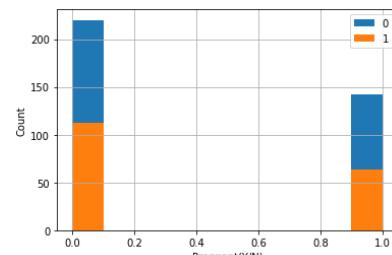


Figure C: The Histogram is plotted on Pregnant(Y/N) feature of dataset to examine the cause of PCOS

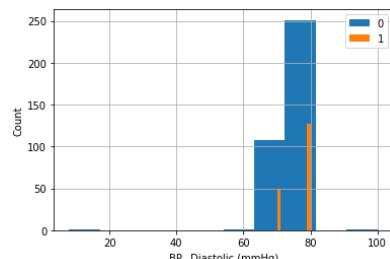


Figure D: The Histogram is plotted on BP_Diastolic(mmHg) feature of dataset to examine the cause of PCOS

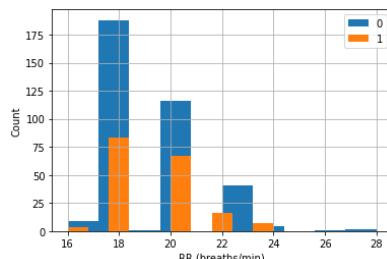


Figure E: The Histogram is plotted on RR(breaths/min) feature of dataset to examine the cause of PCOS

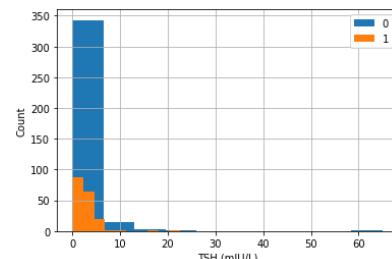


Figure F: The Histogram is plotted on TSH(mIU/L) feature of dataset to examine the cause of PCOS

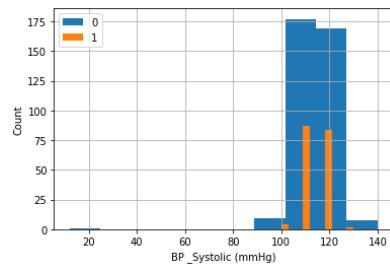


Figure G: The Histogram is plotted on BP_Systolic(mmHg) feature of dataset to examine the cause of PCOS

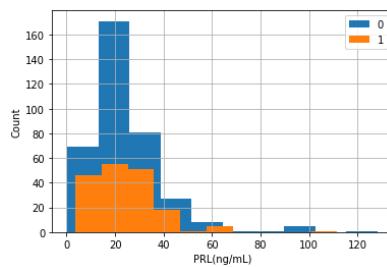


Figure H: The Histogram is plotted on PRL(ng/mL) feature of dataset to examine the cause of PCOS

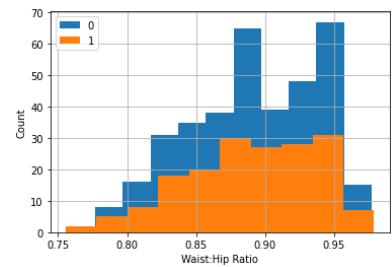


Figure I: The Histogram is plotted on Waist_Hip Ratio feature of dataset to examine the cause of PCOS

FIGURE 7. The histogram analysis analyses the frequency distribution of PCOS Yes or No for selected features.

and mathematical notations for machine learning models are described. The ten predictive machine learning models are under examination for PCOS prediction in our research study.

The stochastic gradient descent (SGD) classifier [33] uses loss functions based on the SGD learning routine for classification. The SGD is used for large-scale learning. The SGD is easy to build and has good efficiency. The SGD efficient optimization model is utilized to minimize a loss function by finding the optimal parameters values of the function. The performance of SGD is based on the loss function. The logistic cost function is expressed in equation 1.

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases} \quad (1)$$

Linear regression (LIR) [34] is the statistical method used for classification that finds the linear relationship between the dependent variable (y) and independent variables (x). A linear relationship analyses how dependent variable values change according to the independent variable values. The LIR model [35] provides a straight line separating the data points. The regression line in the LIR model minimizes the sum of the Square of Residuals, known as the ordinary least

square (OLS). The mathematical notation to express the LIR model is analyzed in equation 2.

$$Y = mX + b \quad (2)$$

The random forest (RF) [36] is a supervised classification model that creates a forest of multiple decision trees. The decision trees are created randomly based on the data samples. Decision nodes represent the features, and tree leaf nodes represent the target output. The majority voting prediction of decision trees is selected as the final prediction. The gini index and entropy are used for data splitting in tree nodes as expressed in equations 3 and 4.

$$Gini\ index = 1 - \sum_{i=1}^n (P_i)^2 \quad (3)$$

$$Entropy(S) = -p_{(+)} \log p_{(+)} - p_{(-)} \log p_{(-)} \quad (4)$$

The bayesian ridge (BR) [37] algorithm uses probability computations for the classification task. The BR model is suitable for real-world problems where the data is insufficient and poorly distributed. The BR model formulates a linear regression model by using the probability distributors.

TABLE 3. The best-fit hyperparameters analysis of all employed machine learning models.

Technique	Hyperparameters
SGD	loss='hinge', penalty='l2', alpha=0.0001, l1_ratio=0.15, max_iter=1000, tol=1e-3, learning_rate='optimal'.
LIR	copy_X=True, fit_intercept=True, positive=False, normalize=False.
RF	max_depth=20, random_state=0, n_estimators=100, criterion='gini', max_features='sqrt', bootstrap=True.
BR	tol=1e-3, n_iter=300, alpha_1=1e-6, lambda_1=1e-6, alpha_2=1e-6, lambda_2=1e-6.
SVM	kernel='linear', C=1.0, degree=3, gamma='scale', tol=1e-3, cache_size=200, decision_function_shape='ovr'.
KNC	n_neighbors=5, weights='uniform', algorithm='auto', metric='minkowski', leaf_size=30, p=2.
MLP	hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, learning_rate='constant'.
LOR	penalty='l2', tol=1e-4, C=1.0, solver='lbfgs'.
GNB	var_smoothing=1e-9.
GBC	loss='log_loss', max_depth=3, learning_rate=0.1, criterion='friedman_mse', n_estimators=100.

The BR model predicts the target (y) by calculating it from a probability distribution instead estimating a single feature value. The mathematical notation to find the y target using the BR model is expressed in equation 5.

$$p(y | X, w, a) = N(y | X_w, a) \quad (5)$$

The support vector machine (SVM) [38] is a supervised machine learning model. The SVM is mainly used for classification and regression problems. The primary aim of the SVM model is to best decision boundary that separates the data points into their relevant category in n -dimensional feature space. The best decision boundary in SVM is known as the hyperplane. The SVM model selects the extreme vectors to create the hyperplane. The vectors are called support vectors. The hyperplane in SVM is used to predict as expressed in equation 6.

$$h(x_i) = \begin{cases} +1 & \text{if } w.x + b \geq 0 \\ -1 & \text{if } w.x + b < 0 \end{cases} \quad (6)$$

The k-neighbors classifier (KNC) [39] is the simplest and non-parametric algorithm machine learning model for classification problems. The KNC model calculates the similarity between the data points and places the new input points into a similar category that is similar to each other. The KNC model saves available data and predicts the new data to its suite category based on similarity. The KNC is a lazy learner model because it performs a prediction at the time of classification. It does not learn immediately from the training data. The time computations are high and have low efficiency. The euclidian distance between data points is found as expressed in equation 7.

$$E(A_1, A_2) = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \quad (7)$$

The MLP classifier (MLP) [40] is a feedforward artificial neural network-based supervised machine learning model. The MLP model [41] is based on types of network layers. The types of layers are input, output and hidden layer. The input layer in the network handles the input data points, and the output layer is responsible for the prediction task. The hidden layer processes the data within the neural network. The MLP uses the back-propagation technique based on data passing in the forward direction. The neurons present in the MLP

network are trained with the backpropagation technique. The neurons use nonlinear activation functions between the output and input layers. The weighted sum of the input features in MLP is calculated as analyzed in equation 8.

$$u(x) = \sum_{i=1}^n w_i x_i \quad (8)$$

The logistic regression (LOR) [42] is a supervised machine learning model for binary classification. The LOR model [43] forecasts the categorical dependent variable using training data of independent variables. The target class must be in the form of a discrete value. The LOR model gives the probabilistic values. The output values lie between 1 and 0. The LOR is similar to the LIR model, with only a difference in their use. The logistic function of s-shaped formed in the LOR model, which forecasts the values 1 or 0. The logistic function is analyzed in equation 9.

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n \quad (9)$$

The gaussian NB (GNB) [44] is a supervised machine learning model. The GNB model is based on the naive bayes methods and theorem. The GNB technique [45] has the powerful assumption that all the predictors are independent of each other. One feature in a class is independent of another feature in the same class. The GNB utilized gaussian distribution and naive assumptions to predict the target class. The target feature prediction by the GNB model is expressed in equation 10.

$$P(Y|features) = \frac{P(Y)P(features|Y)}{P(features)} \quad (10)$$

The gradient boosting classifier (GBC) [46] is an ensemble learning-based Boosting model mainly used for classification and regression tasks. The GBC [47] models work incrementally. The principle of GBC is to build models sequentially by training each base model. The motive is to make a robust model. The several models combine to make a weak learner a robust model. Several gradient-boosted trees are involved in making a GBC. The final powerful model has the correction prediction values. The three main components in GBC as loss function, Weak learner and additive model. for classification, the GBC model prediction for the target class is expressed

TABLE 4. The comparative performance evaluation of employed machine learning models for unseen test data without using the proposed technique.

Technique	Training time(second)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SGD	0.006	79	82	79	79
LIR	0.034	84	85	84	85
RF	0.193	89	89	89	89
BR	0.014	84	85	84	84
SVM	0.666	88	88	88	88
KNC	0.002	70	68	70	68
MLP	0.472	83	83	83	83
LOR	0.042	80	80	80	80
GNB	0.003	81	81	81	80
GBC	0.259	89	89	89	89

in equation 11.

$$y = \frac{\sum \text{Residual}}{\sum [\text{Prev probability} * (1 - \text{Prev probability})]} \quad (11)$$

A. HYPERPARAMETER TUNING

The iterative training and testing process selects the best-fit hyperparameters [48] for all applied machine learning techniques. The hyper-parameters are selected as final, and a machine learning model gives accurate prediction results. The hyperparameter tuning [49] of our research models is analyzed in Table 3. The analysis demonstrates the parameters utilized to achieve the high-performance metrics score. The hyper-parameters proved very beneficial for our employed machine learning models in this research study.

V. RESULTS AND DISCUSSIONS

The results and scientific evaluations of our proposed research study are examined in this section. The Python programming tool and scikit-learn library module are used for building the employed machine learning model. The performance metrics used are the accuracy score, precision score, recall score, and f1-score. The performance metrics are evaluated for scientific validation of our research models. The followings are the essential components of evaluation metrics:

- The predicted values and actual values are positively known as true positive (TP).
- The predicted values and actual values are negative, known as true negative (TN).
- The actual value is negative, and the predicted value is positive, refers as false positive (FP).
- The actual value is positive, and the predicted value is negative, which refers to false negative (FN).

The employed model's accuracy score shows how much the model is good in prediction. The accuracy is also related to the error rate of a model. Higher the accuracy, lower the error rate. The accuracy is determined by dividing the correct number of predictions by the total number. The accuracy score of our proposed model is 100%. Mathematically, the accuracy score is demonstrated as:

$$\text{Accuracy score} = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

The precision score of a learning model is also known as positive predictive value. The precision is measured by the positively predicted label proportion that is positive. The precision, in general, calculates the employed model accuracy in predicting a data sample as positive. The precision score of our proposed model is 100%. The mathematical notations to express precision scores are as follows:

$$\text{Precision score} = \frac{TP}{TP + FP} \quad (13)$$

The recall score of employed models is the measure of how many of the TP were recalled (found) correctly. The recall is also called the sensitivity of a learning model. The recall score of our proposed model is 100%. The mathematical notations to define the recall are as follows:

$$\text{Recall score} = \frac{TP}{TP + FN} \quad (14)$$

The f1 score is the statistical measure that sums up a predictive model's performance by combining the precision and recall values. The f1 measure is the harmonic mean between the recall and precision. The f1 score of our proposed model is 100%. The mathematical equation to calculate the f1 score is expressed as:

$$F1 - score = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (15)$$

The comparative performance metrics analysis of applied learning models is conducted in Table 4. The time complexity computations and performance metrics results are calculated without using our proposed approach. The analysis demonstrated that all applied learning models achieved average scores in predicting PCOS. From the analysis and Figure 8, the highest accuracy, precision, recall, and f1 score is 89%, achieved by RF and GBC techniques. The minimum accuracy score is 70%, the precision score is 68%, the recall score is 70%, and the f1 score is 68% achieved by the KNC technique. The time complexity analysis describes that KNC have less training time of 0.002. However, also have low-performance metrics scores.

The performance metrics comparative analysis of applied learning models is conducted in Table 5. The performance metrics results and time complexity computations are calculated using our proposed approach. The analysis demonstrated that all applied learning models achieved the highest

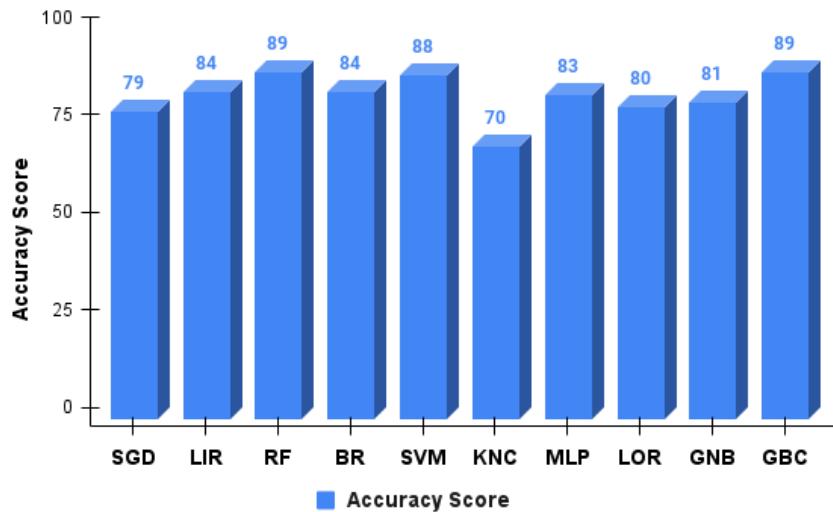


FIGURE 8. The accuracy scores comparative evaluation of employed machine learning models for unseen test data without using the proposed technique.

TABLE 5. Using the proposed technique, the comparative performance evaluation of the employed machine learning model for unseen test data.

Technique	Training time(second)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SGD	0.004	69	68	69	68
LIR	0.024	100	100	100	100
RF	0.147	100	100	100	100
BR	0.004	100	100	100	100
SVM	0.842	100	100	100	100
KNC	0.002	56	53	56	54
MLP	0.592	99	99	99	99
LOR	0.025	100	100	100	100
GNB	0.002	100	100	100	100
GBC	0.071	100	100	100	100

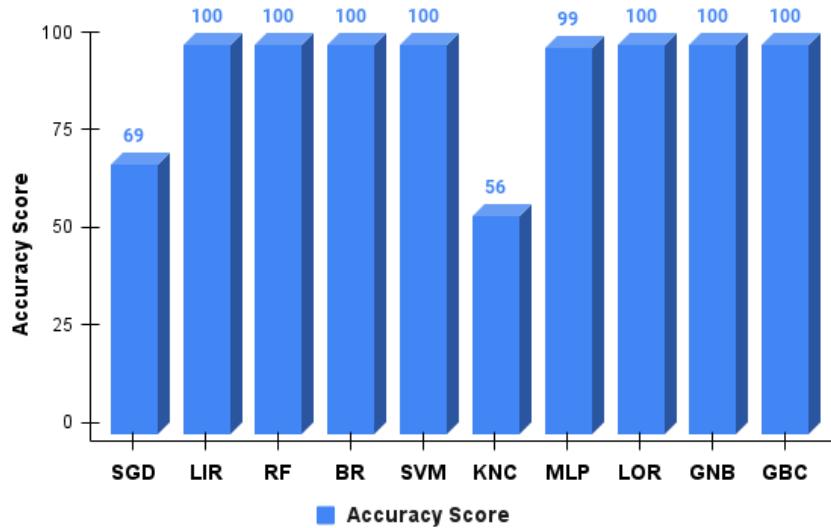


FIGURE 9. Using the proposed technique, the accuracy scores comparative evaluation of employed machine learning models for unseen test data.

performance metrics scores in predicting the PCOS. From the analysis and Figure 9, the highest accuracy, precision, recall, and f1 score is 100%, achieved by LIR, RF, BR, SVM, LOR,

GNB, and GBC techniques. The minimum accuracy score is 56%, the precision score is 53%, the recall score is 56%, and the f1 score is 54%, achieved by the KNC technique. The time

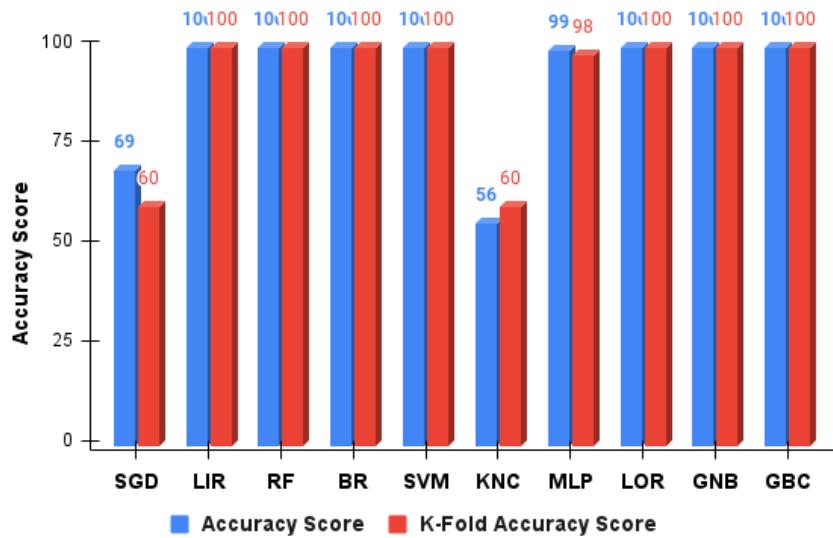


FIGURE 10. The accuracy scores comparative analysis of the K-Fold technique to validate the overfitting of the employed learning techniques.

TABLE 6. The classification report analysis of employed learning models by using the proposed technique.

Target Category	Precision	Recall	F1-score	Support Score
SGD				
0	0.72	0.83	0.77	70
1	0.59	0.44	0.50	39
LIR				
0	1.00	1.00	1.00	70
1	1.00	1.00	1.00	39
RF				
0	1.00	1.00	1.00	70
1	1.00	1.00	1.00	39
SVM				
0	1.00	1.00	1.00	70
1	1.00	1.00	1.00	39
BR				
0	1.00	1.00	1.00	70
1	1.00	1.00	1.00	39
KNC				
0	0.63	0.74	0.68	70
1	0.33	0.23	0.27	39
MLP				
0	1.00	0.99	0.99	70
1	0.97	1.00	0.99	39
LOR				
0	1.00	1.00	1.00	70
1	1.00	1.00	1.00	39
GNB				
0	1.00	1.00	1.00	70
1	1.00	1.00	1.00	39
GBC				
0	1.00	1.00	1.00	70
1	1.00	1.00	1.00	39

complexity analysis describes that GNB has less training time of 0.002. However, the GNB have high-performance metrics scores. The GNB is our proposed model for predicting the PCOS.

The classification report analysis by individual target class for each employed learning model is examined in Table 6.

The classification report values are calculated for the models using the proposed approach. The analysis demonstrates that the KNC and SDG have low accuracy scores in class-wise metrics evaluations. The outperformed GNB model has achieved 100% scores in classification report analysis.

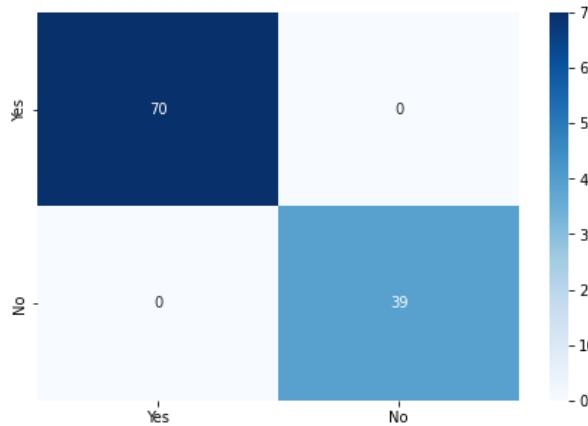
TABLE 7. The K-Fold cross-validation analysis to validate the overfitting of the employed learning techniques.

Sr no	K-Fold	Technique	Accuracy Score (%)
1	10	SGD	60
2	10	LIR	100
3	10	RF	100
4	10	BR	100
5	10	SVM	100
6	10	KNC	60
7	10	MLP	98
8	10	LOR	100
9	10	GNB	100
10	10	GBC	100

To validate the overfitting of employed machine learning models, we have applied the k-fold cross-validation technique as analyzed in Table 7. The 10 folds of the dataset are used for validation. The analysis demonstrates that techniques achieved 100% scores using our proposed approach and 100% accuracy using the k-fold techniques. Figure 10 shows the accuracy of comparative analysis of employed models by using the k-fold validation. The visualized analysis demonstrates that the MLP model achieved 99%, and by using k-fold, 98% accuracy was achieved. The SGD and KNC models achieve the lowest accuracy scores in this analysis. In conclusion, all employed models are validated using k-fold technique. The k-fold analysis demonstrates that our employed machine learning models are not overfitted. Models are in general form and accurate results on unseen test data.

TABLE 8. The performance validation comparative analysis with the past applied state-of-the-art approaches.

Literature	Year	Learning Type	Proposed Technique	Accuracy (%)	Recall (%)	Precision (%)
[22]	2020	Machine Learning	RFLR	91	90	89
Proposed	2022	Machine Learning	CS-PCOS + GNB	100	100	100

**FIGURE 11.** The confusion matrix validation analysis of our proposed model.

The comparative analysis of past applied state-of-the-art studies is examined in Table 8. The comparison parameters are the year, learning type, proposed technique, accuracy score, recall score, and precision score. The analysis demonstrates that using our novel proposed CS-PCOS technique, the outperformed GNB model achieved the highest scores compared with the past proposed techniques. Our proposed model outperformed the state of art studies.

The confusion matrix analysis is conducted to validate our performance metrics scorers as analyzed in Figure 11. The analyzed confusion matrix is for outperformed GNB model. The analysis demonstrates that 70 samples are found as TP, and 39 samples are found as TN. The 0 samples are found for FN and FP in this analysis. The confusion matrix validates our proposed model for achieving the 100% accuracy score in predicting the PCOS.

VI. CONCLUSION

The prediction of PCOS disease using data of 541 patients through machine learning is proposed in this research study. A novel CS-PCOS feature selection technique is proposed. The ten machine learning techniques are SGD, LIR, RF, BR, SVM, KNC, MLP, LOR, GNB, and GBC applied in comparison. The proposed GNB outperformed with a 100% accuracy score and time computation of 0.002 by using the proposed CS-PCOS feature selection techniques. The state of art studies comparison shows that the proposed model outperformed. The proposed model's overfitting is validated using a ten-fold cross-validation technique. Our research study concludes that the dataset features prolactin (PRL), blood pressure systolic, blood pressure diastolic, thyroid stimulating hormone (TSH), relative risk (RR-breaths), and pregnancy are the

most prominent factors having high involvement in PCOS prediction. The study limitations and in future work, we will enhance the dataset by collecting more data on PCOS-related patients and applying data balancing techniques. Also, the deep learning-based will be applied for PCOS prediction.

REFERENCES

- I. Kyrou, E. Karteris, T. Robbins, K. Chatha, F. Drenos, and H. S. Randeva, "Polycystic ovary syndrome (PCOS) and COVID-19: An overlooked female patient population at potentially higher risk during the COVID-19 pandemic," *BMC Med.*, vol. 18, no. 1, pp. 1–10, Jul. 2020.
- B. J. Sherman, N. L. Baker, K. T. Brady, J. E. Joseph, L. M. Nunn, and A. McRae-Clark, "The effect of oxytocin, gender, and ovarian hormones on stress reactivity in individuals with cocaine use disorder," *Psychopharmacology*, vol. 237, no. 7, pp. 2031–2042, May 2020.
- X.-Z. Zhang, Y.-L. Pang, X. Wang, and Y.-H. Li, "Computational characterization and identification of human polycystic ovary syndrome genes," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 12949.
- E. Khashchenko, E. Uvarova, M. Vysokikh, T. Ivanets, L. Krechetova, N. Tarasova, I. Sukhanova, F. Mamedova, P. Borovikov, I. Balashov, and G. Sukhikh, "The relevant hormonal levels and diagnostic features of polycystic ovary syndrome in adolescents," *J. Clin. Med.*, vol. 9, no. 6, p. 1831, Jun. 2020.
- M. Woźniak, R. Krajewski, S. Makuch, and S. Agrawal, "Phytochemicals in gynecological cancer prevention," *Int. J. Mol. Sci.*, vol. 22, no. 3, p. 1219, Jan. 2021.
- D. Dewailly, M. E. Lujan, E. Carmina, M. I. Cedars, J. Laven, R. J. Norman, and H. F. Escobar-Morreale, "Definition and significance of polycystic ovarian morphology: A task force report from the androgen excess and polycystic ovary syndrome society," *Hum. Reproduction Update*, vol. 20, no. 3, pp. 334–352, 2014.
- A. S. Praty and T. T. Shitu, "An efficient decision tree establishment and performance analysis with different machine learning approaches on polycystic ovary syndrome," in *Proc. ICCIT 23rd Int. Conf. Comput. Inf. Technol.*, Dec. 2020, pp. 1–5.
- E. C. Costa, J. C. F. D. Sá, N. K. Stepto, I. B. B. Costa, L. F. Farias-Junior, S. D. N. T. Moreira, E. M. M. Soares, T. M. A. M. Lemos, R. A. V. Browne, and G. D. Azevedo, "Aerobic training improves quality of life in women with polycystic ovary syndrome," *Med. Sci. Sports Exerc.*, vol. 50, pp. 1357–1366, Jul. 2018.
- M. A. Karimzadeh and M. Javedani, "An assessment of lifestyle modification versus medical treatment with clomiphene citrate, metformin, and clomiphene citrate-metformin in patients with polycystic ovary syndrome," *Fertility Sterility*, vol. 94, pp. 216–220, Jun. 2010.
- I. Almenning, A. Rieber-Mohn, K. M. Lundgren, T. S. Løvvik, K. K. Garnaes, and T. Moholdt, "Effects of high intensity interval training and strength training on metabolic, cardiovascular and hormonal outcomes in women with polycystic ovary syndrome: A pilot study," *PLoS ONE*, vol. 10, Sep. 2015, Art. no. e0138793.
- D. Chizen, S. Serrao, J. Rooke, L. McBrearty, R. Pierson, P. Chilibeck, and G. Zello, "The 'pulse' diet & PCOS," *Fertility Sterility*, vol. 102, p. e267, Sep. 2014.
- H. H. Mehrabani, S. Salehpour, B. J. Meyer, and F. Tahbaz, "Beneficial effects of a high-protein, low-glycemic-load hypocaloric diet in overweight and obese women with polycystic ovary syndrome: A randomized controlled intervention study," *J. Amer. College Nutrition*, vol. 31, pp. 117–125, Apr. 2012.
- F. Giallauria, S. Palomba, L. Maresca, L. Vuolo, D. Tafuri, G. Lombardi, A. Colao, C. Vigorito, and F. Orio, "Exercise training improves autonomic function and inflammatory pattern in women with polycystic ovary syndrome (PCOS)," *Clin. Endocrinol.*, vol. 69, pp. 792–798, Nov. 2008.
- F. Saleem and S. W. Rizvi, "New therapeutic approaches in obesity and metabolic syndrome associated with polycystic ovary syndrome," *Cureus*, vol. 9, pp. 1–12, Nov. 2017.

- [15] G. Ladson, W. C. Dodson, S. D. Sweet, A. E. Archibong, A. R. Kunselman, L. M. Demers, N. I. Williams, P. Coney, and R. S. Legro, "The effects of metformin with lifestyle therapy in polycystic ovary syndrome: A randomized double-blind study," *Fertility Sterility*, vol. 95, pp. 1059–1066, Mar. 2011.
- [16] A. Gambineri, L. Patton, A. Vaccina, M. Cacciari, A. M. Morselli-Labate, C. Cavazza, U. Pagotto, and R. Pasquali, "Treatment with flutamide, metformin, and their combination added to a hypocaloric diet in overweight-obese women with polycystic ovary syndrome: A randomized, 12-month, placebo-controlled study," *J. Clin. Endocrinol. Metabolism*, vol. 91, no. 10, pp. 3970–3980, 2006.
- [17] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 156–180, 2021.
- [18] A. Garg and V. Mago, "Role of machine learning in medical research: A survey," *Comput. Sci. Rev.*, vol. 40, May 2021, Art. no. 100370.
- [19] D. Hu, W. Dong, X. Lu, H. Duan, K. He, and Z. Huang, "Evidential mace prediction of acute coronary syndrome using electronic health records," *BMC Med. Informat. Decis. Making*, vol. 19, no. 2, pp. 9–17, 2019.
- [20] M. M. Hassan and T. Mirza, "Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome," *Int. J. Comput. Appl.*, vol. 175, pp. 42–53, Sep. 2020.
- [21] G. Du, L. Ma, J.-S. Hu, J. Zhang, Y. Xiang, D. Shao, and H. Wang, "Prediction of 30-day readmission: An improved gradient boosting decision tree approach," *J. Med. Imag. Health Informat.*, vol. 9, no. 3, pp. 620–627, 2019.
- [22] S. Bharati, P. Podder, and M. R. H. Mondal, "Diagnosis of polycystic ovary syndrome using machine learning algorithms," in *Proc. IEEE Region Symp. (TENSYMP)*, Jun. 2020, pp. 1486–1489.
- [23] S. A. Bhat, "Detection of polycystic ovary syndrome using machine learning algorithms," M.S. thesis, Dublin, Nat. College Ireland, Dublin, Ireland, 2021.
- [24] S. Yang, X. Zhu, L. Zhang, L. Wang, and X. Wang, "Classification and prediction of Tibetan medical syndrome based on the improved bp neural network," *IEEE Access*, vol. 8, pp. 31114–31125, 2020.
- [25] D. Dewailly, M. E. Lujan, E. Carmina, M. I. Cedars, J. Laven, R. J. Norman, and H. F. Escobar-Morreale, "Definition and significance of polycystic ovarian morphology: A task force report from the androgen excess and polycystic ovary syndrome society," *Hum. Reproduction Update*, vol. 20, no. 3, pp. 334–352, 2014.
- [26] A. Saravanan and S. Sathiamoorthy, "Detection of polycystic ovarian syndrome: A literature survey," *Asian J. Eng. Appl. Technol.*, vol. 7, pp. 46–51, Nov. 2018.
- [27] V. Thakre, "PCOCare: PCOS detection and prediction using machine learning algorithms," *Biosci. Biotechnol. Res. Commun.*, vol. 13, no. 14, pp. 240–244, Dec. 2020.
- [28] R. M. Aziz, "Nature-inspired metaheuristics model for gene selection and classification of biomedical microarray data," *Med. Biol. Eng. Comput.*, vol. 60, no. 6, pp. 1627–1646, 2022.
- [29] R. M. Aziz, "Application of nature inspired soft computing techniques for gene selection: A novel frame work for classification of cancer," *Soft Comput.*, vol. 1, pp. 1–18, Apr. 2022.
- [30] 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," *J. Ovarian Res.*, vol. 15, no. 1, pp. 1–13, 2022.
- [31] S. Dhar, S. Mridha, and P. Bhattacharjee, "Mutational landscape screening through comprehensive in silico analysis for polycystic ovarian syndrome-related genes," *Reproductive Sci.*, vol. 29, no. 2, pp. 480–496, 2022.
- [32] P. Kottarathil, "Polycystic ovary syndrome (PCOS) | Kaggle," Hospital, Kerala, India, Tech. Rep. 9.71, 2022.
- [33] J. Huang, S. Ling, X. Wu, and R. Deng, "GIS-based comparative study of the Bayesian network, decision table, radial basis function network and stochastic gradient descent for the spatial prediction of landslide susceptibility," *Land*, vol. 11, p. 436, Mar. 2022.
- [34] S. Ghosal, S. Sengupta, M. Majumder, and B. Sinha, "Linear regression analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases–March 14th 2020)," *Diabetes Metabolic Syndrome, Clin. Res. Rev.*, vol. 14, pp. 311–315, Jul. 2020.
- [35] H. Lee, J. Wang, and B. Leblon, "Using linear regression, random forests, and support vector machine with unmanned aerial vehicle multispectral images to predict canopy nitrogen weight in corn," *Remote Sens.*, vol. 12, p. 2071, Jun. 2020.
- [36] M. A. Khan, S. A. Memon, F. Farooq, M. F. Javed, F. Aslam, and R. Alyousef, "Compressive strength of fly-ash-based geopolymer concrete by gene expression programming and random forest," *Adv. Civil Eng.*, vol. 2021, Jan. 2021, Art. no. 6618407.
- [37] M. H. Na, W. H. Cho, S. K. Kim, and I. S. Na, "Automatic weight prediction system for Korean cattle using Bayesian ridge algorithm on RGB-D image," *Electronics*, vol. 11, p. 1663, May 2022.
- [38] S. Shabani, S. Samadianfar, M. T. Sattari, A. Mosavi, S. Shamshirband, T. Kmet, and A. R. Várkonyi-Kóczy, "Modeling pan evaporation using Gaussian process regression K-nearest neighbors random forest and support vector machines; comparative analysis," *Atmosphere*, vol. 11, p. 66, Jan. 2020.
- [39] L. V. R. Kumari and Y. P. Sai, "Classification of arrhythmia beats using optimized K-Nearest neighbor classifier," in *Intelligent Systems (Lecture Notes in Networks and Systems)*, vol. 185. Midtown Manhattan, NY, USA: Springer, 2021, pp. 349–359.
- [40] R. Pahuja and A. Kumar, "Sound-spectrogram based automatic bird species recognition using MLP classifier," *Appl. Acoust.*, vol. 180, Sep. 2021, Art. no. 108077.
- [41] U. Azmat, Y. Y. Ghadi, T. Al Shloul, S. A. Alsuhibany, A. Jalal, and J. Park, "Smartphone sensor-based human locomotion surveillance system using multilayer perceptron," *Appl. Sci.*, vol. 12, p. 2550, Feb. 2022.
- [42] A. M. Almeshal, A. I. Almazrouee, M. R. Alenizi, and S. N. Alhajeri, "Forecasting the spread of COVID-19 in Kuwait using compartmental and logistic regression models," *Appl. Sci.*, vol. 10, p. 3402, May 2020.
- [43] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A comparative analysis of logistic regression, random forest and KNN models for the text classification," *Augmented Hum. Res.*, vol. 5, pp. 1–16, Mar. 2020.
- [44] D. T. Barus, R. Elfarizy, F. Masri, and P. H. Gunawan, "Parallel programming of churn prediction using Gaussian Naïve Bayes," in *Proc. 8th Int. Conf. Inf. Commun. Technol. (ICoICT)*, Jun. 2020, pp. 1–4.
- [45] L. Cataldi, L. Tiberi, and G. Costa, "Estimation of MCS intensity for Italy from high quality accelerometric data, using GMICEs and Gaussian Naïve Bayes classifiers," *Bull. Earthq. Eng.*, vol. 19, pp. 2325–2342, Apr. 2021.
- [46] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM)," *Diagnostics*, vol. 11, p. 1714, Sep. 2021.
- [47] C. Bowd, A. Belghith, J. A. Proudfoot, L. M. Zangwill, M. Christopher, M. H. Goldbaum, H. Hou, R. C. Penteado, S. Moghimi, and R. N. Weinreb, "Gradient-boosting classifiers combining vessel density and tissue thickness measurements for classifying early to moderate glaucoma," *Amer. J. Ophthalmol.*, vol. 217, pp. 131–139, Sep. 2020.
- [48] J. Isabona, A. L. Imoize, and Y. Kim, "Machine learning-based boosted regression ensemble combined with hyperparameter tuning for optimal adaptive learning," *Sensors*, vol. 22, p. 3776, May 2022.
- [49] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis," *Informatics*, vol. 8, p. 79, Nov. 2021.



SHAZIA NASIM received the master's degree in computer science from Bahauddin Zakariya University, Multan, in 2012. She is currently pursuing the M.S. degree in computer science with the Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan. Her current research interests include data mining and machine learning.



MUBARAK ALMUTAIRI received the B.Sc. degree in systems engineering from the King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, in 1997, the M.Sc. degree in industrial and systems engineering from the University of Florida, Gainesville, FL, USA, in 2003, and the Ph.D. degree in systems design engineering from the University of Waterloo, Waterloo, Canada, in 2007. From 1997 to 2000, he was an Industrial Engineer with Saudi Arabia Oil Company (Aramco). He is currently an Associate Professor with the Computer Science and Engineering Department, University of Hafr Albatin (UHB), Hafr Albatin, Saudi Arabia, where he is also the Dean of the Applied College. His research interests include decision analysis, expert systems, risk assessment, information security, fuzzy logic, and mobile government application.



KASHIF MUNIR received the B.Sc. degree in mathematics and physics from Islamia University Bahawalpur, Pakistan, in 1999, the M.Sc. degree in information technology from University Sains Malaysia, in 2001, the M.S. degree in software engineering from the University of Malaya, Malaysia, in 2005, and the Ph.D. degree in informatics from the Malaysia University of Science and Technology, Malaysia, in 2015. He has been in the field of higher education, since 2002. After an initial teaching experience in courses with the Binary College, Malaysia, for one semester, and at Stamford College, Malaysia, for around four years, he later relocated to Saudi Arabia. He worked with the King Fahd University of Petroleum and Minerals, Saudi Arabia, from September 2006 to December 2014. He moved to the University of Hafr Al-Batin, Saudi Arabia, in January 2015. In July 2021, he joined the Khwaja Farid University of Engineering & IT, Rahim Yar Khan, as an Assistant Professor with the IT Department. He has published journal articles, conference papers, book, and book chapters. He has been in the technical program committee of many peer-reviewed conferences and journals, where he has reviewed many research papers. His research interests include the areas of cloud computing security, software engineering, and project management.



ALI RAZA received the Bachelor of Science degree in computer science from the Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2021, where he is currently pursuing the M.S. degree in computer science. His current research interests include data science, artificial intelligence, data mining, natural language processing, machine learning, deep learning, and image processing.



FAIZAN YOUNAS was born in Pakistan, in 1999. He received the Bachelor of Science degree in computer science from the Khwaja Fareed University of Engineering & Information Technology (KFUEIT), Rahim Yar Khan, Pakistan, in 2021, where he is currently pursuing the M.S. degree in computer science. His research interests include natural language processing (NLP), machine learning, and deep learning.

• • •