Informatics in Medicine Unlocked 47 (2024) 101500

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Empowering early detection: A web-based machine learning approach for PCOS prediction

Md Mahbubur Rahman [a,*], Ashikul Islam [b], Forhadul Islam [b], Mashruba Zaman [b], Md Rafiul Islam [b], Md Shahriar Alam Sakib [b], Hafiz Md Hasan Babu [c]

[a] Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh
[b] Department of Computer Science and Engineering, Dhaka International University, Dhaka, Bangladesh
[c] Department of Computer Science and Engineering, University of Dhaka, Dhaka, Bangladesh

ARTICLE INFO

Keywords:
Machine learning
Mutual information
Django
Polycystic ovary syndrome

ABSTRACT

Nowadays, Polycystic Ovary Syndrome (PCOS) affects many women, making it a prevalent concern. It is a hormonal disorder that causes irregular, delayed, or absent menstrual cycles in the female body. This condition can lead to the development of type 2 diabetes, gestational diabetes, weight gain, unwanted body hair, and various other complications. In severe cases, PCOS can result in infertility, posing a challenge for patients trying to conceive. Statistics show that the incidence rate of PCOS has significantly increased in recent years, which is alarming. If PCOS is identified early, people may follow their doctor's recommendations and live a better life. The dataset used for this research contains records for 541 patients. The aim of this study is to employ machine learning models to identify patterns in this disorder. The information learned is then inputted into various algorithms to assess accuracy, specificity, sensitivity, and precision using different ML models, such as Logistic Regression (LR), Decision Tree (DT), AdaBoost (AB), Random Forest (RF), and Support Vector Machine (SVM) among others. The research utilized the Mutual Information model for feature selection and compared the models to determine the most accurate one. Employing the Mutual Information model for feature engineering, AB and RF achieved the highest accuracy of 94 %.

1. Introduction

Polycystic ovary syndrome commonly referred to as PCOS. PCOS is the most prevalent endocrine condition and involves imbalances sex hormones in the female body. The duration of treatment for PCOS is determined by the patient's symptoms. The cysts that form on the ovaries of some affected persons are the source of the syndrome's name, however they are neither a frequent symptom nor the primary cause of the disease. Surgery, medication, and changes in lifestyle are some potential treatments for PCOS patients. PCOS is a widespread hormonal disorder that affects women of reproductive age who are categorized as assigned female at birth (AFAB) [1]. It can result in ovulation cysts, hirsutism, acne, irregular menstruation, and high testosterone levels. While the exact cause of PCOS remains unknown, potential causes include inflammation, genetics, or insulin resistance. PCOS can increase the risk of diabetes, hypertension, depression, and other health problems [2,3]. The female reproductive system consists of the ovaries, fallopian tubes, uterus, and vagina [5]. The ovaries contain a lifelong supply of eggs that are kept in small, fluid-filled sacs called follicles. Hormones governing the ovaries' activity are produced by the pituitary gland situated at the brain's base. Each month, the gland releases follicle stimulating hormone (FSH) and luteinizing hormone (LH) into the bloodstream, which prompts the ovaries to maturate eggs. As eggs mature, follicles release estrogen. This leads to ovulation when LH levels surge. Unfertilized eggs travel through the fallopian tube. Remaining follicles dissolve, but in case of PCOS, this process can disrupt the menstrual cycle and potentially causing infertility [4–6].

In 1935, Stein and Leventhal discovered that PCOS affects 5–10 % of women aged 12–45 [7]. It impacts over 5 million women globally, with 70 % often remaining undiagnosed [8]. Ethnicity plays a role, with diagnosis rates of 4.8 % among white women, 8 % among African American women, 6.8 % among Spanish women, and 31.3 % among Asian women [9]. PCOS affects 15.3 % of Indian women, significantly impacting their lifestyles and potentially leading to disorders such as anxiety, sleep apnea, and metabolic syndromes. This increases the risk of developing conditions such as diabetes, endometrial cancer, and heart

* Corresponding author.
E-mail address: mmrmbstu@gmail.com (M.M. Rahman).

https://doi.org/10.1016/j.imu.2024.101500
Received 8 January 2024; Received in revised form 15 April 2024; Accepted 15 April 2024
Available online 27 April 2024

diseases [10]. Early diagnosis is crucial due to its impact. Symptoms stem from high androgen levels, causing issues such as weight gain, irregular periods, excessive body hair, and infertility. Hormonal imbalance disrupts ovulation cycles [11,12]. While the exact cause of PCOS remains unclear, insulin insensitivity is known to contribute, as evidenced by skin darkening in certain areas. Additionally, lifestyle factors such as diet and physical inactivity may also trigger PCOS [13,14]. Treatment focuses on managing symptoms: hormonal birth control regulates cycles and reduces testosterone; anti-androgens curb abnormal hair growth and acne; diabetes medications like metformin regulate insulin and cycles; fertility drugs aid ovulation. If medications fail, laparoscopic ovarian drilling can reduce testosterone production, potentially restoring ovulation [15].

With the assistance of artificial intelligence (AI), we are now able to detect and treat complex diseases [16]. Through the utilization of machine learning algorithms, PCOS can be detected earlier, before it advances to a severe state. Our proposed research has developed a web-based interface integrated with machine learning models to predict PCOS at an earlier stage using a user-friendly interface. PCOS presents a binary classification problem, focusing on determining whether a woman has PCOS or not. It provides a clear prediction with a yes or no outcome. Various models, such as LR, Gaussian NB, SVM, KNN, Bernoulli and Multinomial NB are employed for this prediction. These models can produce exact predictions, with each model producing different results, if they are given accurate data. The proposed system preprocesses data through steps such as string-to-integer conversion, filling null values, and removing unnecessary columns in early stage. For the feature selection, the Mutual information model is utilized to identify the best features before feeding the data into the learning stage. The prediction accuracy of our models, including Logistic Regression, Gaussian NB, Decision Tree, MLP, KNN, Linear SVC, Gradient Boosting Classifier, Bernoulli NB and Ada Boost Classifier, surpassed the accuracy of previous model. The highest accuracy was achieved by the Random Forest and Ada Boost, both reaching 94 %. Overview of the primary contributions of the paper.

- In this research, we utilized Standard Scaler and Mutual Information models as feature selection methods to optimize outcomes through the combination of two datasets.
- The proposed research involves the development of a Django-based web interface integrated with a machine learning algorithm. This interface is capable of predicting PCOS based on real-time inputted data, ensuring accuracy in the predictions.
- Thirteen classifiers and various machine learning algorithms, including preprocessing and hyperparameter tuning, are employed in this research. Both Random Forest and AdaBoost Classifier demonstrated the highest accuracy, reaching 94 %.

The entire technique is divided into the following sections: **Section 2** describes the existing research. The research methods and system views are described in **Sections 3** and **4** show the findings and experimental setup and lastly, conclusions added in **Section 5**.

2. Literature review

The collections of healthcare and biological information are expanding swiftly. To examine such substantial and intricate data, artificial intelligence and machine learning procedures have gained popularity [17]. Many academics have suggested AI-based PCOS diagnosis algorithms in recent years, utilizing clinical factors and vital signs as the datasets.

Silva et al. presented the BorutaShap technique and successively trained a random forest model in their article. A dataset included 73 healthy women and 72 PCOS patients. 58 features were ranked based on their importance and relevance. Ultimately, the model was able to achieve an accuracy of 86 % [18]. V.V. Khanna et al. proposed a model for predicting PCOS among fertile patients. They used an open-source dataset of 541 patients from Kerala, India. The authors employed ML models, including NB, DT, LR, KNN, RF, SVM, AdaBoost, Extra Trees and Gradient Boost, to predict PCOS. They also suggested multi-stacking ML. To make model predictions comprehensible, interpretable, and reliable, they employed Explainable AI approaches. The outcome indicated that multi-stacking ML achieved the highest accuracy of 98 % [19]. S. Bharati et al. focused on the data-driven diagnosis of PCOS in women. They utilized an open-source dataset comprising 541 women, including 177 diagnosed with PCOS. The authors employed a univariate feature selection technique along with LR, RF, Gradient Boosting and a hybrid RFLR model. To train and test the models, they divided the dataset using cross-validation and holdout techniques. According to the results, the RFLR model with UFS achieved the highest accuracy of 91.01 % [20]. A. Zigarelli et al. discussed the development of a predictive model for the self-diagnosis of PCOS through machine learning techniques. They employed a publicly accessible Kaggle PCOS dataset containing data from 541 patients, sourced from 10 distinct hospitals in Kerala, encompassing 44 features. The authors adopted the CatBoost classifier and evaluated its efficiency using K-fold validation, attaining an 82.5 % accuracy rate for invasive procedures and 90.1 % accuracy rate for non-invasive clinical indicators [21]. Y. A. Abu Adla et al. developed a model that utilizes machine learning techniques to automate the diagnosis of PCOS. They utilized a dataset containing information from 541 patients and 39 features, which were ranked based on their importance. To reduce the number of features, the authors suggested a hybrid feature selection strategy incorporating both filters and wrappers. Additionally, they used various machine learning models with the selected features to predict PCOS. SVM showed the highest accuracy of 91.6 % [22].

M. M. Hassan et al. suggested a model for identifying PCOS patients. To identify PCOS patients, researchers used a variety of machine learning techniques, including logistic regression, random forest, SVM, CART, and naïve bayes classification. Upon analyzing the outcomes, it was found that the random forest algorithm performed exceptionally well, achieving 96 % accuracy in PCOS diagnosis on the tested dataset [23]. A. Denny et al. presented a method for the early identification and prediction of PCOS using minimal yet promising clinical and metabolic characteristics, which serve as early markers for the illness. Principal Component Analysis (PCA) was employed by the authors to minimize the number of characteristics. To predict PCOS, they utilized LR, KNN, RF, NB and SVM, employing specific features. The outcome demonstrated that RF achieved the highest accuracy of 89.02 % [24]. S. Tiwari et al. proposed a model to diagnose PCOS based on a clinical dataset provided by kottarathil, available through its Kaggle repository. The authors employed correlation feature selection methods to choose a subset of features from the database. Using various machine learning models, including XGBoost, SVM, LR, RF, DT, KNN, Quadratic Discriminant Analysis, GB, AdaBoost, Linear Discriminant Analysis, and CatBoost, they identified the best model based on correlation thresholds. The trials showed that the Random Forest approach outperformed other well-known machine learning algorithms, achieving an accuracy of 93.25 % [25]. Bharadwaj et al. offered a thorough examination of the many clinical characteristics and how each affects a patient's likelihood of developing PCOS. They made use of an open-source Kaggle PCOS dataset, which included 44 features with the target label of PCOS diagnosis and the medical records of 541 individuals. The multicentric dataset was gathered from ten different Keralan hospitals. Machine learning classifiers and the Pearson correlation method for feature selection made up the architecture. The average size of the left and right follicles, the quantity of left follicles, hair growth, and prolactin levels were found to be the most important characteristics. A 93 % accuracy was achieved using the Multilayer Perceptron and SVM radial basis function kernel [26]. H. D. Mehr et al. used classical and ensemble classifiers to diagnose PCOS in their study on the Kaggle PCOS dataset. They considered the benefits of ensemble classifiers and feature selection techniques. Additionally, utilizing the dataset with all features and

Table 1
Comparison table of existing works.

Reference	Dataset	Method	Accuracy
Silva et al. [18]	73 Healthy Women and 72 PCOS Patients	BorutaShap, RF	86.00 %
V.V. Khanna et al. [19]	Dataset of 541 Patients from Kerala, India.	Multi-Stacking ML	98.00 %
S. Bharati et al. [20]	Dataset of 541 Patients.	RF, LR	91.01 %
A. Zigarelli et al. [21]	Dataset of 541 Patients from Kerala, India.	CatBoost	90.10 %
Y. A. Abu Adla et al. [22]	Dataset of 541 Patients from Kaggle.	Linear SVM	91.60 %
M. M. Hassan et al. [23]	Dataset of 541 Patients from Kaggle.	RF	96.00 %
A. Denny et al. [24]	Dataset of 541 Patients from Clinical Examinations.	RF	89.02 %
S. Tiwari et al. [25]	A Clinical Dataset from Kottarathil.	RF	93.25 %
Bharadwaj et al. [26]	Dataset of 541 Patients from Ten Different Keralan Hospitals.	MLP, SVM, RBF	93.00 %
H. D. Mehr et al. [27]	Dataset of 541 Patients from Kaggle.	RF	98.89 %

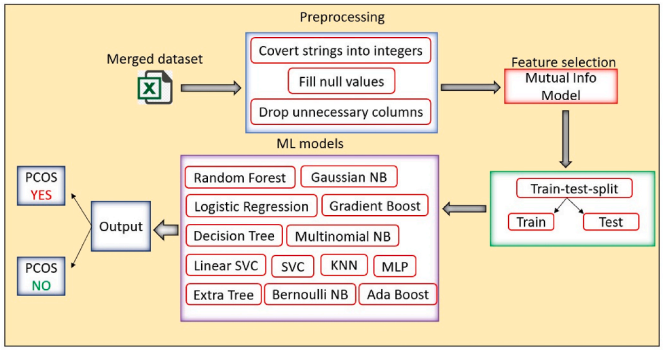


Fig. 1. Work flow of our proposed model.

reduced feature sets produced by filter, embedding, and wrapper feature selection techniques, the effectiveness of several classifiers, including AdaBoost, Random Forest, Multi-Layer Perception and Extra Tree, was examined. With an accuracy of 98.89 %, the Random Forest classifier outperformed other classifiers by utilizing the reduced subset of data based on the integrated feature selection approach [27].

Upon literature review, diverse machine learning classifiers such as RF, LR, CNN, SVM, K-Means, among others, demonstrated inconsistent outcomes in addressing Polycystic Ovary Syndrome. This variability in results poses a puzzling scenario, with each method offering distinct findings. To this end, the study employs a simple machine learning approach centered on feature extraction, exhibiting superior accuracy compared to previous research endeavors. Table 1 provides a concise summary of the related works in this domain.

3. Proposed methodology

The dataset acquired from Kaggle curated accurately, undertaking essential preprocessing steps. These steps included addressing null values, removing duplicate and deleting columns, and systematically converting all data from object type to numerical format. These steps ensured a robust foundation for further analyses. Now the dataset is ready for the application of various machine learning models. Before diving into this phase, the dataset is partitioned into distinct training and testing subsets. As part of our methodology, we introduced a noteworthy feature selection technique known as Mutual Information. This method, known for its efficacy, allowed us to discern the most influential

Table 2
Description of the first dataset with hyperparameter values.

Column	Description	Mutual Information Score
Sl. No	Serial no of dataset	
Patient File No.	Patient number	
PCOS (Y/N)	PCOS Status	
Age (Yrs.)	Age of the Patient	0.023806
Weight (Kg)	Weight in Kilogram	0.030802
Height (Cm)	Height in Centimeter	0.033325
BMI	BMI of Patients	0.018678
Blood Group	Patient's Blood Group	0.000000
Pulse rate (Bpm)	Pulse Rate	0.000000
RR (Breaths/min)	Respiratory Rate or Breathing rate (per min)	0.012108
Hb(G/dl)	Hemoglobin quantity	0.000000
Cycle(R/I)	Hormone problems happens during the reproductive years	0.067988
Cycle length (Days)	Number of days period lasts	0.069582
Marriage Status (Yrs.)	How long the Patient have been married	0.000000
Pregnant(Y/N)	Pregnant Status	0.002934
No. of abortions	Number of abortions	0.003673
FSH (mIU/mL)	FSH stands for Follicle-Stimulating Hormone quantity	0.031054
LH (mIU/mL)	Luteinizing Hormone quantity	0.034230
FSH/LH	Ration of FSH according to LH	0.066481
Hip (Inch)	Hip size in inches	0.047835
Waist (Inch)	Waist in inches	0.024143
Waist: Hip Ratio	Waist and Hip ratio	0.033307
TSH (mIU/L)	Thyroid-Stimulating Hormone quantity	0.019867
AMH (ng/mL)	Anti-Müllerian Hormone quantity	0.076651
PRL (ng/mL)	Prolactin Hormones quantity	0.061434
Vit D3 (ng/mL)	Vitamin D3	0.022579
PRG (ng/mL)	Progesterone hormones quantity	0.017166
RBS (mg/dl)	Random Blood Sugar amount	0.000000
Weight gain(Y/N)	Weight Gaining	0.082270
hair growth(Y/N)	Unwanted hair growth	0.115225
Skin darkening (Y/N)	Skin color Darkening	0.110074
Hair loss(Y/N)	Hair loss	0.002739
Pimples(Y/N)	Pimple appearing	0.023504
Fast food (Y/N)	Fast Food Habit	0.578986
Reg. Exercise(Y/N)	Regular Exercise activity	0.030983
BP _Systolic (mmHg)	Systolic Blood Pressure	0.021730
BP _Diastolic (mmHg)	Diastolic Blood Pressure	0.005281
Follicle No. (L)	Number of follicles in the left ovary	0.233207
Follicle No. (R)	Number of follicles in the right ovary	0.256563
Avg. F size (L) (mm)	Average follicle size in the left ovary	0.010435
Avg. F size (R) (mm)	Average follicle size in the right ovary	0.000000
Endometrium (mm)	Thickness of the endometrial lining in the uterus	0.029409

features. Specifically, we chose the top 12 features, focusing on their score. Our intent is to evaluate each model based on key performance metrics such as accuracy, precision, F1 score, and roc score. This comprehensive comparative analysis will give us the model that exhibits optimal results. The culmination of our investigation will guide our decision-making process, leading us to select the model that demonstrates superior performance. This chosen model will be seamlessly integrated into our user interface, ensuring a refined and effective user experience. Working process of our proposed method is shown in Fig. 1.

3.1. Experimental setup

The proposed classification structure for Polycystic Ovary Syndrome was built using machine learning models from scikit-learn library.

Table 3
Description of the second dataset hyperparameter values.

Column	Description	Mutual Information Score
Sl. No	Serial no of dataset	
Patient File No.	Patient number	
PCOS (Y/N)	PCOS Status [0 = Negative, 1 = Positive]	
I beta-HCG (mIU/mL)	beta-human chorionic gonadotropin (HCG) in the blood	0.000000
II beta-HCG (mIU/mL)	beta-human chorionic gonadotropin (HCG) in the blood	0.001751
AMH (ng/mL)	Anti-Müllerian Hormone quantity	0.076651

Jupyter Notebook was utilized to train and evaluate the model on a 64-bit Windows 10 PC. The computer is equipped with 4 GB of RAM and runs on a 2.11 GHz Intel 10th Generation Core i5 processor. The performance of thirteen models was assessed in terms of Accuracy, Sensitivity, Precision, and Recall. Mutual Information models were applied to each classifier to compare performance. Plot diagrams were generated to identify the optimal model for our dataset. The findings were compared between scenarios with and without the Mutual information model.

3.2. Performance matrix

A confusion matrix acts as a crucial tool in evaluating the efficiency of a machine learning model when applied to a specified dataset for classification purposes. It holds a significant role in measuring the model's performance by summarizing its accuracy in predicting categorical labels for input instances. In this study, when assessing metrics like Precision, Recall, False Positive Rate, True Negative Rate, and F1-Score, the confusion matrix comes into play. This matrix presents a comprehensive breakdown of four crucial measures: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These metrics provide a clear insight into the model's capability to correctly identify and classify positive and negative outcomes, aiding in the assessment and enhancement of the model's classification precision. Equation (1) represents the computing formula of F1-Score.

$$\mathbf{F1-Score} = 2\left(\frac{\boldsymbol{Sensitivity} * \boldsymbol{Precision}}{\boldsymbol{Sensitivity} + \boldsymbol{Precision}}\right) \quad (1)$$

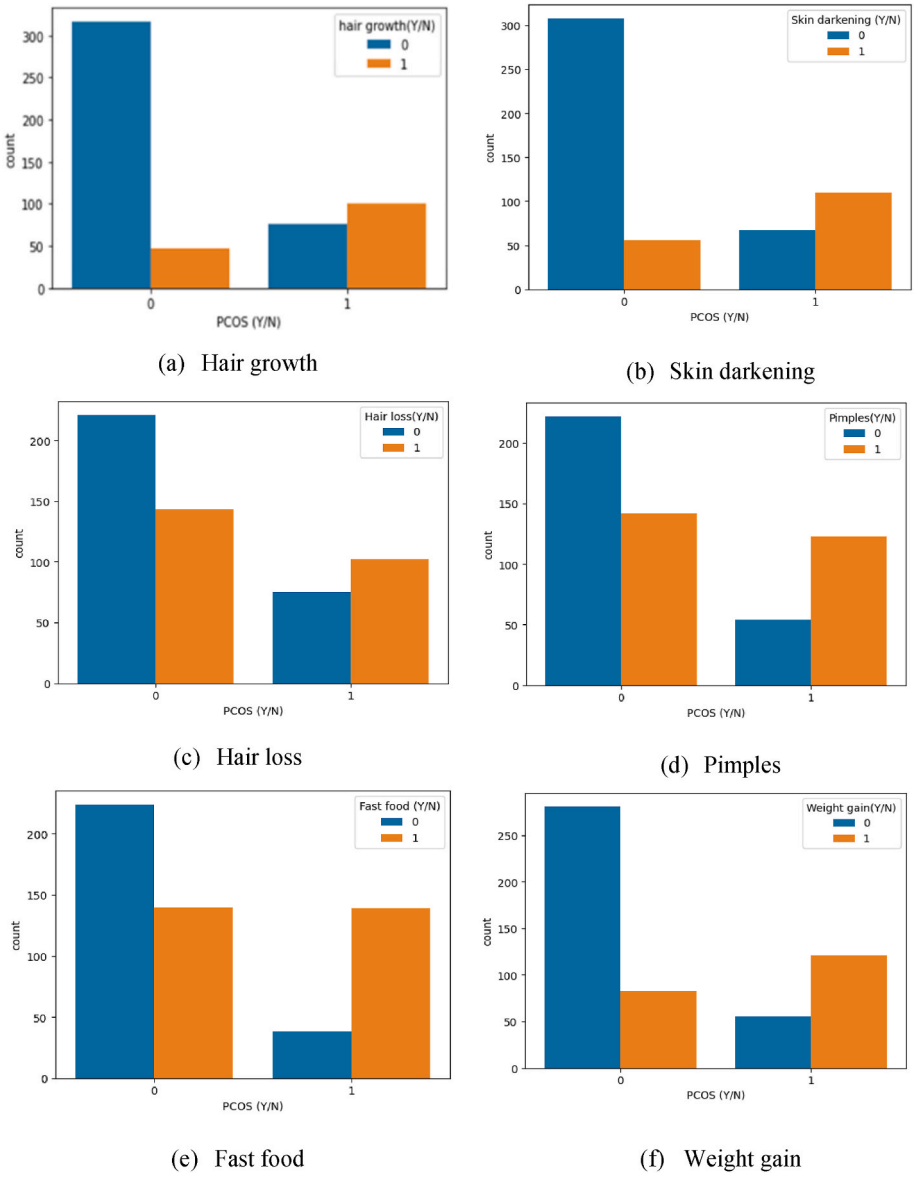


Fig. 2. Different symptoms of PCOS patients.

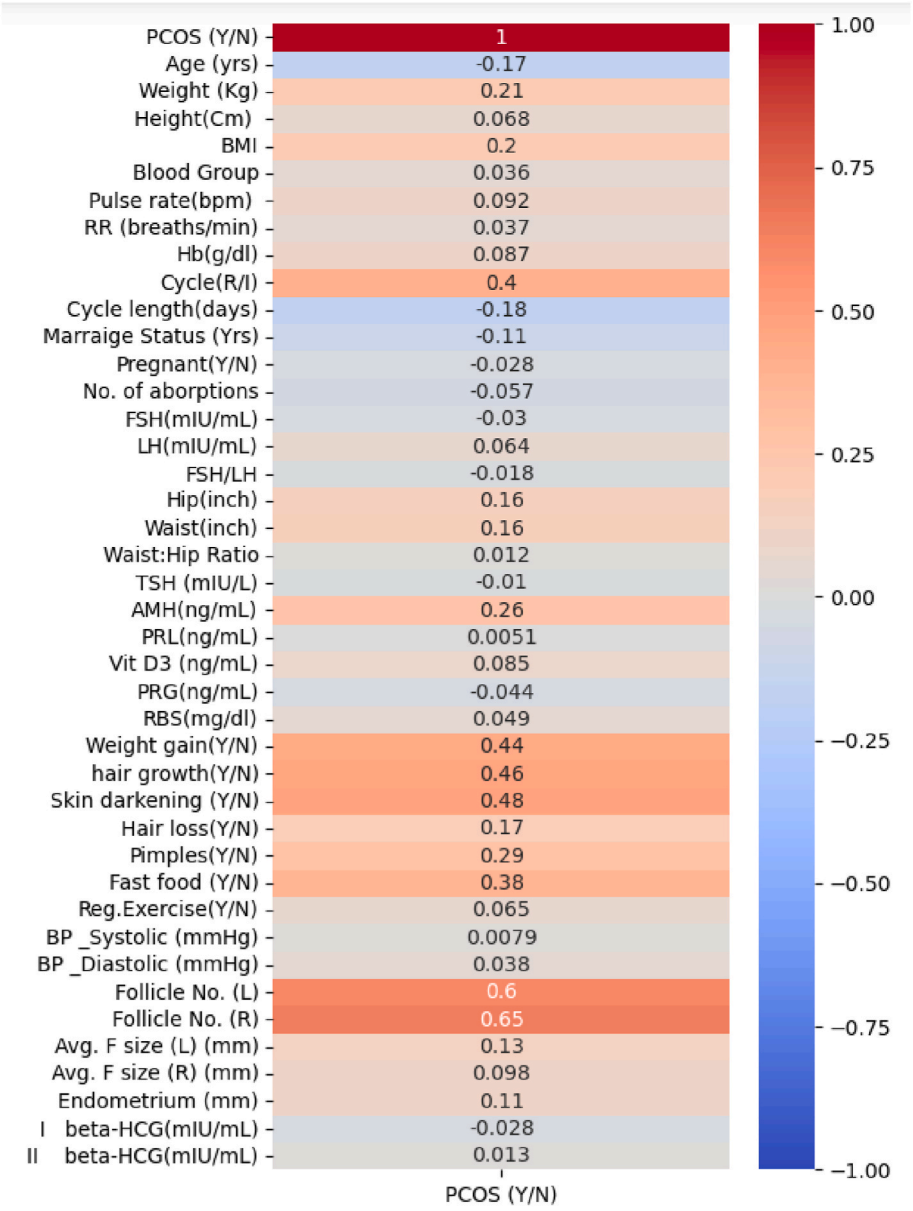


Fig. 3. The heat map color density of used data. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.3. Dataset

This dataset is sourced from the Kaggle, containing 541 patient records with 43 unique columns in both datasets combined. There are 364 PCOS-negative cases and 177 PCOS-positive cases. The 'PCOS (Y/N)' column represents the outcome, where 0 indicates a negative result and 1 indicates the patient has PCOS. In the entire dataset, 0 represents a 'negative' or 'no' while 1 represents 'positive' or 'yes'. Columns include weight, height, number of abortions, blood group, pulse rate, cycle length (days), pregnancy status (y/n), hair loss (y/n), pimples (y/n), etc. Table 2 shows the dataset along with descriptions of the columns. Proposed research worked with two tables containing records of the same patients. Table 3 includes an additional feature added to the main table during the preprocessing step.

3.4. Data preprocessing

In order to feed the dataset into various machine learning models, preprocessing of the data is necessary to achieve optimal results. Preprocessing methods include filling null values, removing duplicate values, eliminating unnecessary columns, and applying encoding methods, among others. The second dataset contains 6 columns, while the first dataset has 43 columns. These datasets contain results from various tests conducted on the same 541 patients. Four features are common to both datasets, so duplicates were removed after merging. As a result, the dataset consisted of 541 patients and 45 columns.

An unnamed column and columns named 'Sl. No.' and 'Patient File No.' were dropped as they had no impact on the output. The final data frame consisted of 42 columns and 541 rows. The column 'AMH (ng/mL)' contained strings, so it was converted to numeric using the 'to_numeric (.)' method from the pandas library. Additionally, the columns 'Marriage Status (yrs.)', 'Fast Food (Y/N)', and 'AMH (ng/mL)' contained null values, which were filled with the median value of each column. Before applying the machine learning models, the standard scaler method was applied to the dataset, except for the target column. The standard scaler converts the values into a range of numeric number

without losing any information, facilitating backend calculations for training the data model.

3.5. Data splitting

Before conducting the split, we merged two CSV files using the Pandas library, resulting in a new data frame with two additional unique columns. Initially, we extracted the "PCOS (Y/N)" target column and stored it as y, forming a data frame. The remaining columns were placed in x, creating two datasets: x for independent columns and y for the target or dependent column. Subsequently, we employed the train_test_split method from the scikit-learn library's model selection module to divide x and y into training and testing sets. The data was partitioned in an 80:20 ratios, with 80 % allocated for training and the remaining 20 % for testing. The entire dataset is eventually split for training and testing into four subsets: x_test, x_train, y_train, and y_test.

3.6. Hyperparameter tuning

Hyperparameters are valuable attributes that are typically utilized before applying classification models. Unlike traditional hyperparameter tuning methods such as GridSearchCV, Mutual Information (Mutual info) is the most fitting method in the hyperparameter realm. Mutual Information leverages information-theoretic principles to assess the dependence between hyperparameters and the model's performance metrics. This approach aids in identifying the most influential hyperparameters and their optimal values, leading to enhanced model performance. In the context of hyper parameter tuning, Mutual Information introduces a data-driven perspective, allowing for a more adaptive and insightful adjustment of hyperparameter values to improve overall model effectiveness. Based on the dataset, a value of 'k' is identified to determine the top features among all values. In Mutual Information, features are selected based on input and output combinations. In this feature selection metric, the top features of the dataset are selected using the 'mutual_info_regression' and 'mutual_info_classif' functions. In the proposed research, the top 12 parameters are selected from 45 features based on this model's achieved value. Tables 2 and 3 shows the hyperparameter values of our dataset.

3.7. Data visualization

There are two major symptoms of PCOS: irregular menstrual cycle, characterized by long delays between cycles, and obesity or excessive weight gain. Additionally, other symptoms can be identified though medical tests conducted at diagnostic centers. The dataset contains results from various diagnostic tests and surveys. In Fig. 2, a comparison bar chart illustrates key symptoms of PCOS. Factors such as hair growth, skin darkening, pimples, fast food consumption, regular exercise, and weight gain are examined. The chart displays the frequency of these symptoms among women with and without PCOS.

The bars indicate that women without PCOS are less likely to experience skin darkening, hair loss, and unwanted hair growth, while women with PCOS commonly exhibit these symptoms. Moreover, there is a notable association between PCOS and fast food consumption, as well as excessive weight gain, with PCOS patients showing a higher prevalence of these factors compared to those without PCOS.

The heat map holds significant importance in identifying connections among the features. Here, we observe colors ranging from blue to red. The deeper the red shade within a cell, the stronger the positive relationship between the corresponding features. Consequently, as one value increases, the other tends to increase as well. Conversely, deeper blue hues indicate a more negative relationship between the paired features. Thus, as one increases, the other tends to decrease. Cells appearing in white possess no discernible correlation between them. Therefore, cells closer to deep red or deep blue shades stand out as our most crucial features. Fig. 3 shows Follicle No. (L), Follicle No. (R), Weight gain, hair growth, and Skin darkening are the more responsible attributes for PCOS.

3.8. Machine learning models

Machine learning models empower prediction systems by leveraging data patterns to make accurate forecasts across diverse domains, enhancing decision-making and efficiency [28,29]. The effectiveness is assessed using 13 classifiers, specifically Logistic Regression, Decision Tree, Ada Boost, XGBoost, Bernoulli NB, Support Vector Machines, Random Forest, K-Nearest Neighbor, Naïve Bayes, Gradient Boosting, and Multilayer perceptron, all detailed in this part. Following that, the top-performing model, showing the utmost accuracy among these classifiers, is assessed.

3.8.1. Multi-layer perceptron

The Multi-layer perceptron, also known as MLP, is a feedforward neural network consisting of multiple layers interconnected with each other. It includes node values, weight values of the input layer, activation function values, and Sigmoid function values to calculate the output. Node values propagate from the input layer to the hidden layer and then to the output layer, indicating its functionality in the forward direction [30,31].

3.8.2. Linear support vector machine

The Support Vector Machine, also known as SVM, is a supervised machine learning classifier. Essentially, it identifies the closest or most similar data points within a class using a hyperplane [32]. When the object's features are two-dimensional, the hyperplane is a single line. However, if the number of object features exceed two, it becomes a multidimensional line. As the number of object features increases, it becomes more challenging to visualize [33]. SVM is primarily used for classification and regression tasks. It is also employed in image classification, text classification, email spam detection, face recognition, handwriting recognition, and code classification. The formula for the hyperplane involves the weight vector (w), bias (b), and input feature vector for a data point (x), as shown in **Equation (2)**.

$$\mathbf{f}(\mathbf{x}) = \mathbf{sign}(\mathbf{w} * \mathbf{x} + \mathbf{b}) \tag{2}$$

3.8.3. K-nearest neighbor

One of the basic and simplest ML model is KNN for supervised learning. KNN is a non-parametric algorithm because it does not make any assumptions about original data. The KNN algorithm makes predictions based on similar data in the dataset. The KNN algorithm is used for classification and regression. It is also used for text categorization, image recognition, and clustering. There are three types of distance matrices [34]. We can measure distance in the KNN algorithm using these formulas. Here are two types of equations, one for 2D data and another for multidimensional data. 'n' is the number of dimensions. 'x_i' is a data point in the training dataset. 'x_j' is the value of the feature for data point 'x'. 'x_{ij}' is the value of the j_{th} feature for data point 'x_i'. **Equation (3)** and **(4)** represent the 2D and multidimensional data.

$$\sqrt{(\mathbf{x_1} - \mathbf{x_{i1}})^2 + (\mathbf{x_2} - \mathbf{x_{i2}})^2} \tag{3}$$

$$\sqrt{\sum_{\mathbf{J}=1}^{\mathbf{n}} (\mathbf{x_J} - \mathbf{x_{ij}})^2} \tag{4}$$

3.8.4. Gradient boosting

Gradient Boosting is a versatile machine learning model that improves prediction accuracy by combining the strengths of weak models. Mainly, Gradient Boosting is used for decision tree classifiers [35]. It is a powerful boosting algorithm capable of both classification and regression tasks. It corrects errors at every step. By utilizing this boosting algorithm, we can achieve the highest level of accuracy. It can handle both

structured and unstructured data [36]. However, it may require a large amount of memory and is susceptible to issues such as overfitting, imbalanced data, complexity, and interpretability.

3.8.5. Logistic regression

Logistic regression is a widely used machine learning algorithm in supervised learning. Its primary function involves predicting the categorical dependent variable based on a provided set of independent variables [37]. The outcome must be a categorical or discrete value [38]. The logistic regression **Equation (5)** is expressed as: P(Y = 1|X), which denotes the likelihood that the dependent variable Y pertains to class 1 when considering the values of the independent variables X. The letter 'e' represents the base of the natural logarithm, roughly equivalent to 2.72. β_0 signifies the intercept or bias term, depicting the value of Y when all independent variables hold a value of zero. The coefficients β_1, β_2, …, β_p correspond to the independent variables X_1, X_2, …, X_p utilized in the model.

$$\mathbf{P(Y=1|X)} = \frac{\mathbf{1}}{\mathbf{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_P X_P)}}} \tag{5}$$

3.8.6. Extra tree

The Extremely Randomized Trees Classifier, also known as Extra Tree, is a member of the ensemble learning group within machine learning. It functions as an extension of the random forest algorithm and serves purposes in both classification and regression tasks [39]. The Extra Trees Classifier, similar to decision trees, does not have a mathematical formula for prediction like linear models [40]. There is no simple mathematical equation to represent this process. But there is a formula for calculating the entropy. In **Equation (6),** 'c' is the number of unique class labels and 'p_i' is the proportion of rows with output label is 'i'.

$$\text{Entropy(S)} = \sum_{\mathbf{i=1}}^{\mathbf{c}} - \mathbf{p_i \log_2(p_i)} \tag{6}$$

3.8.7. Multinomial naïve bayes

The Multinomial Naive Bayes algorithm is a probabilistic learning technique commonly applied in Natural Language Processing (NLP). Founded on the Bayes theorem, this algorithm forecasts the classification tag of textual content, such as an email or a newspaper article [41]. It computes the probability of each classification tag for a provided sample and subsequently outputs the tag with the highest probability [42]. In the subsequent **Equation (7),** P(B) signifies the prior probability of B. P(A) refers to the prior probability of class A, while P(B|A) indicates the probability of predictor B occurring given class A.

$$\mathbf{P(A|B) = P(A) * P(B|A)/P(B)} \tag{7}$$

3.8.8. Bernoulli naïve bayes

The Bernoulli Naive Bayes classifier is a variation of the Naive Bayes algorithm, which is a popular machine learning algorithm used for classification and text analysis tasks. It is particularly well-suited for problems where the features are binary or represent the presence or absence of certain attributes. It operates swiftly, effectively manages irrelevant features, and produces easily interpretable outcomes [43]. Consider a random variable X with the probability of success denoted by p and the probability of failure denoted by q. Here are **Formulas (8)** and **(9)**.

$$\mathbf{p(x) = P[X=x]} = \begin{cases} \mathbf{q=1-p} & \mathbf{x=0} \\ \mathbf{p} & \mathbf{x=1} \end{cases} \tag{8}$$

$$\mathbf{X} = \begin{cases} \mathbf{1} & \textbf{Bernoulli trial} = \mathbf{S} \\ \mathbf{0} & \textbf{Bernoulli trial} = \mathbf{F} \end{cases} \tag{9}$$

3.8.9. Gaussian naïve bayes

The Naive Bayes classifier, a supervised machine learning technique, is utilized in tasks such as text classification. Additionally, it falls under the category of generative learning algorithms, aiming to model the distribution of inputs within a specific class or category [44]. Using a Naive Bayes classifier and a Bayesian probabilistic model, it categorizes a subject i with a vector of attributes x_i into class c so that the posterior probability $P(c|x_{i1}, \ldots, x_{iM})$ is maximized [45]. This classifier is defined by **Equation (10)**. $P(X|Y)$ represents the probability input hypothesis for a particular set of data, based on probabilities $P(Y|X), P(X)$ and $P(Y)$.

$$\mathbf{P(X|Y)} = \frac{\mathbf{P(Y|X) \times P(X)}}{\mathbf{P(Y)}} \tag{10}$$

3.8.10. Decision tree

The Decision Tree is a non-parametric learning technique used in classification and regression for supervised learning. The nodes of a Decision Tree represent the input attributes. Internal nodes represent input factors, branching represents outcomes, and leaf nodes identify classes in the Decision Tree [46]. The entropy of each class is calculated using **Equation (11)**. **Equation (12)** provides an estimate of the entropy based on two features.

$$\mathbf{H(S)} = \sum_{\mathbf{i=1}}^{\mathbf{c}} \mathbf{-T_i \log_2 T_i} \tag{11}$$

$$\mathbf{H(S,D)} = \sum_{\mathbf{c \in D}}^{\mathbf{c}} \mathbf{X_{(C)} \cdot Y_{(c)}} \tag{12}$$

3.8.11. Random forest

Random Forest comprises an ensemble of decision trees. It selects the optimal subset of features by evaluating information gain. Through majority voting among multiple decision trees, it classifies an instance. The ultimate prediction is derived by combining the mean prediction for regression or the mode of classes for classification predictions obtained from all trees [47]. In other words, we can calculate the node probability by dividing the total number of samples collected by the total number of samples reaching the node. The greater the value, the more significant the attribute is. In **Equation (13)** the impurity is represented by the P_{ij}.

$$\mathbf{P_{ij} = L_j N_j - L_{left(j)} N_{left(j)} - L_{right(j)} N_{right(j)}} \tag{13}$$

3.8.12. Support vector machine

One of the main algorithms employed by data scientists is called Support Vector Machines. Although it can be used for classification and regression problems, its widespread use is due to the model's high accuracy and quick computation. One of the following kernel functions, such as linear, polynomial, radial basis, or quadratic, can be used to optimally classify an instance with an unknown class [48]. SVM, used by the kernel technique, implicitly transforms inputs into high-dimensional feature spaces, allowing for efficient non-linear classification. **Equation (14)** shows the calculation process of SVM.

$$\boldsymbol{D} = \left\{ \left(\boldsymbol{x^1}, \boldsymbol{y^1}\right), \ldots, \left(\boldsymbol{x^1}, \boldsymbol{y^1}\right) \right\}, \boldsymbol{x} \in \mathbb{R}, \boldsymbol{y} \in \tag{14}$$

3.8.13. AdaBoost

The AdaBoost classifier, functioning as a meta-estimator, begins by training a classifier with the initial dataset and enhances accuracy by combining multiple classifiers. Its fundamental idea involves refining data training and adjusting classifier weights in each step to ensure accurate predictions, especially for uncommon instances [49]. Moreover, AdaBoost assigns weights to classifiers based on their accuracy, granting more weight to those with higher precision in every iteration. The weighting vector w_n is determined using **equation (15)**. **Equation (16)** must be used to determine the value of w_n, which represents the initial value for the observation weights.

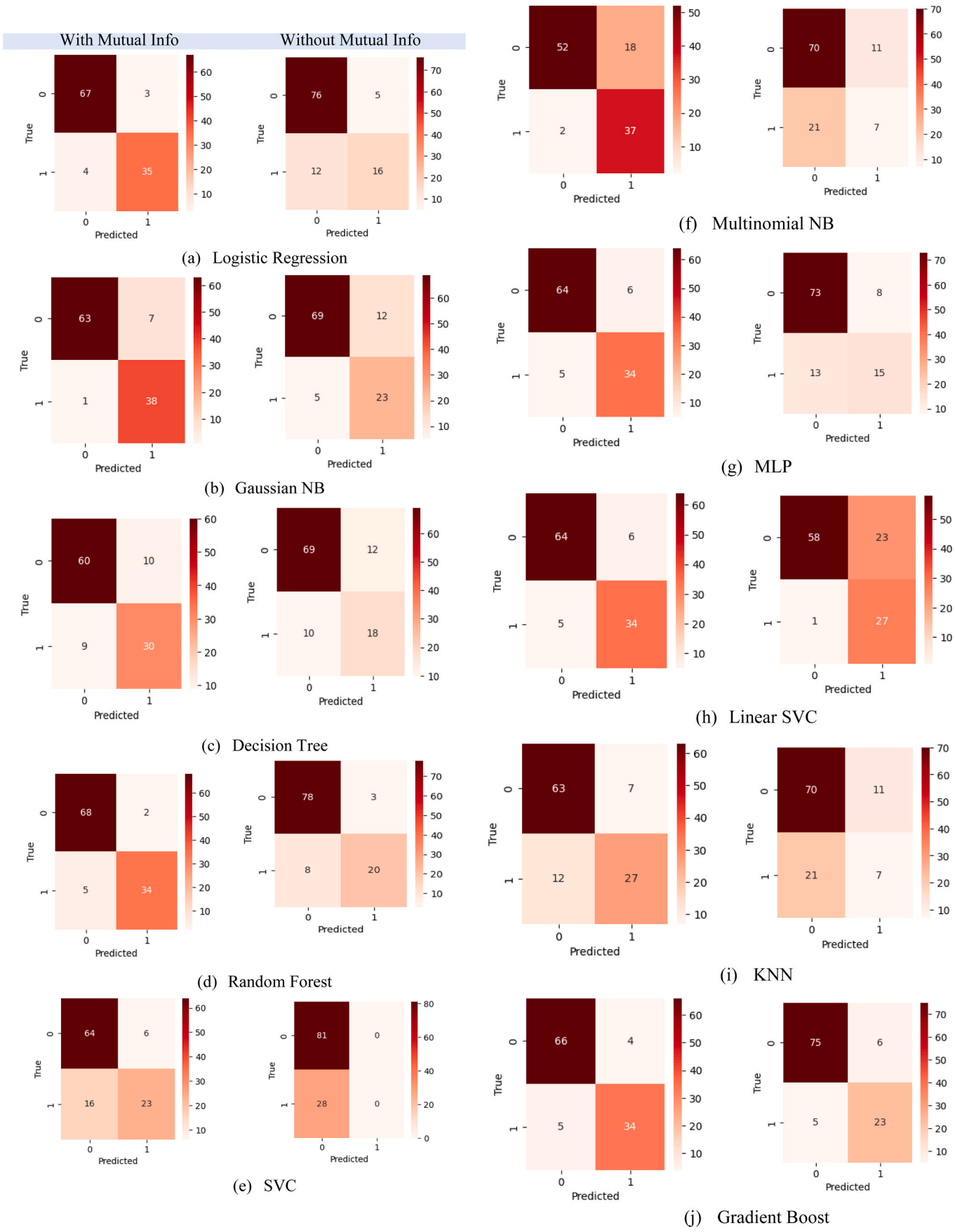
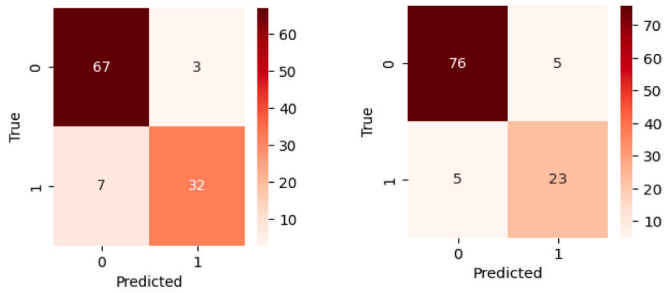
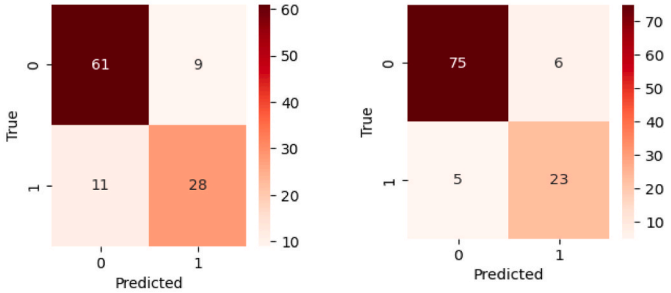


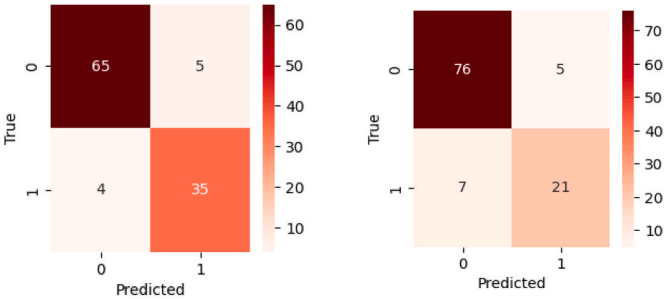
Fig. 4. Confusion matrix with and without Mutual Information model.



(k) Extra Trees



(l) Bernoulli NB



(m) AdaBoost

Fig. 4. (*continued*).

$$err_m = \frac{\sum_{i=1}^{N} w_i I\left(y_i \neq G_i(x_i)\right)}{\sum_{i=1}^{N} w_i} \tag{15}$$

$$a_m = \log\left(\frac{1 - err_t}{err_t}\right) \tag{16}$$

4. Result analysis and web interface

In this segment, we will outline the outcomes, deliberations, and web interface of the suggested approach. Here, we present the classification results for various machine learning models, including Logistic Regression, Decision Tree, AdaBoost, XGBoost, Bernoulli Naive Bayes, Support Vector Machines, K-Nearest Neighbor, Naive Bayes, Random Forest, Gradient Boosting, and Multilayer Perceptron. Each model underwent individual testing, utilizing the confusion matrix.

4.1. Result and discussion

The Mutual Information was applied to the dataset, targeting the output column PCOS (Y/N). The top 12 ranked features were selected, and all the ML models were applied to these features. The results, with and without the Mutual Information model, are shown in Fig. 4 (a) – (m) for 13 machine learning algorithms using the selected 12 features from the dataset. Using the Mutual Information model, most of the classifiers provided progressive results based on true positive values. Random Forest, Support Vector Classifier, Multinomial Naive Bayes, and Bernoulli Naive Bayes demonstrated better outcomes with Mutual Information, whereas Linear SVC's performance using Mutual Information was unsatisfactory. Other models showed average progression compared to those without Mutual Information. The true negative value of PCOS dramatically dropped for SVC classifiers, while others showed an average progression using the Mutual Information model.

The performance comparison between using and not using the

Table 4
Performance measurement of the PCOS dataset without using Mutual Information model.

Model	Accuracy	Precision	Recall	F1_score	ROC Score
LR	84 %	75.86	68.75	0.72	79.83
GNB	81 %	62.79	84.38	0.72	81.80
DT	83 %	70.59	75.00	0.73	81.01
RF	94 %	96.30	81.25	0.88	89.98
SV	72 %	100.00	3.12	0.06	51.56
MNB	34 %	27.27	75.00	0.40	45.94
MLP	72 %	53.57	46.88	0.50	65.00
Linear SVC	85 %	80.77	65.62	0.72	79.57
KNN	61 %	33.33	31.25	0.32	52.64
GB	92 %	87.10	84.38	0.86	89.59
ET	91 %	82.35	87.50	0.85	89.85
BNB	80 %	62.50	78.12	0.69	79.32
AB	92 %	87.10	84.38	0.86	89.59

Table 5
Performance measurement of the PCOS dataset with using Mutual Information model.

Model	Accuracy	Precision	Recall	F1_score	ROC Score
LR	91 %	87.88	82.86	0.85	88.73
GNB	89 %	84.85	80	0.82	86.62
DT	83 %	75.76	71.43	0.74	80.31
RF	**94 %**	**96.67**	**82.86**	**0.89**	**90.75**
SV	83 %	83.33	57.14	0.68	75.87
MNB	88 %	80.56	82.86	0.82	86.7
MLP	90 %	85.29	82.86	0.84	88.05
Linear SVC	92 %	90.62	82.86	0.87	89.4
KNN	84 %	87.5	60	0.71	77.97
GB	93 %	93.55	82.86	0.88	90.08
ET	91 %	90.32	80	0.85	87.97
BNB	77 %	61.36	77.14	0.68	77.08
AB	**94 %**	**93.94**	**88.57**	**0.91**	**92.93**

Mutual Information model is distinguishable. In terms of accuracy, Random Forest, Gradient Boosting, Extra Trees, and AdaBoost achieve success rates of more than 90 % without the Mutual Information feature engineering process. Conversely, Logistic Regression, Random Forest, Multi-layer Perceptron, Linear SVC, Gradient Boosting, Extra Trees, and AdaBoost demonstrate the highest success rate, exceeding 90 %, when utilizing the Mutual Information model. Based on ROC scores, Random Forest and AdaBoost classifiers yield better outcomes. Tables 4 and 5 display performance matrix values such as accuracy, precision, recall, F1 score, and ROC score with and without the use of the Mutual Information model.

The optimal threshold for a model, determined by the true positive ratio and false positive ratio values using the Mutual Information feature engineering model, exceeds 87 % in more than 9 models. The true positive ratio peaks in the AdaBoost classifier, while the false positive ratio is lowest in the Support Vector classifier. In Support Vector Classifier, K-nearest neighbor, and Multinomial Naive Bayes, the AUC score significantly improves using the Mutual Information model, although the final result remains unsatisfactory. The AUC score for the Decision Tree decreases with the application of the Mutual Information model, but scores for other models are successful with Mutual Information. Both the Random Forest and AdaBoost models demonstrate the highest AUC scores in both scenarios: with and without the Mutual Information model. Fig. 5 illustrates the ROC representation of the 13 machine learning classifiers.

When employing Mutual Information and utilizing the selected features, the accuracy obtained from the models is significantly better compared to not using Mutual Information. Through Mutual Information, the highest accuracy of 94 % was achieved by two models, namely Random Forest and AdaBoost, as shown in Fig. 7. Moreover, seven out of thirteen models achieved an accuracy of 90 % or higher. The lowest accuracy, at 77 %, was observed in the Bernoulli Naïve Bayes model, making it the only model that scored below 80 %. Conversely, without Mutual Information, the highest accuracy of 94 % was attained solely by the Random Forest model, and only four models achieved an accuracy of 90 % or above. The lowest accuracy recorded was 34 % in the Multinomial Naïve Bayes model, with three models scoring below 73 %, as demonstrated in Fig. 6.

A group of researchers has analyzed various machine learning models to efficiently predict PCOS. Most researchers utilized a unique dataset of 541 patients collected from Kaggle, integrating 43 features. In this dataset, traditional approaches such as scaling, null value removal, and data merging were employed to determine the most important features for engineering. Additionally, some researchers incorporated methods like Principal Component Analysis, Gini index, and correlation-based hyperparameter tuning. For the classification of PCOS, Random Forest, Multilayer Perceptron, and Linear SVM emerged as the most performant methods among all state-of-the-art algorithms, with all mentioned studies establishing a notable accuracy ranging from 85 % to

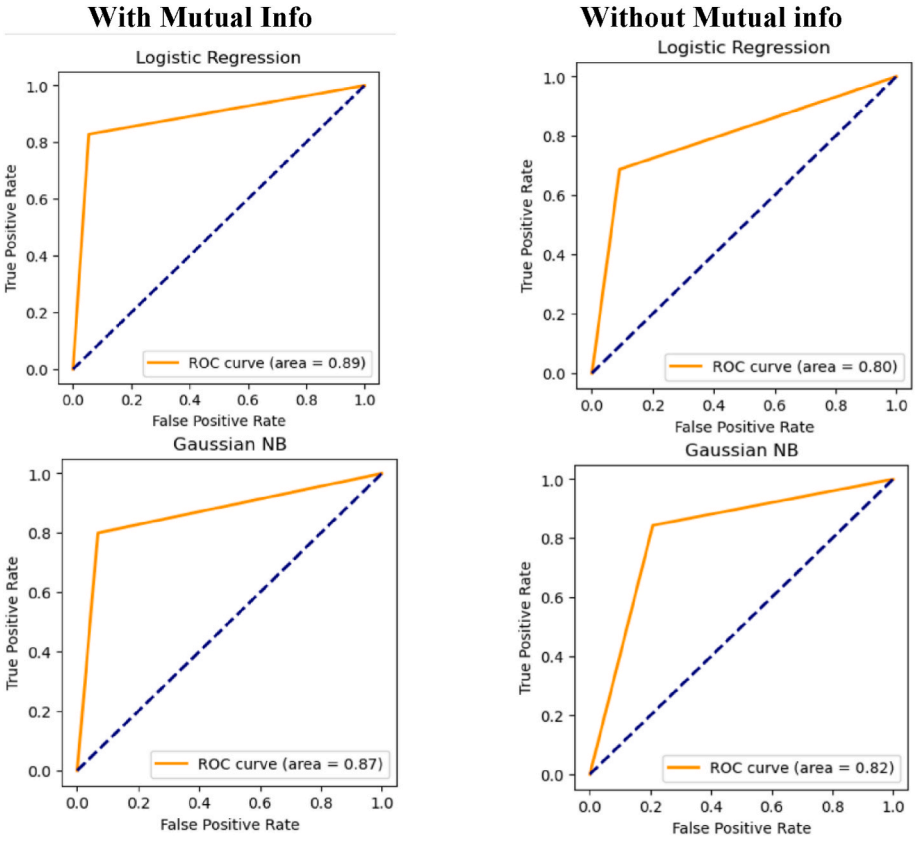


Fig. 5. The ROC curve for both mutual and without Mutual Information model.

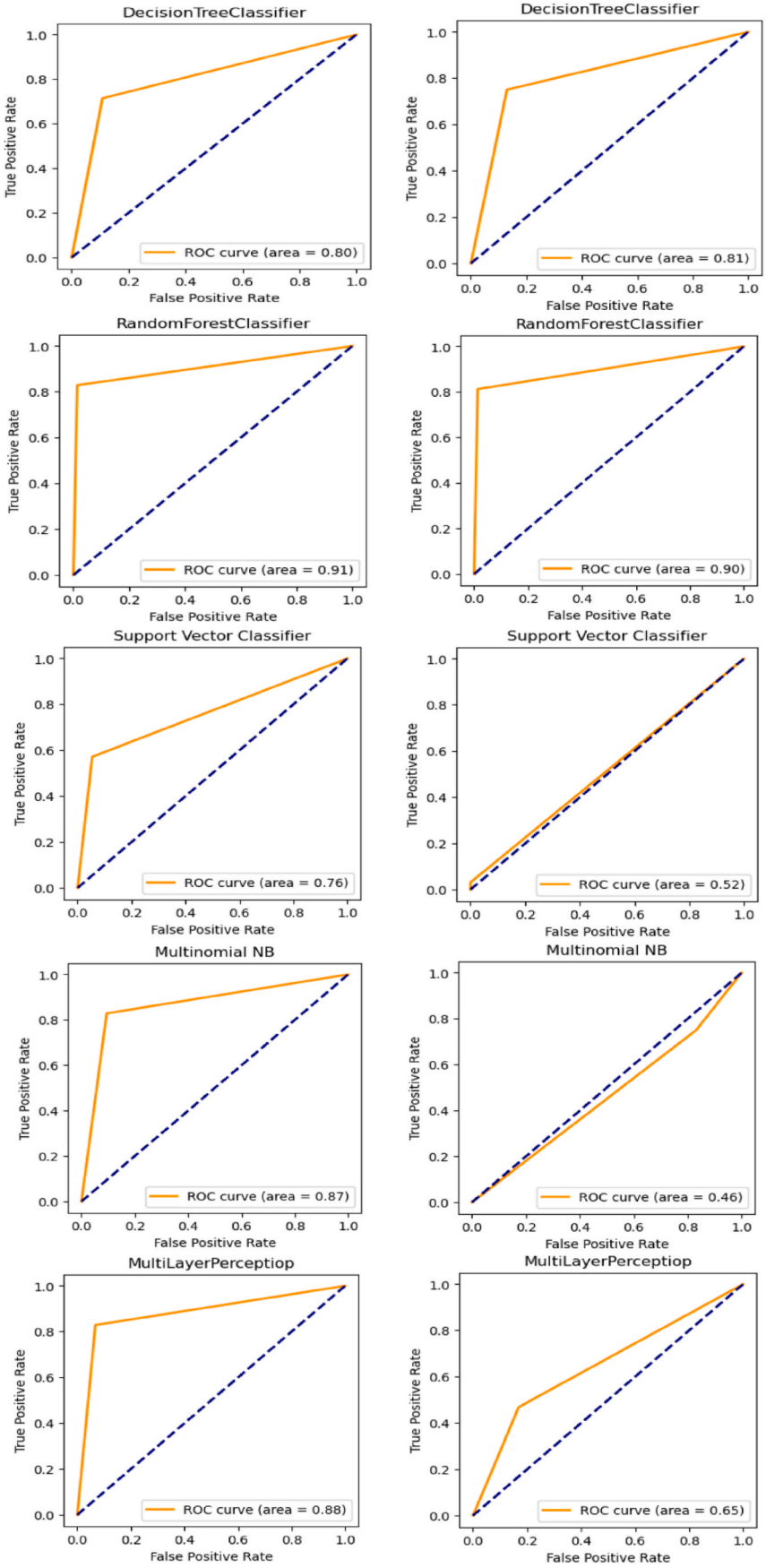


Fig. 5. (*continued*).

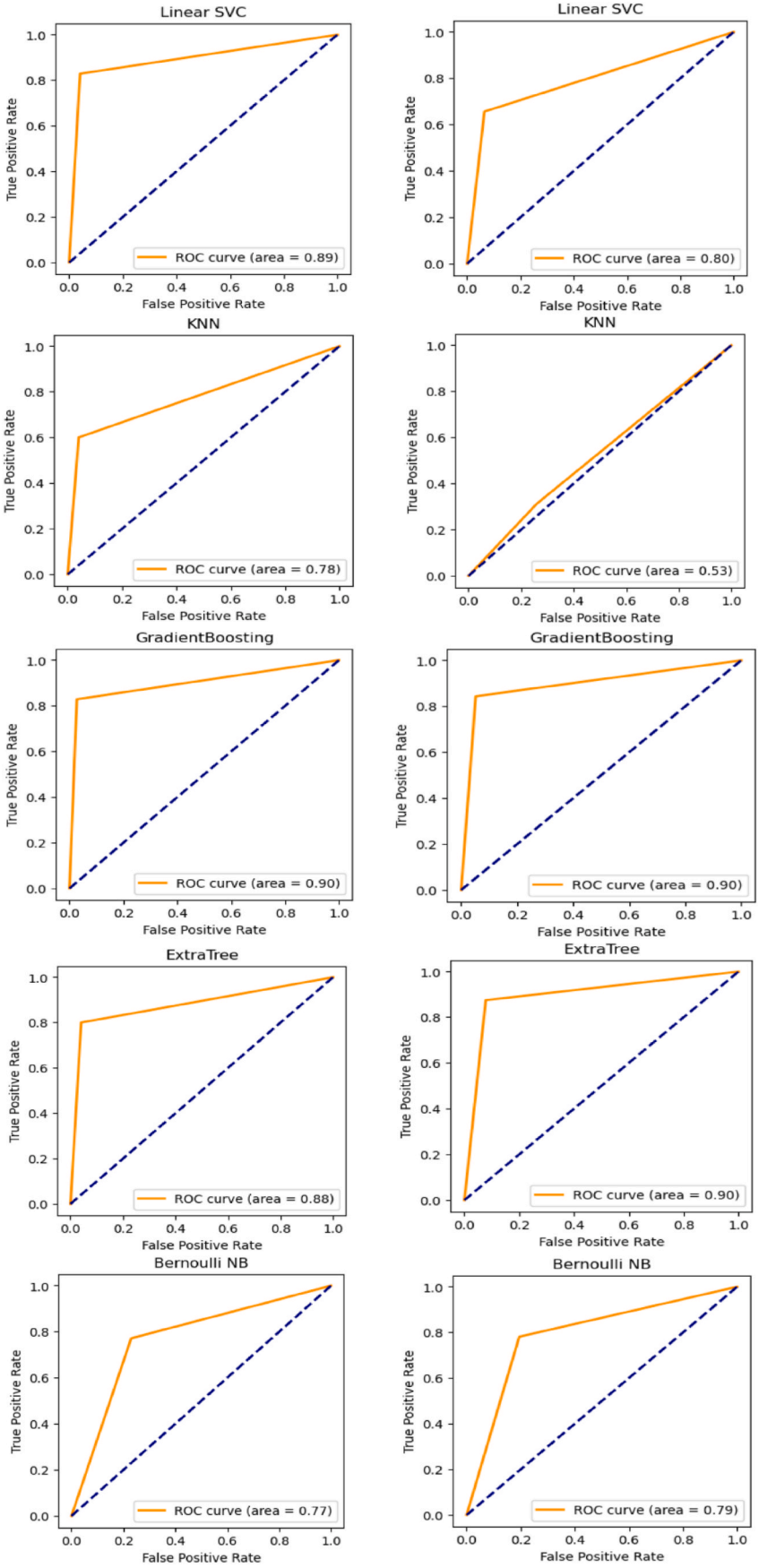


Fig. 5. (*continued*).

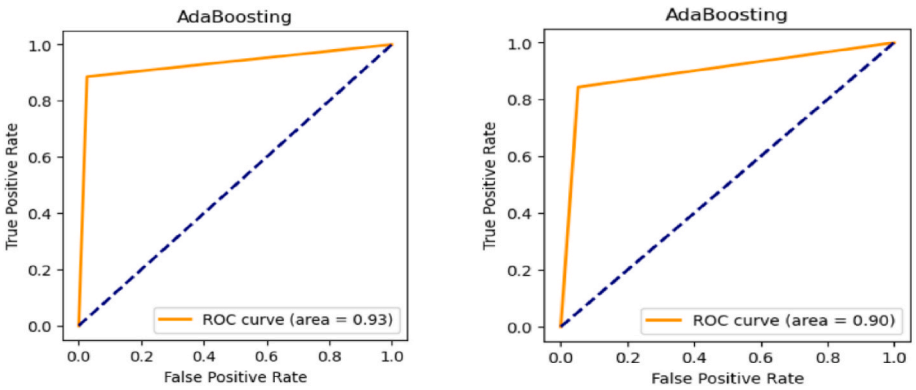


Fig. 5. (*continued*).

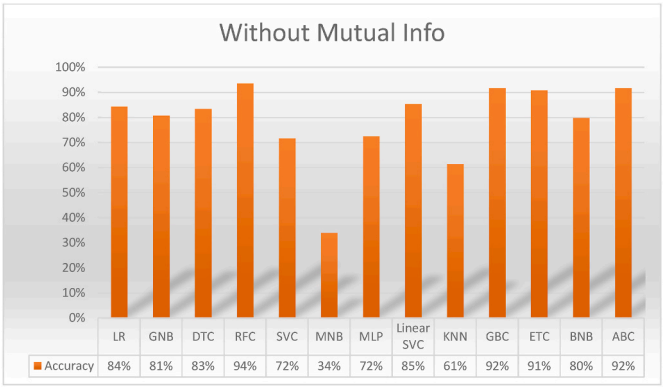


Fig. 6. Accuracy of different classifier without using Mutual Information model.

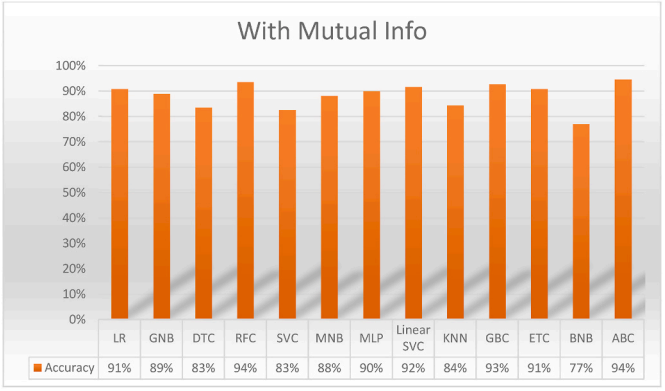


Fig. 7. Accuracy of different classifier using Mutual Information model.

93 %. The proposed research achieved a higher accuracy of 94 % using the same dataset of 541 patients. Mutual Information was employed to select the best-fitted features for predicting PCOS, ultimately identifying 12 features among the 45 features in the merged dataset for the classification of polycystic ovary syndrome. Table 6 represents the comparative analysis of recent works and the proposed work.

4.2. Web interface

The proposed method led to the development of a web-based system for Polycystic Ovary Syndrome detection. The user interface of the PCOS categorization system is constructed utilizing the Django framework. Initially, users need to create an account, and subsequently, they must log in to effectively utilize the system. Fig. 8 illustrates the operational steps of the proposed web-based system. To enhance user security and validity, the proposed system created registration and login forms shown in Fig. 9. While Fig. 10(a) demonstrates the process of uploading a test image for PCOS identification using this system. The user inputs an image, which is then compared to a set of trainable images. Upon completion of the PCOS identification process, the system predicts the types of fruits depicted in the input. The predicted outcome is displayed in Fig. 10(b).

5. Conclusion

PCOS patients may suffer from infertility, becoming unable to bear children if the condition remains unidentified in its early stages. Due to limitations in early detection, the rate of a PCOS patient increase is higher compared to previous years' records. In response, proposed research has developed a machine learning-based web interface prediction system. Early detection could empower patients to take necessary steps advised by their doctor, leading to a healthier life. The goal of our research is to utilize machine learning models for understanding the patterns of this disorder. These models are trained with data to showcase accuracy, specificity, sensitivity, precision, and overall performance using various ML algorithms such as Random Forest, Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, XGBoost Classifier, Support Vector Machines, among others. To achieve feature selection, we employed the Mutual Information method, showcasing a comparison of the highest accuracy attained when not using Mutual Information. Notably, Random Forest and AdaBoost Classifier achieved the highest accuracy of 94 %.

Availability of data and materials

All the materials can be retrieved from https://github.com/shimulmbstu/PCOS.

Ethical Statement for Informatics in Medicine Unlocked

I, Md. Mahbubur Rahman, hereby conscientiously assure that for the manuscript titled " **Empowering Early Detection: A Web-Based Machine Learning Approach for PCOS Prediction.**" the following criteria are fulfilled.

1) This material constitutes the authors' original work and has not been previously published elsewhere.
2) The paper is not presently under consideration for publication elsewhere.
3) The paper accurately reflects the authors' research and analysis in a truthful and comprehensive manner.

Table 6
Comparison analysis of recent works and proposed system.

Title	Number of data	Number of parameters used	Hyperparameter tuning	Method	Accuracy
Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models [18]	145	14	Scaling and merging	BorutaShap, RF	86.00 %
Diagnosis of polycystic ovary syndrome using machine learning algorithms [20]	541	10	Univariate feature selection algorithm	RFLR	91.01 %
Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: observational study [21]	541	10	Principal component analysis	CatBoost	90.10 %
Automated detection of polycystic ovary syndrome using machine learning techniques [22]	541	10	Hybrid feature selection approach	Linear SVM	91.60 %
Detection and prediction system for polycystic ovary syndrome using machine learning techniques [24]	541	8	Principal component analysis	RF	89.02 %
SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning [25]	541	10	Correlation-based feature selection	RF	93.25 %
Manoeuvre of Machine Learning Algorithms in Healthcare Sector with Application to Polycystic Ovarian Syndrome Diagnosis [26]	541	5	Gini index	MLP, RBF	93.00 %
Predicting polycystic ovary syndrome (PCOS) with machine learning algorithms from electronic health records [50]	30,601	3	Statistical feature selection	Multilayer perceptron	85 %
Proposed Method	**541**	**12**	**Mutual Info**	**RF, AB**	**94 %**

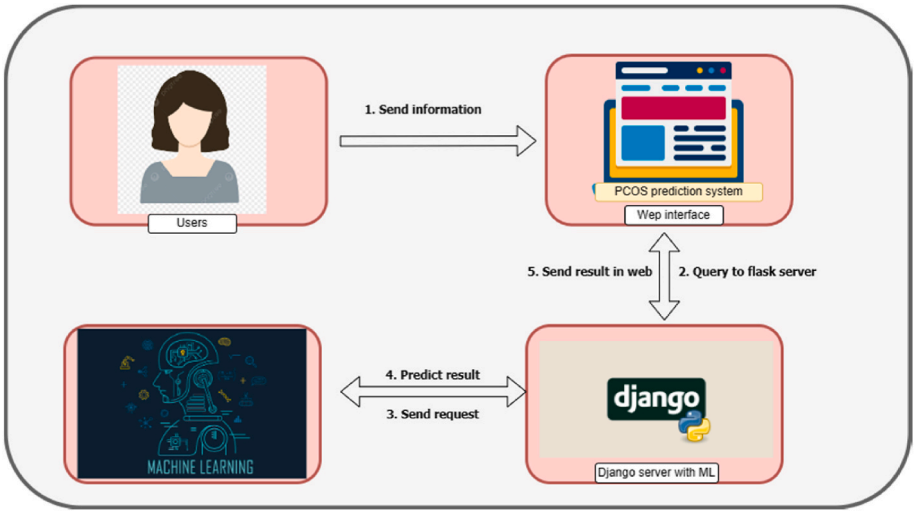


Fig. 8. Working procedures of web-based interface.

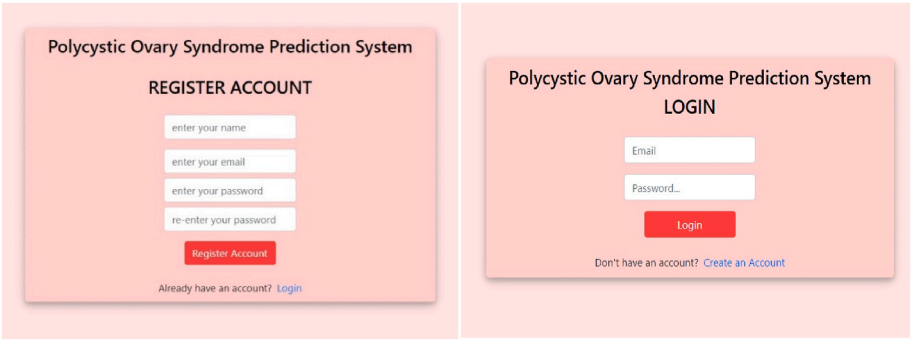


Fig. 9. User registration and login form of proposed web interface.

4) Proper credit is given to the significant contributions of co-authors and co-researchers.
5) The results are appropriately situated within the context of prior and existing research.
6) All utilized sources are appropriately disclosed through correct citation. Any verbatim text replication is clearly identified with quotation marks and accompanied by proper referencing.
7) All authors have actively and personally participated in substantial work leading to the paper and will assume public responsibility for its content.

I hereby affirm my agreement with the aforementioned statements and declare that this submission adheres to the policies of Informatics in Medicine Unlocked, as delineated in the Guide for Authors and the Ethical Statement.

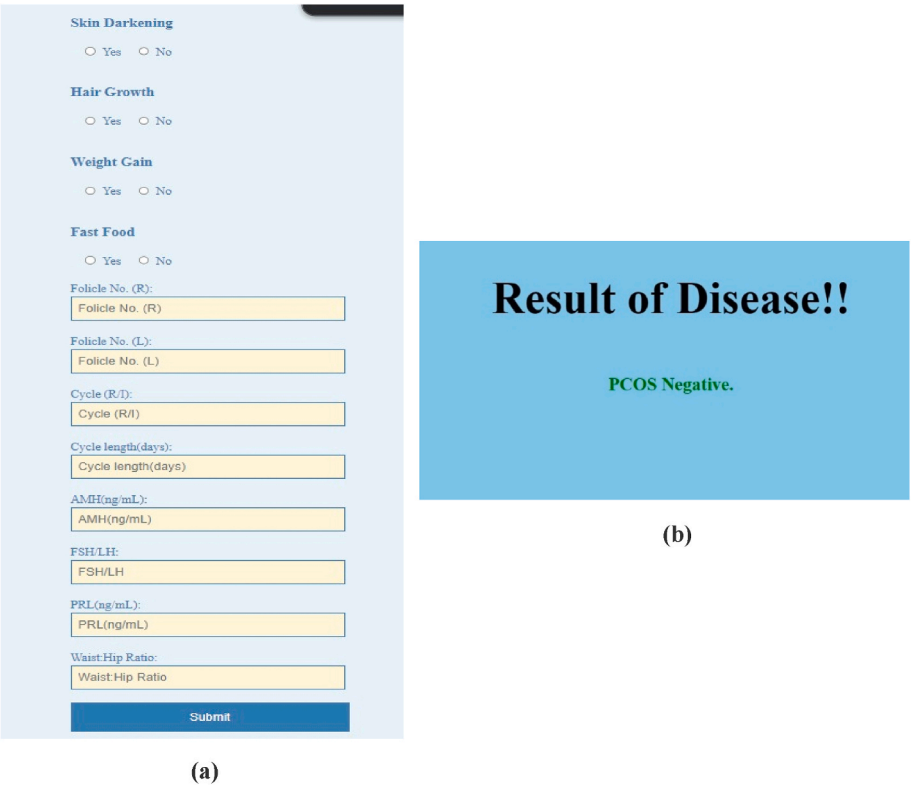


Fig. 10. input and result fields of proposed web interface.

CRediT authorship contribution statement

Md Mahbubur Rahman: Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Formal analysis, Data curation. **Ashikul Islam:** Visualization, Validation, Software, Methodology, Conceptualization. **Forhadul Islam:** Resources, Methodology, Formal analysis, Data curation. **Mashruba Zaman:** Writing – review & editing, Visualization, Investigation, Formal analysis. **Md Rafiul Islam:** Resources, Project administration, Methodology, Conceptualization. **Md Shahriar Alam Sakib:** Visualization, Software. **Hafiz Md Hasan Babu:** Writing – review & editing, Supervision, Project administration, Formal analysis.

Declaration of competing interest

The authors declare no competing financial/non-financial interests.

Acknowledgement

We are very thankful to Dr. SR Begum, Senior Consultant, Department of Obstetrician and Gynecologist (Women), Square Hospital, Bangladesh for her rich confirmation, productive suggestions and consultancy.

References

[1] Oğuz SH, et al. The prevalence, phenotype and cardiometabolic risk of polycystic ovary syndrome in treatment-naïve transgender people assigned female at birth. Endocrine 2023:1–6.
[2] Kulkarni S, et al. Polycystic ovarian syndrome: Current scenario and future insights. Drug Discov Today 2023:103821.
[3] Zhao H, et al. Insulin resistance in polycystic ovary syndrome across various tissues: an updated review of pathogenesis, evaluation, and treatment. J Ovarian Res 2023;16(1):9.
[4] Ma L, et al. The life cycle of the ovary. In: Ovarian aging. Springer; 2023. p. 7–33.
[5] Hajam YA, et al. Herbal medicine applications for polycystic ovarian syndrome. CRC Press; 2023.
[6] Wu T, et al. The cellular and molecular mechanisms of ovarian aging. In: Ovarian aging. Springer; 2023. p. 119–69.
[7] Adashi EY, et al. The polycystic ovary syndrome: the first 150 years of study. F&S Reports 2023;4(1):2–18.
[8] Miles KS. Our pearls matter: PCOS through the lens of women of color and white women. 2023.
[9] Bhat SA. Detection of polycystic ovary syndrome using machine learning algorithms. Dublin: National College of Ireland; 2021.
[10] Benjamin JJ, et al. Stress and polycystic ovarian syndrome-a case control study among Indian women. Clinical Epidemiology and Global Health 2023;22:101326.
[11] Karkera S, Agard E, Sankova L. The clinical manifestations of polycystic ovary syndrome (PCOS) and the treatment options. European Journal of Biology and Medical Science Research 2023;11(1):57–91.
[12] Shankar DY, et al. Overview of polycystic ovary syndrome (PCOS). World Journal of Advanced Engineering Technology and Sciences 2023;8(2):11–22.
[13] Dadoush SFM. Diagnosing and treating the causes of women's polycystic ovary syndrome: clinical and prospective study. African Journal of Advanced Pure and Applied Sciences (AJAPAS) 2023:401–7.
[14] Pramod A. Dietary and physical activity pattern IN PCOS women. St Teresa's College (Autonomous); 2023. Ernakulam.
[15] Della Corte L, et al. Is there still a place for surgery in patients with PCOS? A review. Life 2023;13(6):1270.
[16] Rahman Md Mahbubur. A web-based heart disease prediction system using machine learning algorithms. Network Biology 2022;12(2):64.
[17] Rahman MM, et al. Proposing a hybrid technique of feature fusion and convolutional neural network for melanoma skin cancer detection. J Pathol Inf 2023;14:100341.
[18] Silva I, et al. Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models. J Endocrinol Invest 2022:1–9.
[19] Khanna VV, et al. A distinctive explainable machine learning framework for detection of polycystic ovary syndrome. Applied System Innovation 2023;6(2):32.
[20] Bharati S, Podder P, Mondal MRH. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In: 2020 IEEE region 10 symposium (TENSYMP). IEEE; 2020.
[21] Zigarelli A, Jia Z, Lee H. Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: observational study. JMIR Formative Research 2022;6(3):e29967.
[22] Adla YAA, et al. Automated detection of polycystic ovary syndrome using machine learning techniques. In: 2021 Sixth international conference on advances in biomedical engineering (ICABME). IEEE; 2021.
[23] Hassan MM, Mirza T. Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. Int J Comput Appl 2020;975:8887.
[24] Denny A, et al. i-hope: detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques. In: TENCON 2019-2019 IEEE region 10 conference (TENCON). IEEE; 2019.

[25] Tiwari S, et al. SPOSDS: a smart Polycystic Ovary Syndrome diagnostic system using machine learning. Expert Syst Appl 2022;203:117592.
[26] Bhardwaj P, Tiwari P. Manoeuvre of machine learning algorithms in healthcare sector with application to polycystic ovarian syndrome diagnosis. In: Proceedings of academia-industry consortium for data science: aicds 2020. Springer; 2022. p. 71–84.
[27] Danaei Mehr H, Polat H. Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. Health Technol 2022;12(1):137–50.
[28] Rahman MM, et al. A deep CNN approach to detect and classify local fruits through a web interface. Smart Agricultural Technology 2023;5:100321.
[29] Rahman MM, Khan MSI, Babu HMH. BreastMultiNet: a multi-scale feature fusion method using deep neural network to detect breast cancer. Array 2022;16:100256.
[30] Kruse R, et al. Multi-layer perceptrons. In: Computational intelligence: a methodological introduction. Springer; 2022. p. 53–124.
[31] Xie S, et al. Multi-scale and multi-layer perceptron hybrid method for bearings fault diagnosis. Int J Mech Sci 2022;235:107708.
[32] Najjar E, Breesam AM. Supervised machine learning a brief survey of approaches. Al-Iraqia Journal for Scientific Engineering Research 2023;2(4):71–82.
[33] Çakir M, et al. Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. J King Saud Univ Sci 2023;35(6):102754.
[34] Tyagi A, Singh VP, Gore MM. An efficient automated detection of schizophrenia using k-NN and bag of words features. SN Computer Science 2023;4(5):518.
[35] Liu J, et al. Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach. J Forecast 2023;42(5):1112–37.
[36] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev 2021;54:1937–67.
[37] Madushani JS, et al. Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers. Transport Eng 2023;13:100190.
[38] Praveena R, et al. Prediction of rainfall analysis using logistic regression and support vector machine. In: Journal of physics: conference series. IOP Publishing; 2023.
[39] Roy KS, et al. MalHyStack: a hybrid stacked ensemble learning framework with feature engineering schemes for obfuscated malware analysis. Intelligent Systems with Applications 2023;20:200283.
[40] Berrouachedi A, Jaziri R, Convolutional G Bernard. Extra-trees and multi layer perceptron. In: 2022 IEEE/ACS 19th international conference on computer systems and applications (AICCSA). IEEE; 2022.
[41] Rezaeian N, Novikova G. Persian text classification using naive bayes algorithms and support vector machine algorithm. Indonesian Journal of Electrical Engineering and Informatics (IJEEI) 2020;8(1):178–88.
[42] Liu X, et al. Developing multi-labelled corpus of twitter short texts: a semi-automatic method. Systems 2023;11(8):390.
[43] Guo J, et al. Predicting and extracting thermal behavior rules of hydronic thermal barrier with interpretable ensemble learning in the heating season. Energy Build 2023;301:113699.
[44] Bandi A, Adapa PVSR, Kuchi YEVPK. The power of generative ai: a review of requirements, models, input–output formats, evaluation metrics, and challenges. Future Internet 2023;15(8):260.
[45] Linero AR, Antonelli JL. The how and why of Bayesian nonparametric causal inference. Wiley Interdisciplinary Reviews: Comput Stat 2023;15(1):e1583.
[46] Meng X, et al. Construction of decision tree based on C4. 5 algorithm for online voltage stability assessment. Int J Electr Power Energy Syst 2020;118:105793.
[47] James G, et al. Tree-based methods. In: An introduction to statistical learning: with applications in Python. Springer; 2023. p. 331–66.
[48] Cerulli G. Discriminant analysis, nearest neighbor, and support vector machine. In: Fundamentals of supervised machine learning: with applications in Python, R, and stata. Springer; 2023. p. 147–200.
[49] Ding Y, et al. An efficient AdaBoost algorithm with the multiple thresholds classification. Appl Sci 2022;12(12):5872.