applied system innovation

MDPI

Article

A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome

Varada Vivek Khanna [1], Krishnaraj Chadaga [2], Niranajana Sampathila [1,*], Srikanth Prabhu [2,*], Venkatesh Bhandage [2] and Govardhan K. Hegde [2]

[1] Department of Biomedical Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India
[2] Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India
* Correspondence: niranjana.s@manipal.edu (N.S.); srikanth.prabhu@manipal.edu (S.P.)

Abstract: Polycystic Ovary Syndrome (PCOS) is a complex disorder predominantly defined by biochemical hyperandrogenism, oligomenorrhea, anovulation, and in some cases, the presence of ovarian microcysts. This endocrinopathy inhibits ovarian follicle development causing symptoms like obesity, acne, infertility, and hirsutism. Artificial Intelligence (AI) has revolutionized healthcare, contributing remarkably to science and engineering domains. Therefore, we have demonstrated an AI approach using heterogeneous Machine Learning (ML) and Deep Learning (DL) classifiers to predict PCOS among fertile patients. We used an Open-source dataset of 541 patients from Kerala, India. Among all the classifiers, the final multi-stack of ML models performed best with accuracy, precision, recall, and F1-score of 98%, 97%, 98%, and 98%. Explainable AI (XAI) techniques make model predictions understandable, interpretable, and trustworthy. Hence, we have utilized XAI techniques such as SHAP (SHapley Additive Values), LIME (Local Interpretable Model Explainer), ELI5, Qlattice, and feature importance with Random Forest for explaining tree-based classifiers. The motivation of this study is to accurately detect PCOS in patients while simultaneously proposing an automated screening architecture with explainable machine learning tools to assist medical professionals in decision-making.

Keywords: deep learning; explainable artificial intelligence; local Interpretable model explainer; shapley additive values; machine learning; polycystic ovary syndrome

Citation: Khanna, V.V.; Chadaga, K.; Sampathila, N.; Prabhu, S.; Bhandage, V.; Hegde, G.K. A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. *Appl. Syst. Innov.* **2023**, *6*, 32. https://doi.org/10.3390/asi6020032

Academic Editor: Friedhelm Schwenker

Received: 29 January 2023
Revised: 19 February 2023
Accepted: 21 February 2023
Published: 23 February 2023

1. Introduction

Polycystic Ovary Syndrome (PCOS) consists of symptoms occurring because of abnormally elevated androgen levels [1]. Androgens are widely categorized as male sex hormones. However, when produced in smaller amounts play a vital role in the effective functioning of the female reproductive system. Most patients of PCOS are women of reproductive age experiencing weight gain, irregular menstrual cycles, excessive body hair, hair thinning or male-pattern balding, acne or oily skin, and at times infertility [2]. The hormone imbalance can hinder ovarian follicle development, inhibiting a normal ovulation cycle. 8.2% to 22.5% of women in India, suffer from PCOS [3]. This endocrinopathy can significantly impact the patient's lifestyle, wherein the patient may face depression, anxiety, eating disorders, and sleep apnea. Further, metabolic syndromes can put them at risk of cardiovascular disorders, endometrial cancer, and diabetes mellitus [4,5].

The exact cause of PCOS is still unknown. However, researchers have discovered That cells become resistant to insulin, thereby increasing blood sugar levels. The external manifestation of insulin resistance is usually skin darkening around the neck and armpits [6]. A sedentary, inactive lifestyle and an improper diet can also contribute to a woman getting PCOS.

Artificial Intelligence (AI) is a vast domain characterized by a machine having the ability and intelligence to perceive, infer, process, synthesis, and forecast information without much human interference. AI has been deployed in various sectors, such as banking, finance, agriculture, education, business, engineering, sociology, forensics, and medicine [7]. AI is extensively used for processing and providing insights to improve patient health outcomes [8]. Machine learning and deep learning can efficiently process medical data such as bio-signal and medical images [9] Another rapidly expanding branch of AI with extensive research is fuzzy systems. [10,11] Further, with efficient networks, researchers can scale their machine learning and deep learning architecture [12,13]. Automated diagnosis, patient-triaging, severity prediction, drug discovery, treatments such as computerized drug delivery, precision medicine, prognosis, and decision-making assistance are all possible today owing to the virtue of AI technology [14–17].

Explainable AI (XAI) can be characterized by system transparency and interpretability. In recent years, XAI has helped solve issues of biases, unfairness, safety, and causality [18]. Integrating XAI with AI-based systems for diagnostic, prognostic, and treatment purposes could aid in achieving accountability and increase trust in the medical decision made by the system [19]. Clinical validation of ML and DL models can be made possible with XAI. This study focuses on applications of SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), ELI5, Qlattice and Feature importance with Random Forest for screening PCOS. SHAP uses a game theory approach and provides mathematical values for accuracy and explanation consistency. SHAP provides a feature-importance-based ranking. LIME helps in hypothesis verification and gives insights on potential overfitting to noise. This technique explains local predictions made for each data sample [20]. ELI5 deploys a feature-based weight assignment technique to create a tree map explaining holistic and individual predictions [21]. Qlattice produces a classification model trained on raw data and produces QGraphs followed by a simplified expression to describe the graph [22]. Researchers have established a pathophysiological link between increased LH, insulin resistance, increased estrogen, and decreased FSH levels to PCOS, which is the most common cause of reduced fertility [23]. In this study, we have explored an open-source dataset of fertile PCOS patients and proposed a Machine learning-based multi-level stack to create a PCOS screening decision support. The contributions of this article are as follows: (1) Assessment of the significant PCOS features extracted by three feature selection methods: Salp swarm optimization, Harris hawk optimization, and mutual information. (2) Evaluation of the performance of ML classifiers such as LR, SVM (Kernel: linear, polynomial, Gaussian, Sigmoid), DT, RT, XGBoost, AdaBoost, and ExtraTrees classifiers and creating an improved and reliable classifier by an ensemble stacking of meta-learners. (3) Analysis of our customized multi-level stack against Deep Neural Networks and 1-D CNN models for screening PCOS. (4) Analysis of an XAI layer of our tree-based framework with SHAP, LIME, ELI5, and Qlattice, and feature importance with Random Forest. The rest of the article follows: Section 2 discusses various related work, Section 3 highlights the materials and methods. An in-depth result analysis is conducted in Section 4. Section 5 provides a comparison finding with existing research. The last section comprises of conclusion and future scope.

2. Literature Review

In recent years, many researchers have proposed AI models for PCOS diagnosis, with datasets consisting of clinical parameters and vital signs. Bharadwaj et al. [24] provided a detailed analysis of various clinical features and their contribution to a patient having PCOS. They used an open-source Kaggle PCOS dataset [25], with records of 541 patients and 44 features whose target label was PCOS diagnosis. The dataset is multicentric and was collected from 10 different hospitals in Kerala. The architecture consisted of machine learning classifiers and the Pearson correlation technique for feature selection. They concluded, the most critical features were: the average left and right follicle size, number of follicles in the left, hair growth, and prolactin levels. The SVM Radial Basis Function (RBF)

kernel, and Multilayer Perceptron (MLP) obtained an accuracy of 93%. Zigarelli et al. [26] used the same dataset, adopting the CatBoost classifier with K-fold validation evaluating the classifier performance, achieving 82.5% for the invasive methods and 90.1% accuracy for the non-invasive clinical parameters. Bharati et al. [27] suggested a CatBoost model and a voting hard and voting soft classifier. They extracted the top 13 features by univariate feature selection followed by a hold-out and cross-validation for the data splitting. The soft voting classifier performed with an accuracy of 91.12%. Tiwari et al. [28] developed an ML based Smart PCOS diagnostic system. The architecture compared the performance of various classifiers such as SVM, LR, RF, AdaBoost, DT, KNN, Gradient Boosting, XgBoost, CatBoost, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. They used an open-source dataset and obtained an overall accuracy of 93.25%. Danaei et al. [29] developed an ensemble random forest classifier model and trained features selected by embedded feature selection methods. The model performed with accuracy and sensitivity of 98.89% and 100%, respectively. Bharati et al. [30] proposed an ML model architecture with classifiers such as RF, LR, Hybrid Random Forest- Logistic regression (RFLR), and gradient boosting. RFLR gave the best performance with an accuracy of 91%. Silva et al. [31] proposed a BorutaShap method and sequentially trained a random forest model to identify the most relevant and significant clinical markers. A dataset comprised 72 PCOS patients and 73 healthy women. Fifty-eight features were ranked according to their relevance and significance. The model was able to obtain an overall accuracy of 86%.

3. Materials and Methods

3.1. Dataset Description

In this research, we used an open-source dataset prepared by Kottarathil on Kaggle, having data on 541 fertile women with 43 attributes [25]. This multi-centric data consists of data samples from 10 hospitals in Kerala, India. This dataset has 'PCOS (Yes/No)' as its target variable. Out of 541 patients, 177 were diagnosed with PCOS. The categorical features having Y and N have been encoded into 1 and 0, respectively. Among the 42 features, 24 were non-invasive parameters, and the remaining were either hormonal or gynecological data obtained by invasive vaginal ultrasound and fluid samples. The dataset is described in Table 1.

Table 1. Description of attributes present in our dataset.

Attributes	Meaning	Attributes	Meaning	Attributes	Meaning	Attributes	Meaning
PCOS (Y/N)	Target Variable	Marriage Status (Yrs)	Marital status, if married, the years since marriage	TSH (mIU/L)	Thyroid stimulating hormone	Fast food (Y/N)	Unhealthy eating habits
Age (yrs)	Age of the patient in years	Pregnant (Y/N)	If the woman is pregnant	AMH (ng/mL)	Anti-Mullerian Hormone	Reg.Exercise (Y/N)	Regular Exercise
Weight (Kg)	Weight of the patient in Kgs	No. of abortions	Number of abortions in the lifetime	PRL (ng/mL)	Prolactin levels in blood	BP _Systolic (mmHg)	Systolic Blood pressure is a pressure measured in your arteries when the heart beats

Table 1. *Cont.*

Attributes	Meaning	Attributes	Meaning	Attributes	Meaning	Attributes	Meaning
Height (cm)	Height of the patient in cms	I beta-HCG (mIU/mL)	Case 1: Beta-HCG test; human chorionic gonadotropin (HCG) hormone	Vit D3 (ng/mL)	Vitamin D3 or Cholecalciferol levels	BP _Diastolic (mmHg)	Measurement of artery pressure when heart rests between beats
BMI	Body Mass Index;	II beta-HCG (mIU/mL)	Case 2: Beta-HCG test	PRG (ng/mL)	Serum progesterone levels	Follicle No. (L)	Number of follicles in the left ovary
Blood Group	Blood Group of the patient including A+, A-, B+, B-, O+, O-, AB+, AB-	FSH (mIU/mL)	Follicle Stimulating Hormone	RBS (mg/dL)	Random glucose levels	Follicle No. (R)	Number of follicles in the right ovary
Pulse rate (bpm)	Measure of number of times the heart beats per minute	LH (mIU/mL)	Luteinizing Hormone	Weight gain (Y/N)	Weight gained in the past	Avg. F size (L) (mm)	Average follicle size in the left ovary
RR (breaths/min)	Measure of number of breaths taken per minute	FSH/LH	Ratio of Follicle Stimulating Hormone and Luteinizing Hormone	hair growth (Y/N)	Hair growth	Avg. F size (R) (mm)	Average follicle size in the right ovary
Hb (g/dL)	Haemoglobin range	Hip (inch)	Hip circumference	Skin darkening (Y/N)	Skin discoloration	Endometrium (mm)	Endometrium thickness
Cycle (R/I)	Menstrual cycle; regular or irregular	Waist (inch)	Waist circumference	Hair loss (Y/N)	Balding		
Cycle length (days)	Length of menstrual cycle	Waist:Hip Ratio	Ratio of waist to hip	Pimples (Y/N)	Acne presence		

3.2. Data Pre-Processing

The Kaggle dataset was already pre-processed and required only a few additive steps to make the dataset ready for deployment. The Matplotlib library in Python helped visualize outliers. There were no extreme outliers, and all 541 samples were considered. NumPy helped identify the missing values in the 'Marriage Status (Yrs)', 'Fast food (Y/N)', and 'Marriage Status (Yrs.)' attributes. The median was chosen to replace the missing numerical attributes in the data. 'Fast food (Y/N)' is a categorical variable; the statistical mode of this feature was used for data imputation. We removed attributes 'SI. No' and 'Patient File No' as these had no statistical significance. Descriptive statistical measures of the parameters are tabulated in Table 2.

Table 2. Descriptive statistics of some of the attributes present in the dataset.

Sr. No	Attributes	Mean	Std	Min	25%	50%	75%	Max
1	PCOS (Y/N)	NA	NA	0	NA	NA	NA	1
2	Age (yrs)	31.43068	5.411006	20	28	31	35	48
3	Weight (Kg)	59.63715	11.02829	31	52	59	65	108
4	Height (cm)	156.4848	6.033545	137	152	156	160	180
5	Pulse rate (bpm)	73.24769	4.430285	13	72	72	74	82
6	RR (breaths/min)	19.24399	1.688629	16	18	18	20	28
7	Hb (g/dL)	11.16004	0.866904	8.5	10.5	11	11.7	14.8
8	Cycle (R/I)	NA	NA	2	NA	NA	NA	4
9	Cycle length (days)	4.94085	1.49202	0	4	5	5	12
10	Pregnant (Y/N)	NA	NA	0	NA	NA	NA	1
11	FSH (mIU/mL)	14.60183	217.0221	0.21	3.3	4.85	6.41	5052
12	LH (mIU/mL)	6.469919	86.67326	0.02	1.02	2.3	3.68	2018
13	FSH/LH	6.904917	60.69198	0	1.42	2.17	3.96	1372.83
14	Waist (inch)	33.84104	3.596894	24	32	34	36	47
15	Waist:Hip Ratio	0.891627	0.046135	0.76	0.86	0.89	0.93	0.98
16	AMH (ng/mL)	5.620634	5.876742	0.1	2.01	3.7	6.9	66
17	PRL (ng/mL)	24.3215	14.97039	0.4	14.52	21.92	29.89	128.24
18	PRG (ng/mL)	0.610945	3.808853	0.047	0.25	0.32	0.45	85
19	Weight gain (Y/N)	NA	NA	0	NA	NA	NA	1

Further, we created two sets of frameworks, one with the train-test split ratio of 70:30 and the other with 80:20. This in the upcoming steps would help us analyze the best-performing framework. The split was followed by feature scaling. We deployed the standard scalar algorithm for feature scaling of the training dataset [32]. Standardization has no bounding range, and the data is unaffected by outliers after standardization.

There was an imbalance in the target PCOS (Y/N) class, wherein only 177 out of 541 samples had PCOS. We performed Borderline Synthetic Minority Oversampling Technique (SMOTE) for data balancing on the training dataset to mitigate the potential risk of improper model training. In this method, oversampling creates synthetic points from the minority class. When compared to traditional SMOTE, it solves the issue of misclassified outliers [33]. After Borderline-SMOTE balancing, we obtained 364 and 364 counts for both the PCOS and Non-PCOS classes.

3.3. Feature Selection

This section encompasses an exploration of three different techniques used for feature selection. Three separate frameworks were created for these techniques. We have deployed two Wrapper method algorithms, Harris Hawks and Salp Swarm Optimization, and compared their results with Mutual information. These algorithms helped extract the significant features, reducing the overall data size [34]. We used a 'Feature Selection wrapper class' algorithm from a GitHub toolbox [35].

3.3.1. Harris Hawks Optimization (HHO)

Harris Hawks Optimization is a population-based algorithm that uses stochastic components to find the optimum parameters of a dataset. The algorithm considers the ability of Harris Hawks to form chasing patterns based on the prey's escaping dynamics [36]. This method consists of two phases exploration and exploitation.

In the exploratory phase, the Harris Hawk must monitor a particular area and prey. In HHO, the Harris Hawks can be considered search agents or candidate solutions, out of which the best solutions become near optimum. The exploratory phase constitutes the hawks or search agents monitoring all search space, enhancing the randomness of HHO.The transition from the exploratory to the exploitatory phase is based on the target's energy [37]. The target comprises the optimal solution (Best features). After deploying the HHO optimizer as a feature selection technique, 15 of the most significant features were extracted. Out of which 8 were non-invasive features. The selected features are presented in Table 3.

Table 3. Description of the feature selected by three algorithms.

Sr. No.	Feature Selection Algorithm	Number of Features Selected	Features
1	Harris Hawk Optimization	15	Weight (Kg), Pulse rate (bpm), RR (breaths/min), Hb (g/dL), Pregnant (Y/N), FSH (mIU/mL), LH (mIU/mL), Waist (inch), AMH (ng/mL), PRG(ng/mL), Weight gain (Y/N), hair growth(Y/N), BP _Systolic (mmHg), Follicle No. (R), Avg. F size (L) (mm)
2	Salp Swarm Algorithm	19	Age (yrs), Weight (Kg), Height (cm), Pulse rate (bpm), RR (breaths/min), Cycle (R/I), Cycle length (days), Pregnant (Y/N), FSH (mIU/mL), FSH/LH, Waist (inch), Waist:Hip Ratio, PRG (ng/mL), Weight gain (Y/N), Hair loss (Y/N), Pimples (Y/N), Follicle No. (L), Follicle No. (R), Avg. F size (R) (mm)
3	Mutual Information	15	Follicle No. (L), Follicle No. (R), Skin darkening (Y/N), Fast food (Y/N), hair growth (Y/N), Cycle (R/I), FSH/LH, Cycle length (days), Weight gain (Y/N), AMH (ng/mL), PRL (ng/mL), Pimples (Y/N), BP _Systolic (mmHg), Waist (inch), Age (yrs)

3.3.2. Salp Swarm Optimization (SSA)

Salp swarm optimization algorithm is another bio-inspired wrapper method where the Salps's behavior of swarming, navigating, and ocean foraging behavior is studied to model this optimizer [38]. SSA, provides an improved approach of initial random solutions and the convergence towards the optimum compared to other existing techniques [39]. The mathematical modeling of Salp chains is initiated by dividing the population into two groups, one being the leader and the remaining are followers. The leader would be positioned at the front of the chain to guide the followers. Salps forage a food source, which is the 'target' (final set of features). SSA too has an exploratory and exploitation phase. The position of the followers is a function of the position of the leader and the location of the target.

Salp Swarm Algorithm is known for its swarm intelligence and provides benefits such as robustness, sensitivity, and computational efficiency. Nineteen features were selected with SSA out of which 13 parameters were non- invasive. The selected features are presented in Table 3.

3.3.3. Mutual Information (MI)

The Mutual Information algorithm is a widely known filter method of feature selection. This filter method considers the statistical characteristics of the dataset. Mutual Information is based on the entropy measure; entropy quantifies the uncertainty in the features [40].

Mutual information helps quantify the information shared between two features F and S and is denoted by $MI(F;S)$ is defined by Equation (1).

$$MI(F;S) = \sum_{s \in S} \sum_{f \in F} p(f,s) \log \frac{p(f,s)}{p(f)p(s)} \quad (1)$$

If the Mutual information value is very high F, S are closely related, and if zero they are completely unrelated.

This principle of filtering method of feature selection, when applied to the data, gives how closely related the individual features are to the target [41]. The features are then ranked according to their individual contribution toward the target variable, as seen in Figure 1. Out of these, the top 15 features (10 non-invasive and 5 invasive features) were selected.

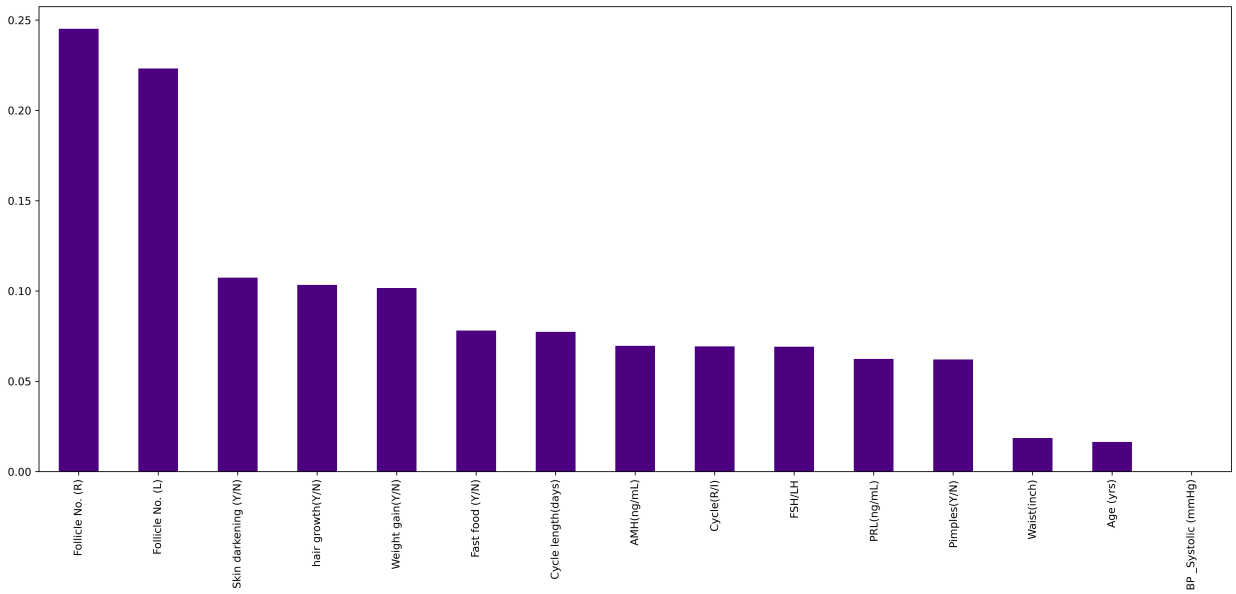

Figure 1. Features ranked from most significant to least significant by Mutual Information Algorithm.

4. Results

4.1. Performance Metrics

We have evaluated and compared our proposed models using standard classification performance measures such as confusion matrix, accuracy, precision, recall, F-1 scores, AUC-ROC score (Area Under the Receiver Operating Characteristics curve), and the precision-recall curve. Our classifiers aim to predict whether a particular patient has PCOS or not.

4.2. Model Evaluation with Machine Learning

This research evaluates and analyzed 12 individual machine learning models: LR, DT, RF, SVM (Kernels- linear, polynomial, gaussian, and sigmoidal), NB, KNN, AdaBoost, XGBoost, and Extratrees. GridSearchCV provided an optimal selection of classifier hyper-parameters. Figure 2 depicts the architecture of the model. After an exhaustive search, examining every possible combination of parameters, this tuning method predicted the optimal hyperparameter, as shown in Table 4.

We took an ensemble learning approach to build a unique and superior model to screen PCOS. The stacking algorithm intends to learn the best combination of the outputs from multiple weak meta-learners for creating the best-performing model. Exploiting this principle, we built three stacks. STACK-1 aggregated seven classifiers: LR, SVM (with linear, polynomial, gaussian, and sigmoidal kernels), NB, and KNN models. STACK-2 represented a tree-based classifier ensemble consisting of DT, RF, AdaBoost, XGBoost, and Extratrees. The multi-level stack of STACK-1 and STACK-2 created STACK-3. Figure 3 pictorially represents the creation of our unique ensemble model.

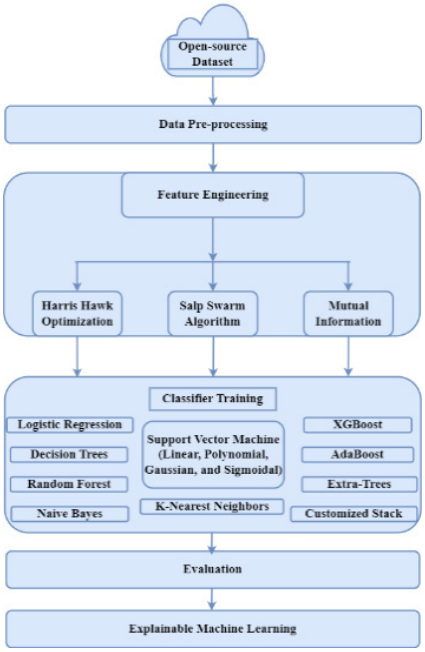

Figure 2. Architecture of the proposed model.

Table 4. Description of the best selected parameters for Machine learning sub-classifiers.

Machine Learning Classifier	Best Parameter Specifications
Logistic Regression	{'C': 0.1, 'penalty': 'l2'}
Decision Tree	{'criterion': 'gini', 'max_depth': 50, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 30, 'splitter': 'best'}
Random Forest	{'bootstrap': True, 'max_depth': 80, 'max_features': 2, 'min_samples_leaf': 3, 'min_samples_split': 10, 'n_estimators': 100}
Support Vector Machine- linear kernel	{decision_function_shape = ovo, gamma = auto, kernel = linear}
SVM-Polynomial kernel	{kernel = poly, max_iter = 200}
SVM-Gaussian kernel	{kernel = rbf, max_iter = 200}
SVM-Sigmoidal kernel	{kernel = sigmoid, max_iter = 1800}
Naïve bayes	{'var_smoothing': 8}
K-Nearest Neighbors	{'n_neighbors': 3}
AdaBoost	{'learning_rate': 0.1, 'n_estimators': 1000}
XGBoost	{'colsample_bytree': 0.3, 'gamma': 0.1, 'learning_rate': 0.05, 'max_depth': 5, 'min_child_weight': 1}
Extratrees	{'min_samples_leaf': 30, 'min_samples_split': 35, 'n_estimators': 50}

A combination of various feature selection techniques, data-split ratios, and meta-learners was considered, the results of which can be seen in Table 5. The recall metric gives insights into the False Negatives of the classifiers. The following were the pipelines producing the best recall: HHO-trained classifiers RF had a 95% score, SSA pipeline with XGBoost gave 97%, and MI with STACK-1 framework had a sensitivity of 100%. Among all architectures, classifiers trained on Mutual information and 80:20 data split had the highest metric scores. Table 6 presents the performance metrics representing models trained on Mutual Information data. It was observed that STACK-3 was either superior or comparable to the individual meta-learners. However, the results for STACK-3 varied with different feature selection techniques. Figure 4 represents the AUC-ROC and PR curves comparison

of STACK-3 trained by HHO, SSA, and MI. The STACK-3 PR scores for HHO, SSA and MI were 0.98, 0.99, and 1, respectively.

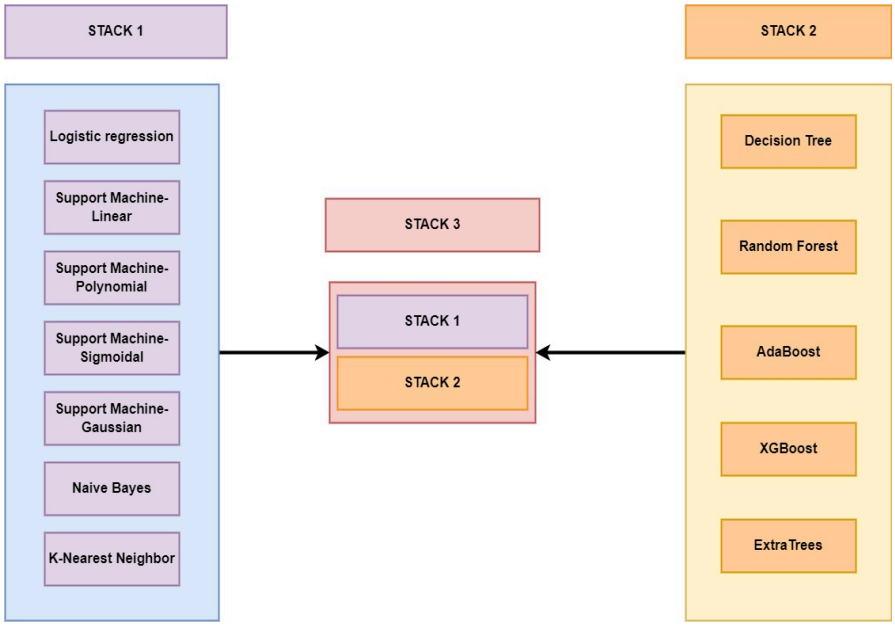

Figure 3. Pictorial representation of the development of customized stacks.

Table 5. Comparison of the accuracies of 15 Machine learning models with the of feature engineering techniques and data split ratios.

Model \ Feature Selection	Harris Hawk Optimization		Salp Swarm Algorithm		Mutual Information	
	70:30	80:20	70:30	80:20	70:30	80:20
Logistic Regression	0.89	0.89	0.87	0.89	0.95	0.93
Decision Trees	0.87	0.84	0.82	0.89	0.84	0.89
Random Forest	0.93	0.97	0.93	0.96	0.96	0.97
Support Vector Machine Linear	0.85	0.93	0.88	0.93	0.94	0.94
Support Vector Machine Polynomial	0.86	0.91	0.93	0.92	0.94	0.95
Support Vector Machine Gaussian	0.88	0.88	0.93	0.92	0.95	0.97
Support Vector Machine Sigmoid	0.87	0.92	0.85	0.9	0.91	0.94
Naïve Bayes	0.79	0.83	0.82	0.57	0.89	0.91
K-Nearest Neighbors	0.88	0.9	0.88	0.86	0.93	0.95
AdaBoost	0.91	0.95	0.92	0.95	0.93	0.95
XGBoost	0.9	0.95	0.94	0.97	0.94	0.96
ExtraTrees	0.83	0.89	0.88	0.9	0.91	0.94
STACK-1	0.89	0.9	0.93	0.89	0.95	0.98
STACK-2	0.92	0.93	0.92	0.98	0.95	0.97
STACK-3	0.92	0.92	0.94	0.95	0.95	0.98

Table 6. Performance of the models trained on 80:20 split Mutual information engineered data.

Model MI	Precision	Recall	F1-Score	Accuracy	AUC	Precision-Recall Score
Logistic Regression	0.92	0.92	0.92	0.93	0.99	0.98
Decision Trees	0.91	0.83	0.87	0.89	0.96	0.94
Random Forest	0.97	0.97	0.97	0.97	1.00	1.00
Support Vector Machine Linear	0.92	0.94	0.93	0.94	0.98	0.98
Support Vector Machine Polynomial	0.94	0.95	0.95	0.95	0.98	0.93
Support Vector Machine Gaussian	0.93	1.00	0.96	0.97	1.00	0.99
Support Vector Machine Sigmoid	0.92	0.94	0.93	0.94	0.98	0.98
Naïve Bayes	0.91	0.98	0.91	0.91	0.98	0.98
KNN	0.91	0.98	0.95	0.95	0.98	0.97
AdaBoost	0.91	0.98	0.95	0.95	0.99	0.99
XGBoost	0.93	0.98	0.95	0.96	1.00	1.00
ExtraTrees	0.91	0.95	0.95	0.94	0.99	0.99
STACK-1	0.96	1.00	0.98	0.98	0.99	0.99
STACK-2	0.95	0.98	0.97	0.97	1.00	1.00
STACK-3	0.97	0.98	0.98	0.98	1.00	1.00

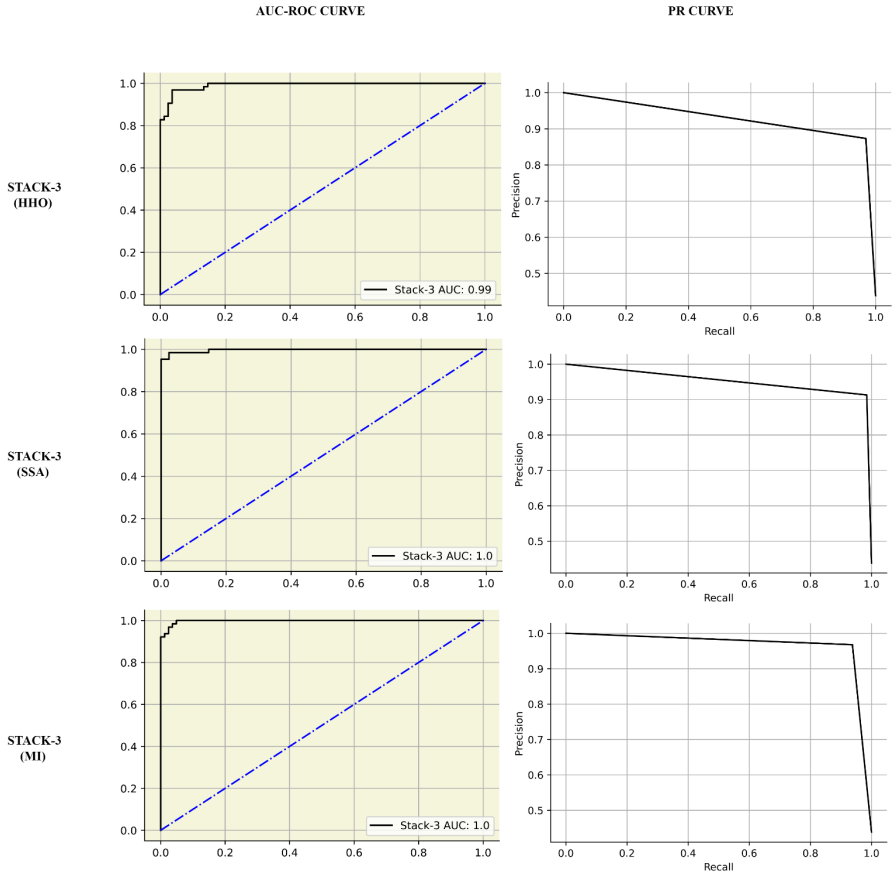

Figure 4. Performance evaluation of STACK-3 by AUC Curves and Precision Recall curves of STACK-3 (multilevel-stacking) for feature selection. The first, second and third rows indicate the performance of STACK-3 with HHO, SSA and MI respectively.

4.3. Model Evaluation using Deep Learning

This study evaluates two Deep Learning models, Deep Neural Network (DNN) and Convolutional Neural Network (CNN). ML uses simpler principles for creating predictive

models and can perform a task without explicit programming. In contrast, DL uses complex algorithms that mimic the human brain's capability to learn and train. DL is based on neural networks that detect patterns in structured and unstructured datasets. Deep learning models help avoid an extra step of feature selection [42]. We aim to compare the performance of these state-of-the-art models against the results obtained with ML models trained on three different engineered datasets.

4.3.1. Deep Neural Network (DNN)

We created a customized DNN model with five hidden layers. The Rectified Linear Activation function (ReLU) activated the sensory layer and the associator layers. The sigmoid activation function was used in the output layer. Details of the DNN design can be seen in Table 7. We used Adaptive Moment Estimation (Adam) optimizer, this uses a stochastic gradient descent approach to optimize and train DL models [43]. We chose binary cross-entropy as our loss function. This loss function calculates the cross entropy between the true and predicted classes, producing a binary output of 0 or 1. The customized DNN model achieved a training accuracy of 82%. The test data was then fed to the model, and an overall prediction accuracy of 93.85% was obtained. The plots of model accuracy and model loss for both the training and testing data were obtained, as shown in Figure 5.

Table 7. Deep Learning models' comparison with the proposed model.

Models	Accuracy	Precision	Recall	F1-Score	Optimizer	Batch Size	Epochs	Network Description			
								Layer No.	**Role**	**Activation Function**	**Number of Nodes**
DNN	0.94	0.89	0.98	0.93	Adam	26	200	Layer 1	Input layer	ReLU	19
								Layer 2	Hidden layer 1	ReLU	12
								Layer 3	Hidden layer 2	ReLU	9
								Layer 4	Hidden layer 3	ReLU	7
								Layer 5	Hidden layer 4	ReLU	5
								Layer 6	Hidden layer 5	ReLU	3
								Layer 7	Output layer	Sigmoidal	1
								Layer No.	**Role**	**Activation Function**	**Number of Filters/ Units**
1-D CNN	0.90	0.86	0.84	0.85	Adam	10	200	Layer 1	Layer Conv 1D -1 (Input)	LeakyReLU	32 filters
								Layer 2	Layer Conv 1D-2	LeakyReLU	64 filters
								Layer 3	Layer Conv 1D-3	LeakyReLU	128 filters
								Layer 4	Max Pooling	NA	NA
								Layer 5	Dropout	NA	NA
								Layer 6	Flatten the output	NA	NA
								Layer 7	Dense layer 1	LeakyReLU	256 units
								Layer 8	Dense layer 2	LeakyReLU	512 units
								Layer 9	Dense layer-3 (output)	Sigmoidal	1 unit
STACK-3 (MI)	0.98	0.97	0.98	0.98	GridSearchCV	80:20 train-test split	-	NA			

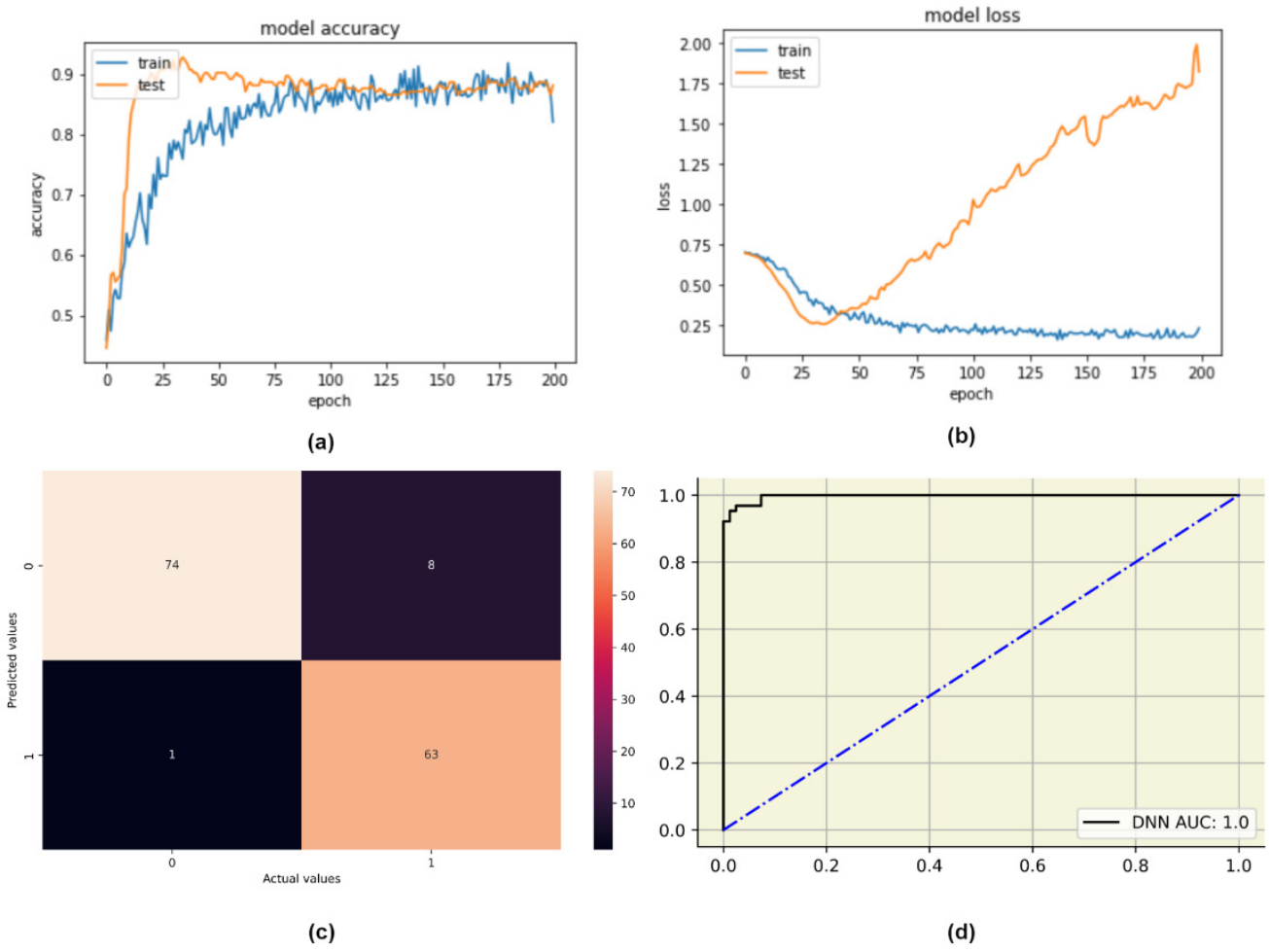

Figure 5. DNN evaluation plots: (**a**) Change in model accuracy with the increase in epochs, (**b**) Change in model loss with the increase in epochs, (**c**) Confusion matrix, (**d**) AUC-ROC plot for DNN.

4.3.2. 1-Dimensional Convolutional Neural Network (1D-CNN)

Convolutional Neural Networks are widely used to create architectures for solving computer vision problems. CNN kernels can carry out feature extraction in images because of two primary principles: local connectivity and spatial locality [44]. When considering a tabular dataset, 1-D CNN is used. However, the CNN kernel expects the dataset's features to have spatial locality correlation, which may not be observed in all datasets. Hence, a basic principle of reordering the features is used for the spatial correlation to be maintained. However, CNN is not the primary choice when processing tabular data as it has no locality characteristics. In the 1-D CNN model, a more extensive set of locally connected features is created by a fully connected layer followed by subsequent layers [45].

The details of the customized 1-D CNN model are given in Table 7. 1-D CNN achieved a training accuracy of 97%. Figure 6 represents the model accuracy and loss plot, followed by the confusion matrix and the ROC curve for the customized 1-D CNN model. On performance evaluation, our 1D-CNN model achieved accuracy, precision, recall, and f1-score of 90%, 86%, 84%, and 85%, respectively. Further, the AUC-ROC score obtained was 1.

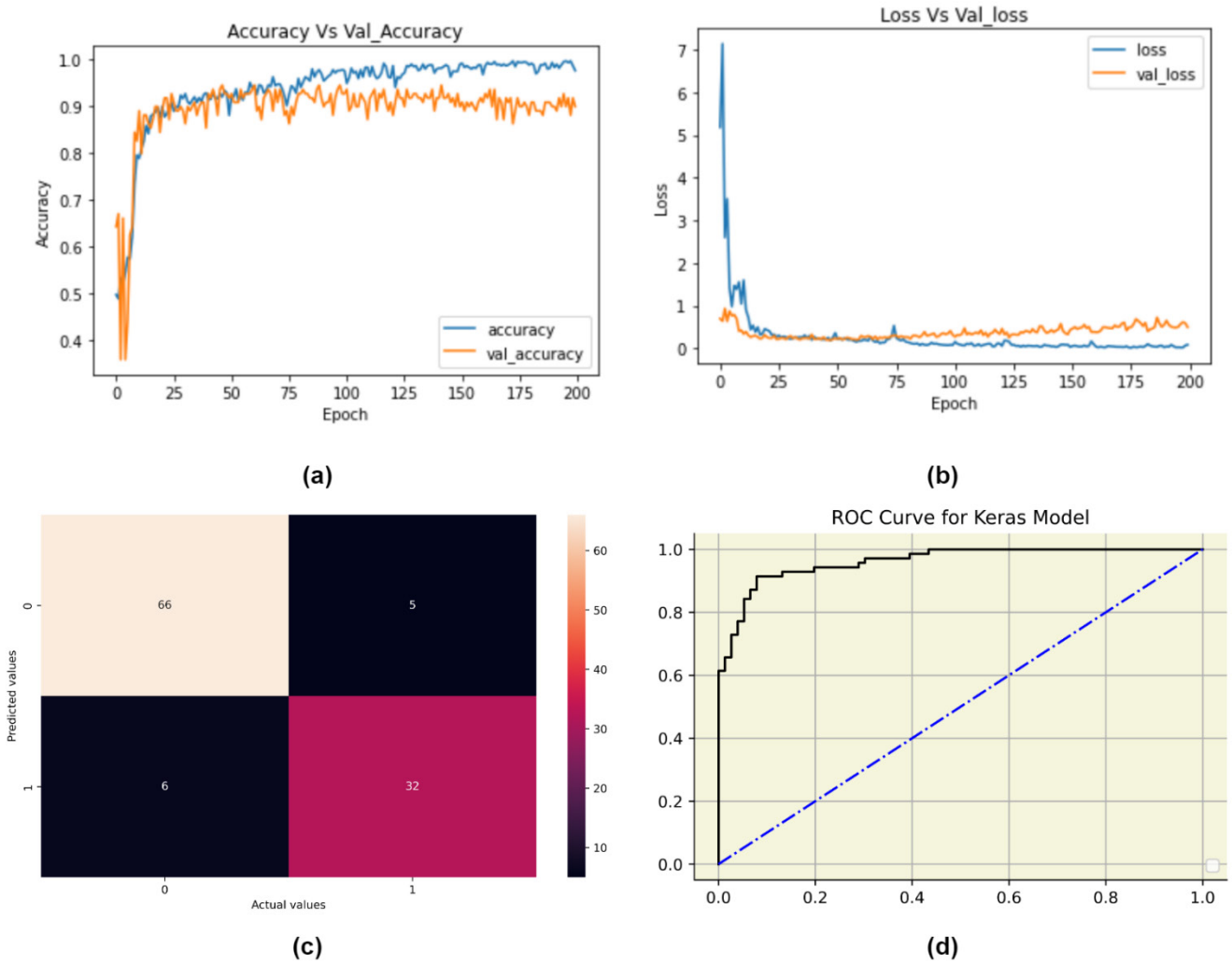

Figure 6. 1D-CNN evaluation plots: (**a**) Change in model accuracy with the increase in epochs, (**b**) Change in model loss with the increase in epochs, (**c**) Confusion Matrix of CNN model and (**d**) ROC curve for the CNN model.

4.4. Explainable AI (XAI)

This section explores various XAI techniques used to explain predictions and debugging classifiers. Due to incompatibility issues with the sklearn, StackingClassifier function, the XAI tools were unable to explain our best-performing pipeline. However, tree-based meta-classifiers performed comparable to the multi-level stack, and this section helps interpret these classifiers. The obtained visualizations have assisted in providing probabilities of the patient being predicted PCOS positive to make classifier predictions meaningful and understandable.

4.4.1. Shapley Additive exPlanations (SHAP)

Lundberg et al. [46] proposed a "Unified Approach" SHAP architecture where each feature gets assigned an 'importance value'. SHAP quantifies the contributions of each feature in the model prediction, however, does not evaluate the quality of the decisions made by the classifiers. Nevertheless, Important insights into the classifier predictions makes them, meaningful [47]. In this study, the classifiers trained after mutual information performed the best. XGBoost was among the best performing ML meta-learners with an accuracy and recall of 96% and 98%, respectively. We used this model to demystify its predictions using SHAP values.

SHAP Violin Plot

A SHAP Violin plot was created for one of our best-performing tree-based model, XGBoost. With tree-based models built on principles of ensemble modeling, it can become impossible to know the model-rationale in decision-making [48]. As shown in Figure 7, the horizontal axis represents the SHAP values, and the color depicts the higher or lower value of the data points. The features are arranged in the order of their importance (The best feature is present on the top). Bright red indicates a higher value, and bright blue indicates a lower feature value. These violin plots help spot outliers and produce a more accurate representation of densities than kernel density estimated from only a few points [49]. In this violin plot, it is evident that the most crucial feature, Follicle number (number of follicles in the right ovary), a higher number of these follicles contributes positively to the PCOS prediction. It is revealed that PCOS prediction of XGBoost is based on the patients experiencing hair growth, weight gain, skin darkening, an irregular menstrual cycle, and observance of more follicles in both left and right ovaries. As observed, this Violin plot agrees with the classic symptoms of PCOS [1].

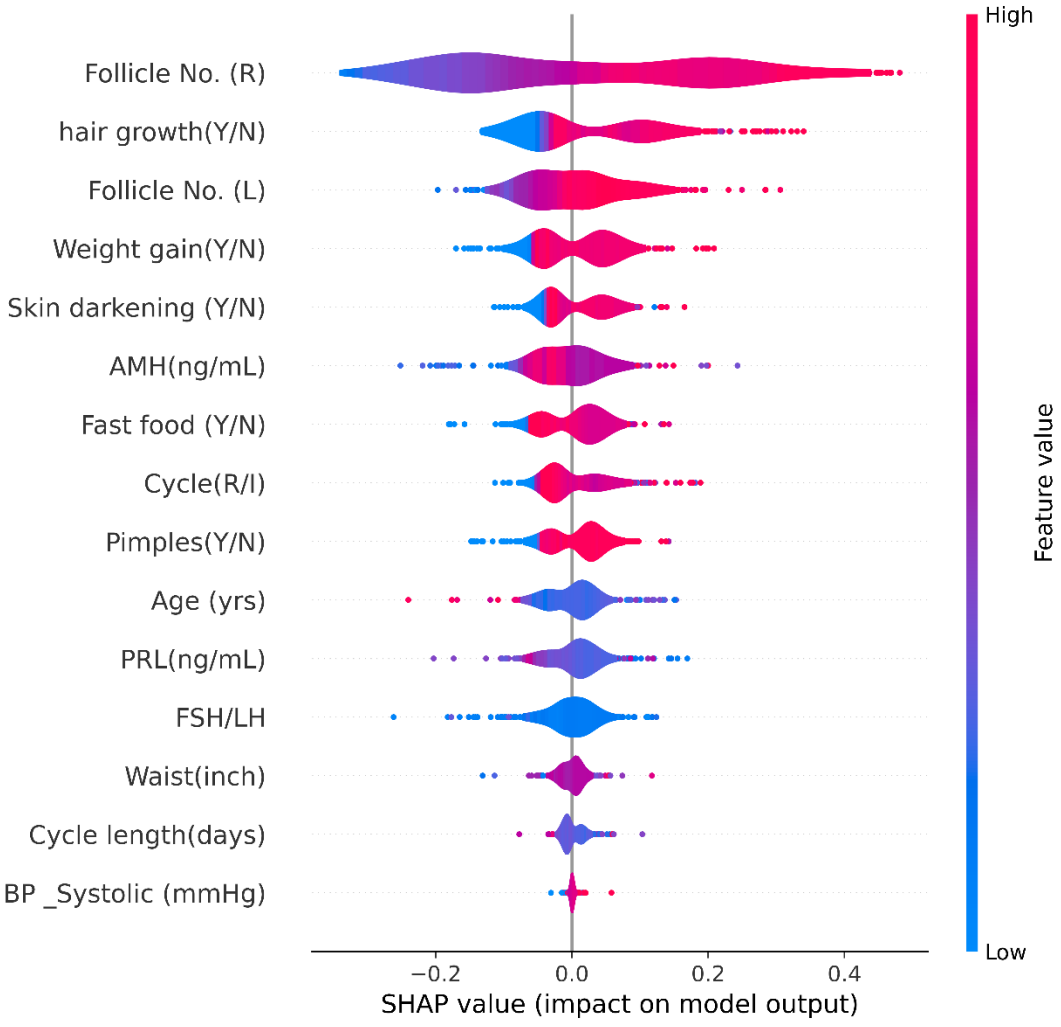

Figure 7. SHAP Violin Plot.

SHAP Waterfall Plot

A SHAP Waterfall plot can be observed in Figure 8, where the horizontal axis represents the expected values of the classifier output and the shifts toward the right or left signify the positive or negative contribution of individual features [50]. In Figure 8, random patient data was taken for analysis and the dataset showed that this patient has been diag-

nosed with PCOS. The feature values of this data point are in grey along the *y*-axis and the features are arranged from top to bottom, from the most to the least significant. Here, *f(x)* and *E[(f(x)]* denote the prediction and expected values, respectively. In this SHAP waterfall plot, the selected patient has 11 and 9 follicles in their left and right ovaries, respectively. This indicates a higher-than-normal follicular number, and these two features have the most significance in PCOS detection. This patient experiences pimples and skin darkening along with an increase in PRL levels. These are an indication of the presence of PCOS [2]. With a waterfall plot it is easy to visualize the shifts of values from a prior expectation to the final prediction

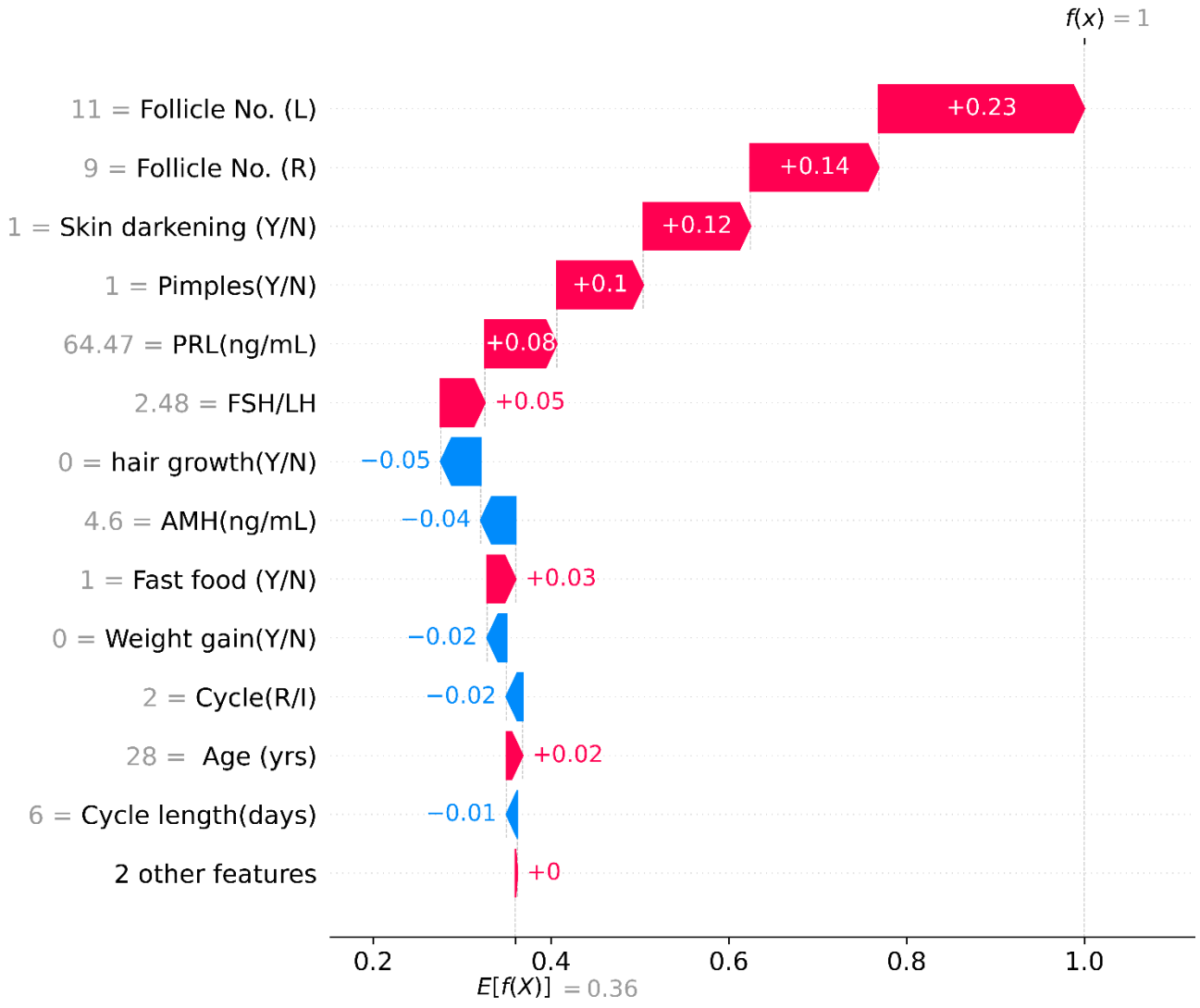

Figure 8. SHAP Waterfall Plot.

SHAP Force Plot

The SHAP Force plot helps the user identify the most significant features during prediction for a single observation. Higher scores, in red, lead to the prediction being 1 (PCOS-positive), and lower scores, in blue, lead to 0 (PCOS-negative). Features having a higher impact on the prediction are positioned near the dividing boundary, here Follicle No. (R) is the most significant [51]. The size of the bar quantifies the impact of the feature. Figure 9 indicates the SHAP force plot for a randomly selected patient sample. In this plot, the XGBoost prediction is explained by the right and left follicular numbers being normal and the patient not showing signs of weight gain, excessive hair growth, and acne.

Multiple invasive and non-invasive parameters have been analyzed for XGBoost to predict the patient as PCOS-negative.

Figure 9. SHAP Force plot.

SHAP Dependance Plot

A dependence plot is a scatter plot representing how a model prediction depends on a feature. Each point on the plot represents one patient. x-axis represents the feature value, and y-axis represents the SHAP values. The color corresponds to the second feature. Vertical patterns indicate interactions between the features. Figure 10. represents the SHAP dependence plot of the feature 'Follicle No. (R)'. Four plots depict the interaction between 'Follicle No. (R)' and non-invasive features such as weight gain, hair growth, skin darkening, and pimples. In all plots, it is observed that with a higher than eight follicular number, the patient is more likely to develop these external PCOS manifestations. SHAP's correlation does not imply causation. However, PCOS Hormonal imbalance manifests as non-invasive parameters such as obesity, hyperpigmentation, and hirsutism [52]. SHAP dependence plot help gain insights into the feature interactions and their effect on the prediction.

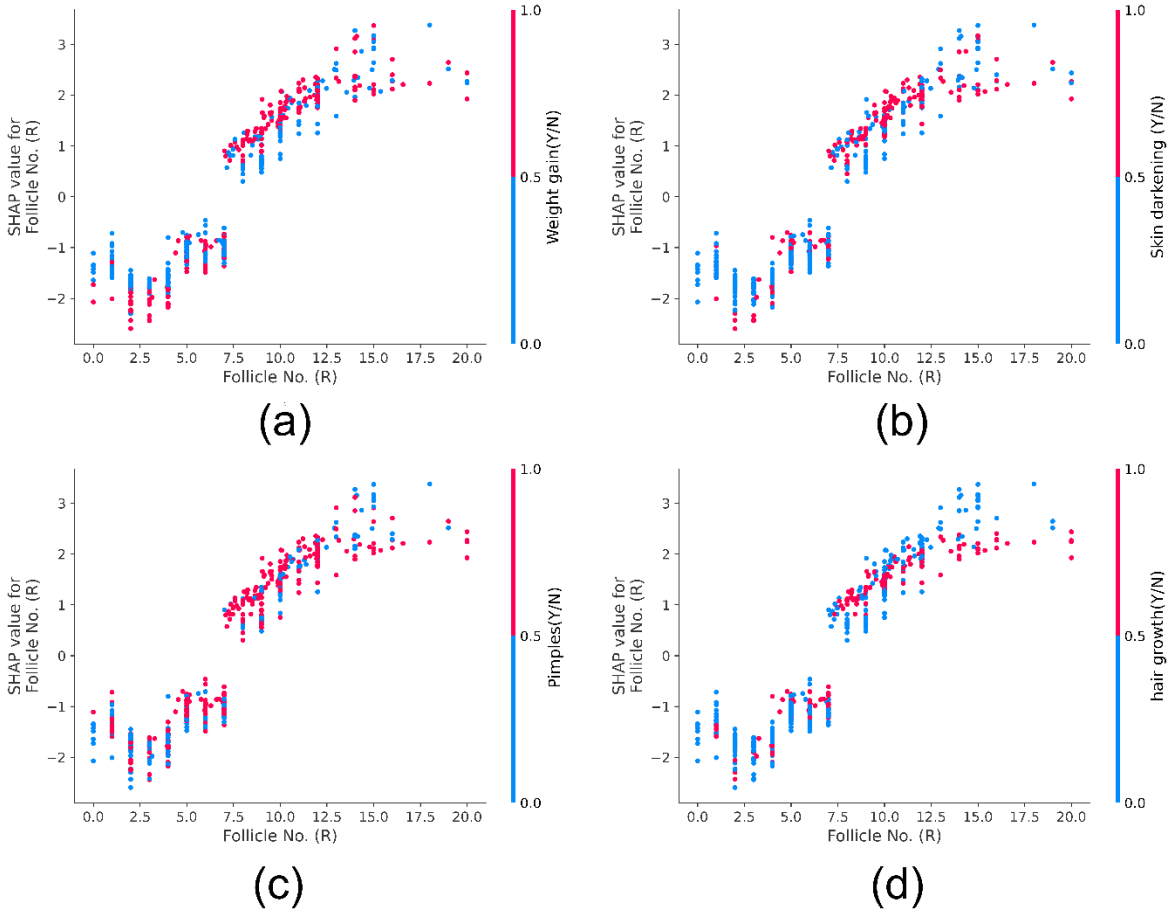

Figure 10. SHAP Dependence plot for Follicle No. (R). Sub-figure (**a**) indicates the plot of Follicle No. (R) against Weight gain (Y/N), (**b**) indicates the plot of Follicle No. (R) against Skin Darkening (Y/N), (**c**) indicates the plot of Follicle No. (R) against Pimples (Y/N) and (**d**) indicates the plot of Follicle No. (R) against hair growth (Y/N).

4.4.2. Local Interpretable Model-Agnostic Explanations (LIME)

Ribeiro et al. [53] introduced a model-agnostic tool for the interpretability of all supervised machine-learning models. In this study, LIME used a random forest classifier trained after mutual information. RF had an accuracy and precision of 97% and 95%, respecively. LIME, like SHAP, arranges the features according to their significance and indicates the probabilities of certain RF predictions.

In Figure 11, two plots for LIME are given. In the first plot, Random Forest has classified the selected a PCOS-positive prediction with a 0.98 probability. LIME explains this RF prediction with various insights into the patient's condition of excessive hair growth, weight gain, skin darkening, pimples, and having more than nine follicles in the left ovary. The second plot depicts a patient with a 0.98 probability of being predicted PCOS negative, explaining that the most classic symptoms of PCOS are absent. No signs of Skin darkening or unwanted hair growth with a normal range of follicles in ovaries. These results agree with other PCOS medical literature [4,5].

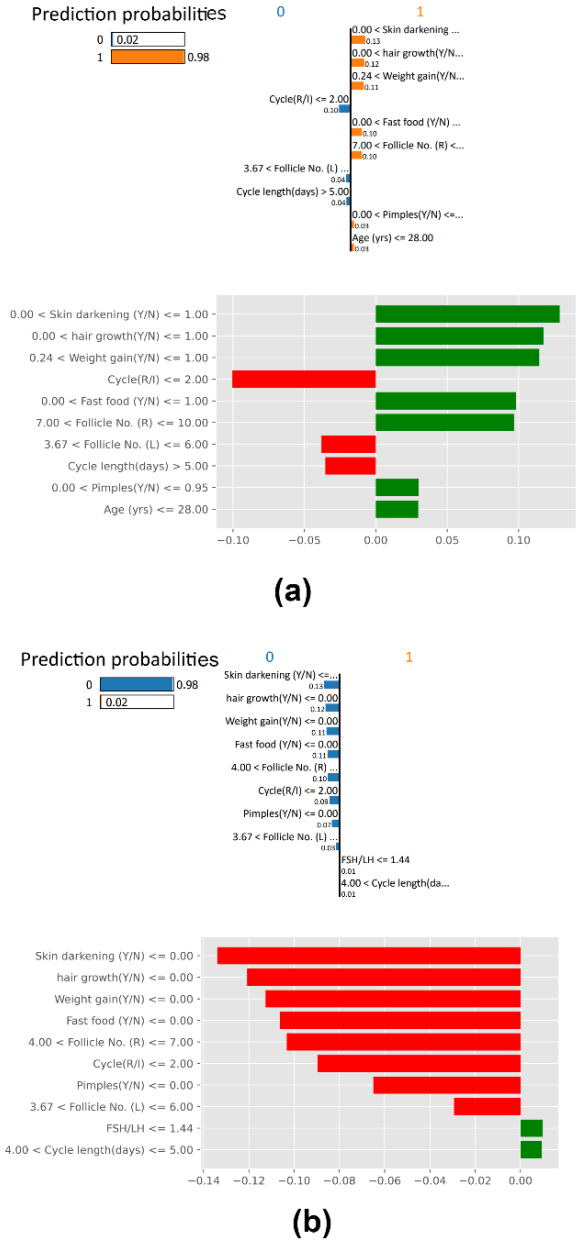

Figure 11. LIME Plot (**a**) PCOS-positive case, (**b**) PCOS-negative case.

Among most XAI models available, LIME is very intuitive and produces simple plots for local interpretability, revealing how a prediction is made.

4.4.3. ELI5

ELI5 is a python package for inspecting and interpreting ML classifiers. With tree-based models, ELI5 uses the gini index for preparing decision trees by weights [54]. Figure 12a shows the tabulated weights calculated for each parameter. The features are ordered and assigned weights based on their importance (with the most critical parameter being on top). A multi-node decision tree map is created based on this table. ELI5 gives a transparent and in-depth visualization of how a decision tree classifier made a prediction, based on the nodes. Here, Follicle no. (R) is the root node, followed by further splitting based on the conditional nodes at each tree level to get the eventual classification at the leaf nodes. ELI5, like LIME, provides local interpretations for each observation, as seen in Figure 12b. The selected patient in the plot (b) is classified as PCOS positive with the explanation given by individual feature contributions to the output. The patient faced unwanted hair growth and pimples and had a significantly high number of follicles in the left ovary. ELI5 assists in debugging ML models, providing insights on permutation and feature importance.

Weight	Feature
0.4676	x1
0.1343	x4
0.0869	x8
0.0581	x0
0.0515	x6
0.0447	x9
0.0356	x5
0.0274	x7
0.0217	x10
0.0153	x11
0.0127	x12
0.0124	x14
0.0118	x13
0.0115	x2
0.0084	x3

(a)

y=1.0 (probability **1.000**) top features

Contribution?	Feature	Value
+0.521	<BIAS>	1.000
+0.348	hair growth(Y/N)	1.000
+0.174	Follicle No. (L)	11.197
+0.157	Pimples(Y/N)	1.000
+0.065	Fast food (Y/N)	1.000
+0.038	Waist(inch)	37.885
-0.304	Follicle No. (R)	6.459

(b)

Figure 12. ELI5 plots: (**a**) Global prediction: Depicts the weights calculated for the features based on GINI index (x0, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14 correspond to 'Follicle No. (L)', 'Follicle No. (R)', 'Skin darkening (Y/N)', 'Fast food (Y/N)', 'hair growth(Y/N)', 'Cycle (R/I)', 'FSH/LH', 'Cycle length(days)', 'Weight gain(Y/N)', 'AMH(ng/mL)', 'PRL(ng/mL)', 'Pimples(Y/N)', 'BP_Systolic (mmHg)', 'Waist(inch)', ' Age (yrs)'. (**b**) Local prediction, table of the most critical features in the classifier prediction, the prediction is 1 (the patient has PCOS).

4.4.4. QLattice

Qlattice is a supervised ML method, created by Abzu, inspired by Richard Feynman's path intergral equation [55]. Qlattice is tool that automatically generates a predictive model

and gives insights on its interpretability and explainability even with less data samples [56]. Figure 13 depicts the Qgraph for the created Qlattice model, this connects all the inputs to the outputs. In these models weights are initially randomly assigned however with the constant training, optimization is performed to minize the loss function. The QGraph here takes the most significant features as follicle no. (R), weight gain (Y/N) and hair growth (Y/N) as its inputs. The white boxes represent the interactions, that build a function by taking inputs and predict the PCOS outcome. Equation (2), represents a general expresion created for the model. This XAI model agrees with the medical symptoms of PCOS, when selecting its critical parameters for detecting PCOS [5].

$$logreg(0.54\ (Follicle\ No.\ (R)) + 1.85\ (Weight\ gain\ (Y/N)) \\ +2.37\ (hair\ growth\ (Y/N)) - 6.38) \tag{2}$$

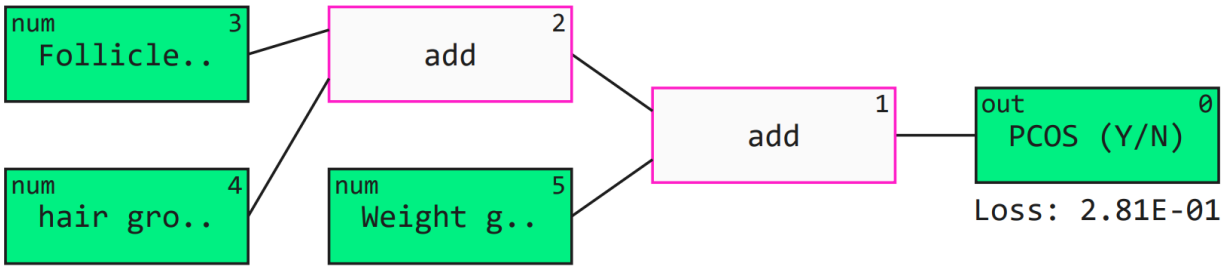

Figure 13. QGraph to interpret the results obtained by the trained models.

4.4.5. Feature Importance with Random Forest

Random forest is an ensemble of decision trees and each internal node of the tree makes decisions on how the dataset gets divided [57]. In our feature importance algorithm with RF, gini impurity was calculated, based on which measurements are made to understand how well each feature decreases the impurity of the split. We used the scikit-learn implementation of RF. Figure 14 represents a graph with features ordered in descending order from left to right according to their importance value, wherein the number of follicles in the right and left ovaries contribute significantly to the PCOS prediction by the RF classifier. Further, Skin darkening, excessive hair growth, weight gain, abnormal AMH levels and FSH/LH values can lead to a PCOS-positive prediction.

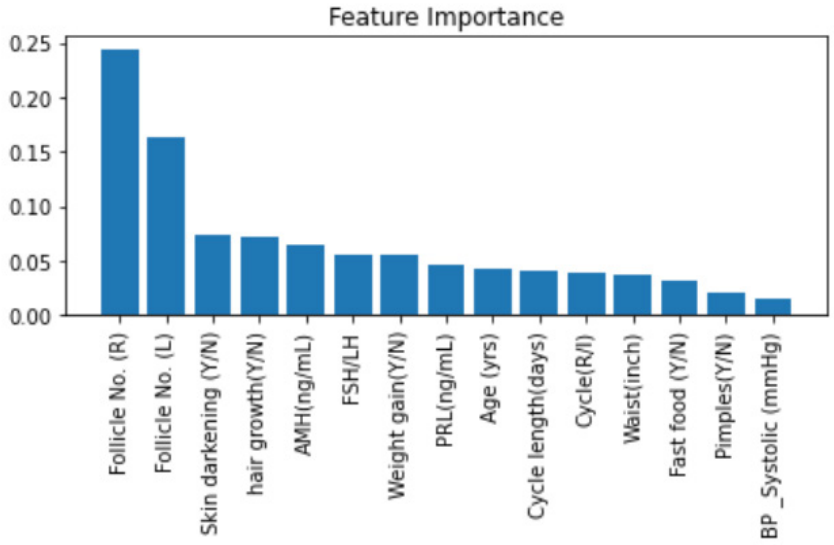

Figure 14. Feature Importance with Random Forest.

5. Discussion

Polycystic Ovary Syndrome (PCOS) is a hormonal disorder that causes ovary enlargement and symptoms such as obesity, acne, hirsutism, oligomenorrhea, acanthosis nigricans, and male-pattern baldness. It is often underdiagnosed or misdiagnosed, leading to the risk of severe health conditions such as type 2 diabetes, hypertension, cardiovascular disorders, infertility, and uterine and endometrial cancer. We created multiple ML and DL pipelines in this study and evaluated their performance to find the best-performing classifier.

The ML frameworks were created by considering all combinations of two data split ratios, three feature extraction methods, and 10 classifiers. Models trained with 80% (80:20 split) of data samples performed significantly better than the ones trained on 70% of samples, regardless of the feature selection technique. As seen in Table 5, the impact of various feature selection techniques gives insights into the training of the classifiers. Interestingly, all tree-based classifiers such as DT, RF, AdaBoost, XGBoost, and Extra-trees had comparable performances irrespective of the features selected. However, Classifiers trained on MI-extracted features achieved the highest performance metrics among other deployed feature selection techniques. These models had higher sensitivity, assuring a significant reduction in false negatives.

Further, the accuracies of most models trained on the HHO and SSA methods provided a similar performance, except for the NB classifier, whose accuracy significantly worsened with SSA data. Stack models, a blend of individual meta-learners based on how they get trained, would define how they work in an ensemble. This can be observed in the STACK-3 performance under each feature selection technique. HHO features trained the poorest STACK-3 classifier with an accuracy of 92%, and MI features trained the best stack achieving 98% accuracy. Figure 4 depicts the improvement in the AUC-ROC score and PR curve for STACK-3 from HHO to MI-trained stack.

Different feature-selection pipelines had different 'best-performing' classifiers. With HHO, the stochastic-based method for feature selection, XGBoost outperformed all other models with prediction accuracy, recall, and AUC score of 97%, 100%, and 99%, respectively. When considering the model performance on data feature engineering with SSA, STACK-2 achieved the best accuracy, recall, and AUC score of 98%, 98%, and 99%, respectively. Classifiers trained on MI-extracted features achieved the highest performance metrics among other deployed feature selection techniques. Out of all combinations of classifiers and feature extraction techniques, STACK-3 trained with an 80:20 split on MI-engineered data had the highest performance, achieving accuracy, precision, recall, f1-score, AUC score, and Precision-Recall scores of 98%, 97%, 98%, 98%, 1, and 1, respectively. As indicated in Table 6, it should be noted meta-learners like RF, SVM-RBF, and XGBoost trained on an 80:20 split on MI-engineered data have performed significantly better than other models, with close to 100% sensitivity. These meta-learners could be the ideal choice when designing individual PCOS classifiers. The ML-based pipelines were compared with customized DL architectures. DNN obtaining accuracy and recall of 94% and 98%, respectively. 1-D CNN performed comparatively poor with an accuracy and recall of 90% and 94%. Both the tree-based meta-learner (MI) pipelines such as RF, XGBoost and STACK-3 have outperformed these complex deep learning architectures. Our proposed best-performing multi-level stack drastically lowered the number of false negatives PCOS-predictions (with only 3 out of the processed 146 samples).

Further, we added a layer of Explainable machine learning to the high-performing tree-based meta-learner frameworks. We explored multiple SHAP with visualizations to explain global and local XGBoost predictions based on feature importance. LIME created an intuitive and interpretable probability-based visual for local predictions of Random Forest. ELI5 ranked and assigned weights based on feature importance and created a tree map for debugging the decision tree classifier. Further, we deployed a Qlattice algorithm that automatically created a predictive model along with a Qgraph to explain the classifier. Feature Importance with Random Forest helped get insights into global predictions made by RF. Notably most tools explained the predictions based having the following features

with highest significance: Follicle No. (L), Follicle No. (R), Skin darkening (Y/N), Weight gain (Y/N), hair growth(Y/N), Pimples (Y/N). Research has indicated the significance of these features in diagnosis [4]. XAI tools hence help visualize and interpret local predictions just as practitioners would consider during primary detection of PCOS.

Multiple articles proposed machine learning models for screening PCOS using the same open-source Kaggle dataset used in this study. Neto et al. [58] published a study for detecting PCOS patients with data mining tools. Classifiers such as SVM, MLP, RF, LR, and NB, along with Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, were adopted. RF, along with data sampling techniques, performed the best with accuracy, sensitivity, and precision of 95%, 94%, and 96%, respectively. Nandipati et al. [59] This study compared Python (Spyder IDE) and Rapid Miner for diagnosis. Spyder IDE used a correlation matrix and recursive feature elimination with a logistic regression model, while RapidMiner used forward selection (with NB) and backward elimination (with DT) with a cross-validation operator. The engineered data trained seven ML models, KNN, SVM, RF, NB, Auto-MLP, AdaBoost, and Bagging with DT. Ten features were selected, with RapidMiner obtaining the highest average accuracies of the ML models at 85.97%. Vedpathak et al. [60] proposed an ML PCOS detection architecture named 'PCOcare'. Out of the 42 features, 30 were selected by the Chi-square method. They deployed RF, SVM (with Linear kernel and radial basis function kernel), NB, LR, and KNN. RF performed the best with accuracy, sensitivity, and precision metric scores at 91%, 94%, and 88%. Hdaib et al. [61] trained multiple classifiers, KNN, NN, NB, SVM, LR, classification tree, and Linear discriminant (LD). After feature engineering, they selected 43 features, and LD performed the best with 92.60% accuracy. ÇİÇEK et al. [62] proposed a LIME PCOS diagnostic model, which used RF as its classifier. A chi-square test was deployed, and features with a p-value greater than 0.05 were considered statistically significant. RF obtained an accuracy of 86.03%. LIME described local predictions for the first five patients. We have tabulated the remaining model performance comparison in Table 8.

Table 8. Comparison of existing models with the proposed study for PCOS diagnosis.

Sr. No.	Model	Classifier	Feature Section	Accuracy	Sensitivity	Precision	F1-Score	AUC Score	Explainable AI Techniques
1	[24]	RF, XGBoost, MLP, SVM	Pearson correlation	93%	_	_	_	0.962	_
2	[26]	CatBoost	K-fold validation	82.5% (invasive) and 90.1% (non-invasive parameters)	_	_	_	_	_
3	[30]	RF, LR, Hybrid RF and Logistic regression (RFLR) and gradient boosting	_	91.01%	90%	_	_	_	_
4	[28]	SVM, LR, RF, AdaBoost, DT, KNN, Gradient Boosting, XgBoost, CatBoost, Linear Discriminant Analysis, Quadratic Discriminant Analysis	Pearson correlation	93.25%	92.68%	98.28%	0.954	_	_

Table 8. *Cont.*

Sr. No.	Model	Classifier	Feature Section	Accuracy	Sensitivity	Precision	F1-Score	AUC Score	Explainable AI Techniques
5	[29]	Ensemble RF, MLP, AdaBoost and Extra tree	Filter, embedded and wrapper feature selection	98.89%	100%	98.30%	_	_	_
6	[27]	CatBoost, voting hard, voting soft	K-fold method (13 features)	91.12%	_	_	_	0.92	_
7	Proposed model	LR, DT, RF, SVM (Linear, gaussian, polynomial, sigmoidal), NB, KNN, AdaBoost, XGBoost and stacking models	SSO, HHO and MI	98%	98%	97%	0.98	1	SHAP (Local and global interpretation) on XGBoost, LIME on RF, ELI5 and Qlattice

Multiple studies mentioned above have RF as the best-performing classifier. However, our RF meta-learner pipeline outperforms these architectures with accuracy, precision, and recall of 97%, 97%, and 97%, respectively. Overall, the proposed customized multilevel stack is superior to most studies conducted on this dataset. Our ensemble, where various models work towards the same problem of screening PCOS among fertile women, provides improved performance, reduction in false negatives, and reliability compared to all other models. At the time of writing this manuscript, no other authors had used SHAP, ELI5, Qlattice and feature importance with Random Forest. on this dataset. Rather than a mystified 'black box', we believe a meaningful and interpretable model would be of more practical use in clinical settings. With this motivation, we created this architecture to contribute towards de-mystifying artificial intelligence in healthcare.

6. Conclusions and Future Scope

Polycystic Ovary Syndrome is a complex endocrine disorder that affects women in their reproductive years. About 8 to 20% of Indian women suffer from PCOS, and if undiagnosed or misdiagnosed, women with this syndrome become more likely to develop infertility issues, cardiovascular disorders, type 2 diabetes, and uterine and ovarian cancer. In this study, we aimed to screen PCOS among fertile women. This study evaluates multiple frameworks and proposes a meta-learner-based multi-level stack ML classifier trained on MI-engineered data to be the best-performing pipeline. This pipeline proved superior against state-of-the-art neural networks such as DNN and 1D-CNN. Further, we deployed a layer of explainability and transparency to PCOS screening with tools like SHAP, LIME, ELI5 (for the best-performing tree-based classifiers), Qlattice and Feature Importance with Random Forest.

A user interface could be built to deploy PCOS screening in real-time, and this framework could be scaled to predict PCOS among a larger population. Non-invasive features could be used to create a patient-centric application to provide risk analysis to female patients susceptible to PCOS and link them to their Electronic Health Record for medical evaluation. Figure 15 represents the future scope of this study. However, various external validations and rigorous testing and scalability assesment should be conducted prior to the deployment of this framework in medical facilities. The dataset should be exhaustive and of high quality. It is critical to bridge the gap between medical and informatics professionals by ensuring the model is as meaningful to the healthcare personnel as it is to a Machine Learning expert.

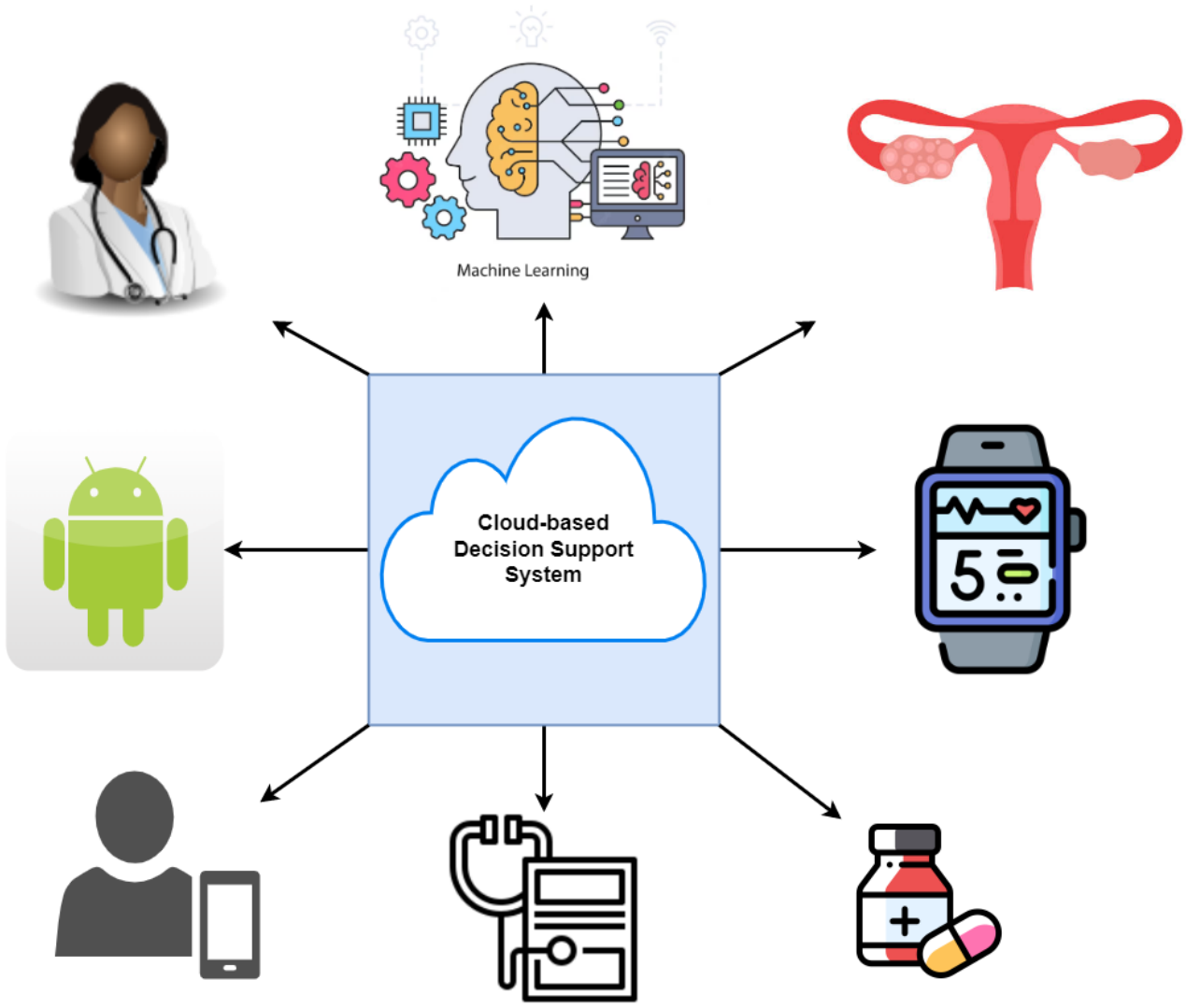

Figure 15. Future Vision of this study.

Author Contributions: Conceptualization: K.C. and N.S.; methodology: V.V.K. and K.C.; software: V.V.K.; validation: S.P.; formal analysis: V.B.; investigation: G.K.H.; resources: K.C.; data curation: V.V.K.; writing—original draft preparation: V.V.K.; writing—review and editing: K.C. and N.S.; visualization: V.V.K. and K.C.; supervision: N.S. and S.P.; project administration: G.K.H. and V.B.; funding acquisition: N.S. and G.K.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available dataset was analyzed in this study. This data can be found on Kaggle: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos, accessed on 7 December 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Azziz, R.; Carmina, E.; Dewailly, D.; Diamanti-Kandarakis, E.; Escobar-Morreale, H.F.; Futterweit, W.; Janssen, O.E.; Legro, R.S.; Norman, R.; Taylor, A.E.; et al. The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: The complete task force report. *Fertil. Steril.* **2009**, *91*, 456–488. [CrossRef] [PubMed]
2. Ndefo, U.A.; Eaton, A.; Green, M.R. Polycystic ovary syndrome: A review of treatment options with a focus on pharmacological approaches. *Pharm. Ther.* **2013**, *38*, 336.

3. Mohan, V.; Mehreen, T.S.; Ranjani, H.; Kamalesh, R.; Ram, U.; Anjana, R.M. Prevalence of polycystic ovarian syndrome among adolescents and young women in India. *J. Diabetol.* **2021**, *12*, 319–325. [CrossRef]
4. Rojhani, E.; Rahmati, M.; Firouzi, F.; Saei Ghare Naz, M.; Azizi, F.; Ramezani Tehrani, F. Polycystic Ovary Syndrome, Subclinical Hypothyroidism, the Cut-Off Value of Thyroid Stimulating Hormone; Is There a Link? Findings of a Population-Based Study. *Diagnostics* **2023**, *13*, 316. [CrossRef] [PubMed]
5. Mcdonald, T.W.; Malkasian, G.D.; Gaffey, T.A. Endometrial cancer associated with feminizing ovarian tumor and polycystic ovarian disease. *Obstet. Gynecol.* **1977**, *49*, 654–658. [CrossRef] [PubMed]
6. Diamanti-Kandarakis, E.; Christakou, C.D. Insulin resistance in PCOS. *Diagn. Manag. Polycystic Ovary Syndr.* **2009**, 35–61. [CrossRef]
7. Schorr, H.; Rappaport, A. *Innovative Applications of Artificial Intelligence*; AAAI Press: Cambridge, UK, 1989.
8. Benke, K.; Benke, G. Artificial intelligence and big data in public health. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2796. [CrossRef]
9. Szolovits, P.; Patil, R.S.; Schwartz, W.B. Artificial intelligence in medical diagnosis. *Ann. Intern. Med.* **1988**, *108*, 80–87. [CrossRef]
10. Tang, Y.M.; Zhang, L.; Bao, G.Q.; Ren, F.J.; Pedrycz, W. Symmetric implicational algorithm derived from intuitionistic fuzzy entropy. *Iran. J. Fuzzy Syst.* **2022**, *19*, 27–44.
11. Tang, Y.; Pan, Z.; Pedrycz, W.; Ren, F.; Song, X. Based kernel fuzzy clustering with weight information granules. *IEEE Trans. Emerg. Top. Comput. Intell.* **2022**, 1–15. [CrossRef]
12. Mulyanto, M.; Faisal, M.; Prakosa, S.W.; Leu, J.S. Effectiveness of focal loss for minority classification in network intrusion detection systems. *Symmetry* **2021**, *13*, 4. [CrossRef]
13. Chen, M.; Wei, Z.; Ding, B.; Li, Y.; Yuan, Y.; Du, X.; Wen, J.R. Scalable graph neural networks via bidirectional propagation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 14556–14566.
14. Bhardwaj, K.K.; Banyal, S.; Sharma, D.K. Artificial intelligence based diagnostics, therapeutics and applications in biomedical engineering and bioinformatics. In *Internet of Things in Biomedical Engineering*; Academic Press: Cambridge, MA, USA, 2019; pp. 161–187.
15. Liu, L.; Shen, F.; Liang, H.; Yang, Z.; Yang, J.; Chen, J. Machine Learning-Based Modeling of Ovarian Response and the Quantitative Evaluation of Comprehensive Impact Features. *Diagnostics* **2022**, *12*, 492. [CrossRef] [PubMed]
16. Khanna, V.V.; Chadaga, K.; Sampathila, N.; Prabhu, S.; Chadaga, R.; Umakanth, S. Diagnosing COVID-19 using artificial intelligence: A comprehensive review. *Netw. Model. Anal. Health Inform. Bioinform.* **2022**, *11*, 1–23. [CrossRef]
17. Chadaga, K.; Prabhu, S.; Sampathila, N.; Chadaga, R.; Ks, S.; Sengupta, S. Predicting cervical cancer biopsy results using demographic and epidemiological parameters: A custom stacked ensemble machine learning approach. *Cogent Eng.* **2022**, *9*, 2143040. [CrossRef]
18. Hagras, H. Toward human-understandable, explainable AI. *Computer* **2018**, *51*, 28–36. [CrossRef]
19. Islam, M.R.; Ahmed, M.U.; Barua, S.; Begum, S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Appl. Sci.* **2022**, *12*, 1353. [CrossRef]
20. Zhang, Y.; Song, K.; Sun, Y.; Tan, S.; Udell, M. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv* **2019**, arXiv:1904.12991.
21. Vij, A.; Nanjundan, P. Comparing Strategies for Post-Hoc Explanations in Machine Learning Models. In *Mobile Computing and Sustainable Informatics*; Springer: Singapore, 2022; pp. 585–592.
22. Purwono, P.; Ma'arif, A.; Negara, I.M.; Rahmaniar, W.; Rahmawan, J. Linkage Detection of Features that Cause Stroke using Feyn Qlattice Machine Learning Model. *J. Ilm. Tek. Elektro Komput. Inform* **2021**, *7*, 423. [CrossRef]
23. Witchel, S.F.; Oberfield, S.E.; Peña, A.S. Polycystic ovary syndrome: Pathophysiology, presentation, and treatment with emphasis on adolescent girls. *J. Endocr. Soc.* **2019**, *3*, 1545–1573. [CrossRef]
24. Bhardwaj, P.; Tiwari, P. Manoeuvre of Machine Learning Algorithms in Healthcare Sector with Application to Polycystic Ovarian Syndrome Diagnosis. In Proceedings of the Academia-Industry Consortium for Data Science, Wenzhou, China, 19–20 December 2022; Springer: Singapore, 2022; pp. 71–84.
25. Available online: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos?select=PCOS_data_without_infertility.xlsx (accessed on 7 December 2022).
26. Zigarelli, A.; Jia, Z.; Lee, H. Machine-Aided Self-diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study. *JMIR Form. Res.* **2022**, *6*, e29967. [CrossRef] [PubMed]
27. Bharati, S.; Podder, P.; Mondal, M.; Surya Prasath, V.B.; Gandhi, N. Ensemble Learning for Data-Driven Diagnosis of Polycystic Ovary Syndrome. In Proceedings of the International Conference on Intelligent Systems Design and Applications, Online, 12–14 December 2021; Springer: Cham, Switzerland, 2022; pp. 71–84.
28. Tiwari, S.; Kane, L.; Koundal, D.; Jain, A.; Alhudhaif, A.; Polat, K.; Zaguia, A.; Alenezi, F.; Althubiti, S.A. SPOSDS: A Smart Polycystic Ovary Syndrome Diagnostic System Using Machine Learning. *Expert Syst. Appl.* **2022**, *203*, 117592. [CrossRef]
29. Danaei Mehr, H.; Polat, H. Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. *Health Technol.* **2022**, *12*, 137–150. [CrossRef]
30. Bharati, S.; Podder, P.; Mondal, M.R.H. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 5–7 June 2020; IEEE: Dhaka, Bangladesh, 2020; pp. 1486–1489.

31. Silva, I.S.; Ferreira, C.N.; Costa, L.B.X.; Sóter, M.O.; Carvalho, L.M.L.; Albuquerque, J.D.C.; Sales, M.F.; Candido, A.L.; Reis, F.M.; Veloso, A.A.; et al. Polycystic ovary syndrome: Clinical and laboratory variables related to new phenotypes using machine-learning models. *J. Endocrinol. Investig.* **2022**, *45*, 497–505. [CrossRef] [PubMed]
32. Raju, V.G.; Lakshmi, K.P.; Jain, V.M.; Kalidindi, A.; Padma, V. Study the influence of normalization/transformation process on the accuracy of supervised classification. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; pp. 729–735.
33. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In Proceedings of the International Conference on Intelligent Computing, Hefei, China, 23–26 August 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
34. Kumar, V.; Minz, S. Feature selection: A literature review. *SmartCR* **2014**, *4*, 211–229. [CrossRef]
35. Available online: https://github.com/JingweiToo/Wrapper-Feature-Selection-Toolbox (accessed on 27 December 2022).
36. Heidari, A.A.; Mirjalili, S.; Faris, H.; Aljarah, I.; Mafarja, M.; Chen, H. Harris hawks optimization: Algorithm and applications. *Future Gener. Comput. Syst.* **2019**, *97*, 849–872. [CrossRef]
37. Debjit, K.; Islam, M.S.; Rahman, M.A.; Pinki, F.T.; Nath, R.D.; Al-Ahmadi, S.; Hossain, M.S.; Mumenin, K.M.; Awal, M.A. An Improved Machine-Learning Approach for COVID-19 Prediction Using Harris Hawks Optimization and Feature Analysis Using SHAP. *Diagnostics* **2022**, *12*, 1023. [CrossRef]
38. Abualigah, L.; Shehab, M.; Alshinwan, M.; Alabool, H. Salp swarm algorithm: A comprehensive survey. *Neural Comput. Appl.* **2020**, *32*, 11195–11215. [CrossRef]
39. Zivkovic, M.; Stoean, C.; Chhabra, A.; Budimirovic, N.; Petrovic, A.; Bacanin, N. Novel improved salp swarm algorithm: An application for feature selection. *Sensors* **2022**, *22*, 1711. [CrossRef]
40. Vergara, J.R.; Estévez, P.A. A review of feature selection methods based on mutual information. *Neural Comput. Appl.* **2014**, *24*, 175–186. [CrossRef]
41. Liu, H.; Sun, J.; Liu, L.; Zhang, H. Feature selection with dynamic mutual information. *Pattern Recognit.* **2009**, *42*, 1330–1339. [CrossRef]
42. Zhang, L.; Tan, J.; Han, D.; Zhu, H. From machine learning to deep learning: Progress in machine intelligence for rational drug discovery. *Drug Discov. Today* **2017**, *22*, 1680–1685. [CrossRef]
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
44. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Into Imaging* **2018**, *9*, 611–629. [CrossRef] [PubMed]
45. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
46. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
47. Singh, A.; Sengupta, S.; Lakshminarayanan, V. Explainable deep learning models in medical image analysis. *J. Imaging* **2020**, *6*, 52. [CrossRef]
48. Wang, D.; Thunéll, S.; Lindberg, U.; Jiang, L.; Trygg, J.; Tysklind, M. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *J. Environ. Manag.* **2022**, *301*, 113941. [CrossRef]
49. Hintze, J.L.; Nelson, R.D. Violin plots: A box plot-density trace synergism. *Am. Stat.* **1998**, *52*, 181–184.
50. Deb, D.; Smith, R.M. Application of Random Forest and SHAP Tree Explainer in Exploring Spatial (In) Justice to Aid Urban Planning. *ISPRS Int. J. Geo Inf.* **2021**, *10*, 629. [CrossRef]
51. Lubo-Robles, D.; Devegowda, D.; Jayaram, V.; Bedle, H.; Marfurt, K.J.; Pranter, M.J. Machine learning model interpretability using SHAP values: Application to a seismic facies classification task. In Proceedings of the SEG International Exposition and Annual Meeting, Virtual, 11–16 October 2020; OnePetro: Richardson, TX, USA.
52. Zehra, B.; Khursheed, A.A. Polycystic ovarian syndrome: Symptoms, treatment and diagnosis: A review. *J. Pharmacogn. Phytochem.* **2018**, *7*, 875–880.
53. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
54. Agarwal, N.; Das, S. Interpretable machine learning tools: A survey. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, Australia, 1–4 December 2020; pp. 1528–1534.
55. Broløs, K.R.; Machado, M.V.; Cave, C.; Kasak, J.; Stentoft-Hansen, V.; Batanero, V.G.; Wilstrup, C. An approach to symbolic regression using feyn. *arXiv* **2021**, arXiv:2104.05417.
56. Bharadi, V. QLattice Environment and Feyn QGraph Models—A New Perspective Toward Deep Learning. *Emerg. Technol.Healthc. Internet Things Deep. Learn. Model.* **2021**, 69–92.
57. Menze, B.H.; Kelm, B.M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F.A. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinform.* **2009**, *10*, 1–16. [CrossRef] [PubMed]
58. Neto, C.; Silva, M.; Fernandes, M.; Ferreira, D.; Machado, J. Prediction models for Polycystic Ovary Syndrome using data mining. In Proceedings of the International Conference on Advances in Digital Science, Salvador, Brazil, 19–21 February 2021; Springer: Cham, Switzerland, 2021; pp. 210–221.

59. Nandipati, S.C.; Ying, C.X.; Wah, K.K. Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques. *Appl. Math. Comput. Intell.* **2020**, *9*, 65–74.
60. Shreyas, V.; Vaidehi, T. PCOcare: PCOS Detection and Prediction using Machine Learning Algorithms. *Biosci. Biotechnol. Res. Commun.* **2020**, *13*, 240–244.
61. Hdaib, D.; Almajali, N.; Alquran, H.; Mustafa, W.A.; Al-Azzawi, W.; Alkhayyat, A. Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms. In Proceedings of the 2022 5th International Conference on Engineering Technology and its Applications (IICETA), Al-Najaf, Iraq, 31 May–1 June 2022; pp. 532–536.
62. Çiçek, İ.B.; Küçükakçali, Z.; Yağin, F.H. Detection of risk factors of PCOS patients with Local Interpretable Model-agnostic Explanations (LIME) Method that an explainable artificial intelligence model. *J. Cogn.Syst.* **2021**, *6*, 59–63.