Op-RMSprop (Optimized-Root Mean Square Propagation) Classification for Prediction of Polycystic Ovary Syndrome (PCOS) using Hybrid Machine Learning Technique

Rakshitha Kiran P[1]
Assistant Professor, Dept of MCA
Research Scholar Dept of ISE, Dayananda Sagar College of Engineering, Bengaluru-78, affiliated to VTU

Naveen N. C[2]
Professor and Head, Dept of CSE
JSS Academy of Technical Education
Bengaluru-60, affiliated to VTU

***Abstract*—Polycystic Ovary Syndrome is a common women's health problem caused by the imbalance in the reproductive hormones which causes problems in the ovaries. An appropriate machine learning (ML) algorithm can be applied to analyze the datasets and validate the performance of the algorithm in terms of accuracy. In this paper, a unique hybrid and optimized methodology are proposed which uses SVM linear kernel with Logistic Regression functionalities in a different way. The output of this model is passed on to the RMSprop optimizer. Optimization will train the model iteratively to get better output. For this research 1600 datasets were collected from the leading hospital in Bangalore Urban region. This optimized hybrid method is tested over PCOS datasets and it exhibited 89.03% accuracy. The results showed that the optimized-hybrid model works efficiently when compared to other existing ML Algorithms like SVM, Logistic regression, Decision tree, KNN, Random forest, and Adaboost. Also, the results of the optimized-hybrid SVLR model showed good results in terms of F-measure, precision, and recall statistical criteria. Overall this paper summarizes the working of the proposed optimized-SVLR hybrid model and prediction of PCOS.**

Keywords—SVM; decision tree; logistic regression; RM Sprop; frameworks

I. Introduction

In the present world, there are very large innovations in the field of medicine which help clinical experts to predict any disease in a better and faster way. The research by Kirschner MA [2] shows that around 60-70% of the Indian young women population suffers from polycystic ovary syndrome (PCOS). PCOS is an endocrinopathy caused because of an imbalance of the reproductive hormones affecting women of reproductive age, which creates complications in the ovaries [1] [26]. The ovaries produce eggs which will be released every month in case of a healthy menstrual cycle. But ladies with PCOS will end up with irregular menstrual cycles as the egg may not develop as it should or it may not be released during ovulation [4]. Most women develop PCOS in their 20s and 30s but it can arise at any age after puberty [1]. Obesity is one of the root causes of PCOS and other infertility-related problems [5]. Women with PCOS have a number of obstacles to successful lifestyle improvement. For weight management, the intrinsic factors range is changed in PCOS, which is indicated by recent research. This has an impact on how well PCOS-affected women can control their weight, although there is currently insufficient and conflicting evidence.

Computational model-based frameworks constructed with ML techniques are now widely regarded as valuable tools for predicting and analyzing a wide range of diseases. Around-the-clock ML approaches are sufficient to successfully and proficiently predict the disease [7]. ML models, in contrast to traditional techniques, do not require in-depth knowledge of data insights. SVM, Nave Bayes, Decision Tree, and Artificial Neural Network (ANN) is classifier models of ML approaches that are commonly used in medical services.

Due to the expensive computational tasks, and overfitting conditions, high dimensional data can have an impact on classifier accuracy in the majority of the existing research. The multiple characteristics are used by the previous research for classifying the PCOS problem, which affects the effectiveness of the classification results. To overcome these challenges, a novel Op-RMSprop algorithm has been proposed. The overfitting problem is reduced by selecting the most significant data by using the proposed model, and also it enhances the classification performance and processing time.

The proposed research's primary goal is to provide a unique and accurate prediction model which could predict the possibility of PCOS patients becoming infertile soon. Here a hybrid model is built combining the functionality of the two models Support Vector Machine and Logistic Regression Algorithms and the model is optimized using the RMSprop optimizer. One of the most important aspects of ML is optimization [6].The optimization model is created and the features of the optimization method from the input data are learned by the ML-based algorithms [8].The popularization and implementation of ML models are significantly influenced by the efficiency and effectiveness of quantitative optimization algorithms in the era of large databases [9]. For this research RMSProp Optimization model [33] has been used, which is one of the best optimization techniques available and provides optimal output, also this technique reduces the learning rate monotonically.

This paper summarizes various first-order optimization techniques in Section II, a Literature review on the application of ML in the healthcare domain in Section III, criteria for evaluation of ML algorithms in Section IV, the proposed framework in Section V, and Results of the proposed framework in Section VI and finally in Section VII Conclusion.

II. Literature Review

In [16], Prof. Keshavaraj GK, and Prof.SuryaSukurmran explain about Data mining process where knowledge is extracted from huge datasets. Three main modules of Data mining are Clustering/ Classification, Association Rules, and Sequence Analysis. In classification/clustering data set is analyzed and a set of grouping rules is generated which is used to classify future data. In DM information is extracted from data sets and transformed into an understandable structure. It follows a computational methodology for discovering patterns in large data sets involving different approaches like AI, ML, statistics, and database systems. All DM tasks are either automatic or semi-automatic used for the examination of huge amounts of datasets. There are six common classes of tasks in DM they are: Anomaly detection, Regression, Classification, Clustering, Association rule learning, and Summarization. Classification is a prime methodology in data mining and is widely used to predict relationships for data instances. In this paper, a few basic classification techniques like DT induction, KNN classifier, and Bayesian networks are discussed. This paper aims to give an insight into various classification techniques in data mining.

In [17], the author summarizes the caused deaths due to heart disease which has become a major issue. With the rise in heart stroke rates at younger ages, there is a need to put in place an early-stage device to identify the signs of a heart attack and thereby avoid it. It is not possible for a common man to undergo ECG frequently and as a result, there must be an application to detect chances of heart stroke at an early stage. In this paper, the authors have proposed a model which can predict the chances of a heart stroke using basic attributes like BP, age, gender, pulse rate, etc. An Artificial Neural network algorithm has given the most accurate result.

In [18] the author gives insights into the constructive examination of different chronic diseases. ML algorithms provide a large impact on health care by giving an effective inspection of problems for accurate diagnosis. In this paper, the authors talk about the kidney problem which is associated with other numerous factors like hypertension, aging, and diabetes, and its effects on people in the age group 60 and above. The authors used ML techniques to analyze chronic kidney disease (CKD). Around 400 data sets were collected from the UCI repository and Apriori algorithm with 10-fold-cross validation. Six classification different algorithms like ZeroR, Naïve Bayes, J48, OneR, and IBK were applied to the datasets. Data were preprocessed and normalization of missing data was done before analyzing datasets. The results shown were 99% detection accuracy for CKD datasets using the Apriori algorithm [13]. This study examines different ML methods, especially classification and association techniques. The paper also analyses the impacts of utilizing feature determination procedures in amalgamation with classification. This was carried out using the ANACONDA python tool. The outcomes were cross-checked with correctly classified instances, mean absolute value, and kappa statistic, with or without the feature_selection methodology. Datasets are processed with the Apriori_Association algorithm and the best results were achieved with IBk and Apriori associative algorithm with an accuracy of 99%.

Diabetes is considered to be the worst and perhaps most chronic illness that causes sugar levels to increase. If diabetes remains untreated and unidentified, a lot of problems can happen [24]. To address the common yet crucial problem, the ML concept was used. In this paper likelihood of diabetics was predicted using a high precision value obtained from the model built using ML techniques. Thus classification algorithms DT, SVM, and NB are used here, to detect diabetes at an early phase. Using Pima Indians Diabetes Database (PIDD) from the UCI system observation was done. The performance of all algorithms is measured on different parameters such as Precision, Accuracy, F-Measure2, and Recall. The values obtained from the experiments were verified using Working Characteristic Receiver (ROC) [24].

Breast cancer (BC) is the most often observed problem faced by women worldwide, causing cancer-related deaths [20]. To enhance the prediction and probability of survival considerably breast cancer must be detected at an early stage. Accurately classification of benign tumors can help patients avoid unwarranted treatments. In this paper, the author reviews ML techniques in breast cancer detection, and diagnosis. Artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and k-nearest neighbors (k-NNs) were applied to the cancer datasets [20].

Daqqa, Khaled A. S. Abu et al [21], used data mining techniques for getting patterns and models which are undiscovered in datasets. Leukemia is a condition that affects blood status. Blood_Cell_Counter (CBC) is employed to determine Leukemia detection. Leukemia is determined by examining the blood cell relations, gender, age, and also current health condition of patients using ML techniques. Datasets of 4,000 patients were involved in this study. Three classification algorithms like SVM, k-NN, and DTs were used in this study. From the experimental results, while comparing the models 77.30% accuracy is achieved by the DT algorithm than the other methods.

The main research gap of the previous methods includes the lower performance results of the detection and diagnosis of the PCOS problem, high processing time, high overfitting problem, and high computational cost. The proposed model solves these problems, reduces the overfitting risk, and improves the classification performance and processing time of the model.

III. Need for Optimization in Machine Learning

Building a model hypothesis, describing the objective function, and the model parameters are determined by solving the minimum and maximum of the objective function, these steps are important aspects of ML [10]. The first two steps of

these three crucial processes are ML modeling problems, and the desired model is solved by using the third step with optimization methods. The optimization problems are formulated in almost every ML algorithm. The optimization model is created and the features of the optimization method from the input data are learned by the ML based algorithms.

The most commonly used optimization technique is a first-order method. An extensive survey done on optimization methods by Sun et al [33] says that ML frequently employs first-order optimization algorithms. The commonly utilized first-order optimization algorithms that are used in the ML journey are summarized in Table I.

TABLE I. SUMMARIZATION OF FIRST-ORDER OPTIMIZATION TECHNIQUES [33]

Method	Properties	Advantages	Disadvantages
Gradient Descent	Here With every update, the model finds the target and gradually converges to the objective function's optimal values.	the objective function is convex	The cost of calculation is high since it operates by dynamically adjusting elements in the reverse direction of the objective function's gradient.
Stochastic Gradient Descent	Here, each iteration is based on a random sample that updates the gradient (theta) rather than directly calculating the gradient itself.	The cost of calculation is reduced since the time it takes to calculate each update is independent of the total number of training samples.	Determining the suitable learning rate is difficult.
Adagrad	Here the learning rate will be low if the high gradient, and vice versa	The approach can be used to solve problems with sparse gradients. Each parameter's learning rate is adaptively adjusted.	For addressing non-convex problems, it is not suitable.
RMSprop	It changes the way from total gradient accumulation to an exponential moving average.	It can be used to solve non-stationary and non-convex issues. Also, suitable for large and multidimensional space	The update procedure is rehashed around the local minimum within the late training period.
Adam	Combines the momentum method and adaptive methods	With large amounts of data and larger feature space, it is effective for solving most non-convex optimization problems	In some instances, the approach may fail to converge.
ADMM (alternating direction method of multipliers)	The approach addresses optimization problems with linear constraints.	Divide and rule methodology	Difficult to calculate the penalty

From Table I, we can infer that the RMSprop optimization technique is one of the best techniques which can be applied to medical datasets. It's one of the good and fast optimizers when compared to the existing optimizers. RMSProp is an algorithm that aims to find the global minima where the cost function reaches the smallest possible value. The technique relies on the concept of the Exponentially Weighted Average (EWA) of the gradients. The exponentially weighted average (EWA) is used to determine the moving average. It consists in keeping the previous values in a memory buffer. This is achieved by using this recursive formula:

$$V_t = \beta V_t\text{-}1 + (1-\beta)\,\theta_t \quad (1)$$

Where V_t: Moving average value at 't ' i.e. averaging θ_t over $1/1-\beta$ units (approx).

IV. BACKGROUND

PCOS is a famine problem faced mostly by young women between the ages of 19-35yrs. The risk of PCOS is higher if a woman is obese or if her mother, her sister, or her aunt had PCOS [2][27]. A few common signs of PCOS are as follows:

- Women undergoing PCOS may have an Irregular menstrual cycle i.e. they will fail to get periods or may have fewer periods (less than 8 times/ year) or they may have periods every 21 days or more frequently. Few ladies may even stop getting periods [11][25].
- Excessive hair on the facial hair or various body parts that men normally have. This phenomenon is named "Hirsutism", which affects about 70-80% of women having PCOS.
- Acne may be developed on the face, chest, or upper back.
- Hair thinning or loss of hair can happen on the scalp resulting in baldness.
- Increase in the weight or problem in losing weight
- Skin darkening in the neck region or other parts of the body.
- Skin tags are tiny flaps of excess skin in the neck section or armpits.

PCOS is caused because of two main reasons:

1) More production of androgens hormones, at times, is called "male hormones". Every woman contributes towards making a small number of androgens. Androgens are responsible for the overall development of male features such as male-pattern baldness in the female body [2][34]. Women having PCOS have higher androgens. Two major signs of androgens are it stops the egg from releasing during each menstrual cycle and extra growth of hair and acne [12][28].

2) An increase in insulin levels results in PCOS. Insulin is a hormone that is responsible for converting food into energy. Insulin resistance is a condition in which cells of the body fail to respond adequately to insulin, resulting in higher levels of insulin blood [3][35]. Most women with PCOS undergo insulin resistance when they are obese or overweight [14].

3) Multiple key factors include poor eating habits, less physical activity, and having a diabetes family background (usually type 2 diabetes). This can lead to type 2 diabetes over time. Fig. 1 shows the complex interaction with the underlying problem of PCOS. The figure describes the root cause of PCOS [29].

This section presents the results obtained through experimentation. It is important to find the best fitting classification algorithms [19] [22]. There are various criteria to measure the performance of the algorithms; they are listed as follows [23]:

- Classification accuracy: is the potential of the design to effectively foresee the class labels. It is given in the form of percentage.
- Speed: is the time taken to bring up the model.
- Robustness: to predict the system correctly with missing data and noisy observations.
- Scalability: A model can be precise and efficient when managing a growing quantity of data.
- Interpretability: the degree of interpretation that the algorithm provides.
- Rule Structure: Understanding the rule structure of the algorithm.

The second stage is to inspect performance criteria like measurements, the speed, and frequency of the working model, and intelligibility.

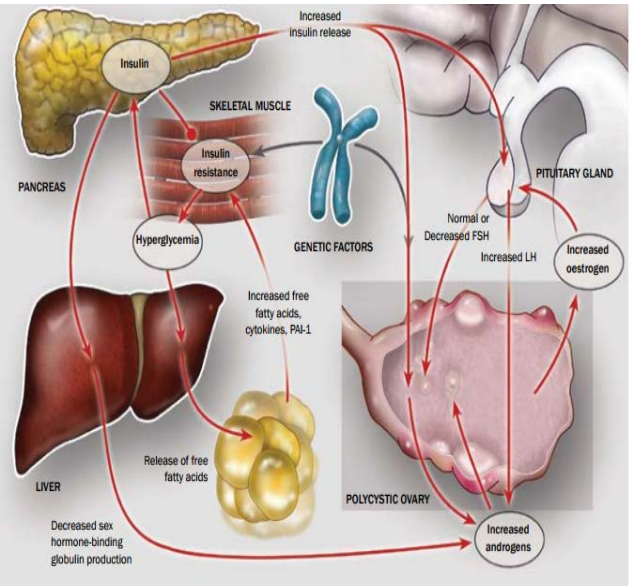

Fig. 1. Complex Interaction Underlying PCOS.

Accuracy (AC): is the percentage of correct predictions. It is calculated as per the confusion matrix.

$$AC = \frac{TN + TP}{TP + FP + FN + TN`} \quad (2)$$

- TN --> true "-"
- TP --> true "+"
- FP --> false "+"
- FN--> false "-"

Precision (P): reflects the fraction of positive observations of "+" observations correctly predicted among the total "+" observations predicted.

$$P = \frac{TP}{TP+FP`} \quad (3)$$

Recall (R): calculates the proportion of accurately projected positives in each class.

$$R = \frac{TP}{TP + FN} \quad (4)$$

F-measure: The criteria for Recall and Precision are taken together than individually. Thus the F-Measure values that are obtained by the Harmonic Mean (HM) of both methods are considered. F-measure thus provides two levels of classification accuracy.

$$F - measure = \frac{2 \times P \times R}{P + R`} \quad (5)$$

Where R is the recall and P is the precision.

ROC area: The curve shows how different classification algorithms perform in terms of prediction value. It is important for selection criterion for finding the correct methodology for classification. If the value approaches 1, it demonstrates that the classification was done properly.

RMSE: The RMSE is determined as the Mean Squared Error's square root. To calculate the dissimilarity between actual values and estimated values, RMSE is used. It shows the difference between expected and observed values' standard deviation. The RMSE value is desirable to a small.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n} (yi - \hat{y}i)}{n}} \quad (6)$$

V. Research Methodology

The main purpose of this paper is to examine classification ML techniques based on the accuracy of prediction for PCOS detection. The paper explores different detection methodologies with different techniques using classification algorithms and analyses them for events that have been accurately described. These classification algorithms give early-stage PCOS analysis standards.

In the proposed methodology, a prediction model is built, and also the comparison of various classification algorithms is shown in Fig. 5. The main goal of this research is to propose the best classification technique using ML to predict PCOS. A comparative study of the proposed method is carried out using other cutting-edge techniques. Fig. 2 describes the different phases in brief:

1) Dataset selection process: The PCOS [30][32] is detected by selecting the dataset for analyzing data to extract necessary information. Datasets to implement the ML technique should be in large numbers to get accurate results. The dataset for this research is obtained from ESIC Hospital, Bangalore urban region. The data was collected from ladies within the age group of 19-40, working in different firms.

2) Data Preprocessing and feature selection step: The dataset retrieved from Survey had 20 attributes, out of which only 17 attributes are applicable for this research. The few missing records, invalid values, duplicate values, and unnecessary fields were removed. Based on the attribute-relation file format, the dataset is created using 17 attributes. It is then transformed into binomial form.

3) Data: A dataset stored in CSV format contained 1800 PCOS patient details containing 17 selected attributes. The final "class" attribute has the value "0/1", which shows that an individual with PCOS, like 1, and normal patients as "0". The PCOS dataset's standards, representations, and attributes are described in Table III. The dataset has 735 “non PCOS” and 1065 “PCOS” cases.

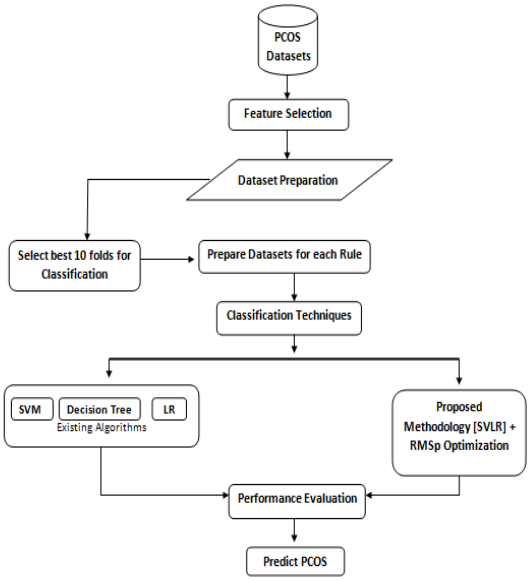

Fig. 2. Proposed Methodology.

A. Proposed Optimized-Hybrid SVLR Classification Technique

This is a hybrid classifier combining the two classifiers i.e. Support vector machine and Logistic regression classifiers. SVM can be used as a regression method by maintaining a few characteristics that characterize the algorithm. SVLR uses the same principles as the SVM for classification but with minor changes. This SVLR technique performs two stage classifications. The training and testing stages have a distinct window to extract the features from the datasets. The dataset after preprocessing is passed on to the SVLR machine model to analyze the dataset. The schematic diagram of the proposed method is given in Fig. 3.

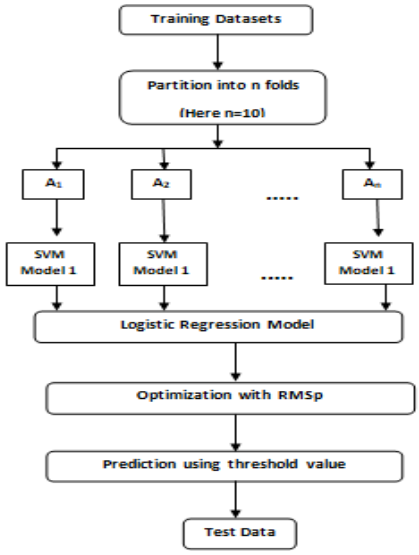

Fig. 3. Working on Proposed Optimized-SVLR Model.

B. SVLR Algorithm

Let's represent the Dataset in the form (A1, A_2, A_N) where N represents N- samples and where= 1, 2,..,10.

Step 1: There is an 'N' number of SVM Classifier models, represented as SV_i running over 'N' sets of datasets. For every run, SVM Classifier will form a Hyperplane to say h_i to SV_i.

Step 2: The distance d_R to the hyperplane must be calculated for ‘N’ samples in the data set. Therefore the vector with ‘n’ dimension where $d=d_{k,1}…d_{k,n}$ is obtained.

Step 3: The ‘d’ vector is then prepared to be given to the Logistic Regression Model which will take all responses of the SVM model.

Step 4: After the evaluation with the LR model, the prediction is done and necessary parameters are evaluated using the threshold value. In the proposed method, the description of the attributes is given in Table II.

TABLE II. DESCRIPTION OF ATTRIBUTES

SL.No	Attributes	Description
1	Are your periods regular? (YES/NO)	two nominal values "yes" and "no"
2	Are you gaining weight Rapidly? (YES/NO)	two nominal values "yes" and "no"
3	Are you facing an excess of facial or Body Hair? (YES/NO)	two nominal values "yes" and "no"
4	Do you have patches of dark areas on your skin? (Yes/No)	two nominal values "yes" and "no"
5	Do You suffer from pimples? (YES/NO)	two nominal values "yes" and "no"
6	Do you face depression and anxiety? (YES/NO)	two nominal values "yes" and "no"
7	Do you have any family history of Hyper Tension? (YES/NO)	two nominal values "yes" and "no"
8	Are you finding any difficulty in maintaining your body weight? (YES/NO)	two nominal values "yes" and "no"
9	Do you have oily skin? (YES/NO)	two nominal values "yes" and "no"
10	Are you losing a lot of hair or has it become thinner in its strength? (YES/NO)	two nominal values "yes" and "no"
11	Do you exercise regularly?	two nominal values "yes" and "no"
12	Are you mentally stressed due to the following exercise? (Are you newly admitted to the hostel?)	two nominal values "yes" and "no"
13	Are you mentally stressed due to the following exercise? (Do you have personal problems?)	having two nominal values "yes" and "no"
14	Are you mentally stressed due to the following exercise? (Peer pressure?)	two nominal values "yes" and "no"
15	Are you mentally stressed due to the following exercise? (Change in dietary habits?)	two nominal values "yes" and "no"
16	How often do you eat fast food?	two nominal values "yes" and "no"
17	Any Family history of diabetes?	two nominal values "yes" and "no"
18	Class Label	PCOS suffering patient-Yes (1) else No(0)

C. RMSprop Optimization

Root Mean Squared Propagation is an extension to the gradient descent optimization algorithm [31] and is designed to speed up the optimization process, by decreasing the number of function evaluations that are needed to improvise the functionality of the optimization algorithm to obtain the best. RMSprop is similar to gradient descent with momentum in that it employs an exponentially weighted average of gradient, but the distinction is that it updates the parameters [15].

Implementation: The algorithm works better results by updating the model parameters such as the Weight (W) and bias (B).

Consider the parameters Wi and Wj which are used to update the parameters W and B. During the backward propagation:
Weight = W — learning rate * Wi

Bias = B — learning rate * Wj

The weighted averages of Wi and Wj's squares are exponentially weighted in the RMSprop algorithm.

$\Delta Wi = \beta * \Delta Wi + (1 — \beta) * Wi_2$

$\Delta Wj = \beta * \Delta Wj + (1 — \beta) * Wj_2$

Here, 'β' momentum is a separate hyperparameter that ranges from 0 to 1.

Then a new weighted average of observed and previous values must be calculated. After that the parameters are updated.

Weight = W — learning rate *Wi / sqrt(ΔWi)

Bias = b — learning rate * Wj / sqrt(ΔWj)

ΔWi is comparatively small so it is divided by Wi and ΔWj is comparatively large, so Wj is divided with a relatively larger number to slow down the changes on a vertical dimension.

VI. EXPERIMENTAL RESULTS

A. Performance Evaluation of Existing ML Algorithms for PCOS Datasets

The PCOS datasets have been tested over different ML Algorithms. The performance measures of the paper have been summarized in Table III.

F-measure, Recall, Precision, ROC area, RMSE, and Accuracy value was calculated and compared.

The highest classification accuracy of 87.39 percent is obtained by Logistic Regression and SVM classifiers, as seen in the tables above. While comparing the RMSE values, also the SVM and LR classifiers achieve the reduced results of 0.35 then the other classifiers, and Decision Tree (DT) obtains the worst performance results with a value of 0.361.The LR algorithm has the largest ROC area, with a value of 0.87, therefore it has the best classification performance. SVM and LR are the best methods when precision and F-measure are considered. According to the recall criterion, the SVM and LR show a good value of 0.874. In terms of recall, precision, F-measure, and accuracy evaluation metrics, SVM and logistic regression are the best algorithms.

TABLE III. EVALUATION OF ML ALGORITHMS

Algorithm	F-measure	Recall	Precision	ROC area	RMSE	Accuracy (%)
KNeighbors Classifier	0.827	0.84	0.827	0.883	0.32	84.032
Support Vector Machine (SVM) Classifier	0.871	0.874	0.869	0.788	0.355	87.395
Decision Tree (DT) Classifier	0.832	0.832	0.82	0.77	0.361	83.1
Random Forest (RF) Classifier	0.717	0.798	0.839	0.933	0.334	79.8319
AdaBoost Classifier	0.844	0.849	0.841	0.884	0.3291	84.87
Gaussian Naïve base Classifier	0.832	0.832	0.832	0.896	0.3365	83.1933
Logistic Regression (LR)	0.875	0.874	0.876	0.844	0.3572	87.395

B. *Performance Evaluation of Optimized RMSprop-SVLR Hybrid Algorithm for PCOS Datasets*

The paper proposes a new optimized and hybrid ML Model which combines the features of both Support Vectors Machine as well as Logistic Regression Algorithms and performs optimization to enhance the output. The optimized-SVLR Model takes PCOS Datasets as input and checks for performance metrics.

The output value is accurately determined by the proposed method as per the results. Fig. 4(b) depicts the RMSE reduction curve of the gradient descent method during training and testing. The expected and observed outputs of the test signal are also shown in Fig. 4(a). The RMSE values can steadily decrease with iteration when relatively modest learning rates are utilized due to an effective network setup.

The average precision score and the recall trade-off were then shown. The Precision-Recall Curve depicts the distribution of values in detail. When there is a major deficiency in the dataset, the Precision-Recall Curve is performed and the PCOS problem is identified based on the performance of the curve because depending on a single metric is damaging and not a sufficient measure of selecting the algorithm

The Precision-Recall Curve and Average Precision Score are depicted in Fig. 5. Precision Score is the AP at the top of each graph. The AP for the Precision-Recall Curve is Average Precision Score is shown below in the figure.

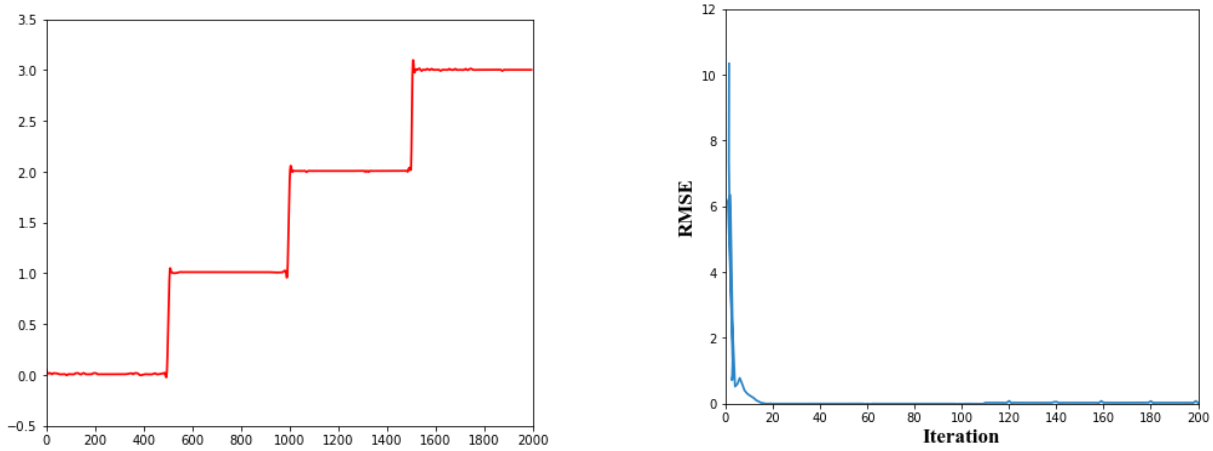

Fig. 4. Numerical Labels are used to calculate the RMSE in the Training Phase. Classification Results using the Provided (a) Over Modeled Labels as Output Labels and (b) RMSE Value based on Numerical Labels during the Training Process.

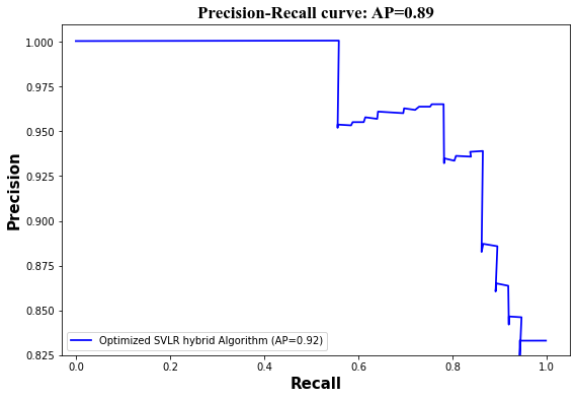

Fig. 5. With Average Precision, the Precision-Recall Curve for Hybrid-SVLR Algorithm.

Table IV shows the performance of an optimized-SVLR algorithm. The optimized SVLR Algorithm is a hybrid algorithm that combines the features of Support Vector Machine and Logistic Regression. The model is optimized with RMSprop optimized algorithm. It is observed that Optimized-SVLR performed well for the given datasets for all the performance measures.

C. Results and Discussion

The higher accuracy of 89% is achieved by the proposed Optimized-SVLR algorithm based on input images with feature extraction and PCOS diagnosis as per the results. The performance of each classifier is shown by the ROC curve for a better representation of this comparison. The true-positive and false-positive rates are compared by forming the curve. Based on this curve, the best classifier has the lowest false positive rates and highest true-positive rates. Fig. 6 illustrates the ROC curve.

As per the results, the developed SVLR model is the best classifier for PCOS diagnosis. Moreover, the Random Forest Classifier and the Gaussian Nave basis Classifier are the second and third high-performance classifiers, respectively.

The accuracy comparison of several existing classifiers with the proposed classifiers is given in Fig. 7. The optimized-hybrid SVLR model produces better accuracy than the prior classifiers; the SVLR model effectively diagnoses the PCOS problem.

TABLE IV. PERFORMANCE MEASURES OF OPTIMIZED SVLR HYBRID ALGORITHM

Optimized SVLR Algorithm	Performance measures
Accuracy	89 %
Recall	0.92
Precision	0.89
ROC area	1.0
RMSE	0.29
F-measure	0.91

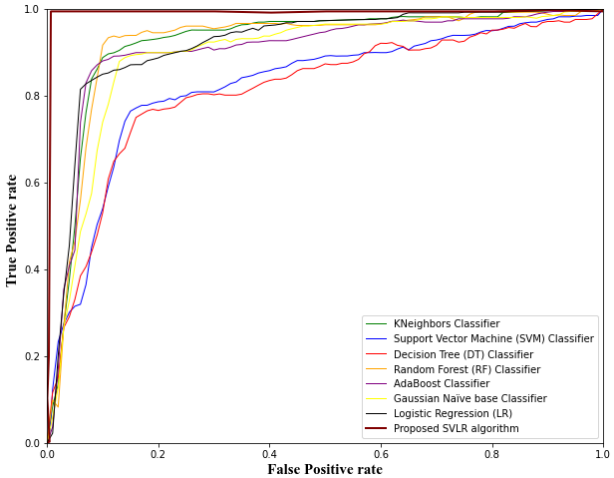

Fig. 6. The ROC Curve of the Classifiers for Diagnosis of PCOS.

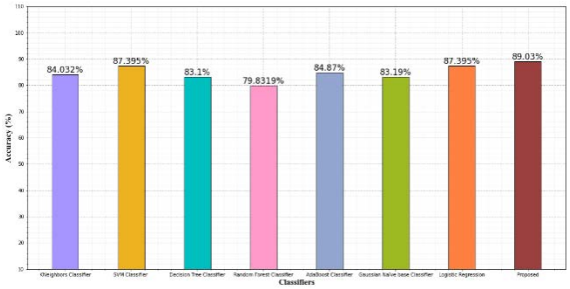

Fig. 7. Accuracy Comparison Graph.

PCOS is a significant medical problem among women caused by the imbalance in the reproductive hormones which causes problems in the ovaries. An appropriate ML algorithm can be applied to analyze the datasets and validate the performance of the algorithm in terms of accuracy. A different hybrid and optimized methodology are proposed in this paper, which uses SVM linear kernel with Logistic Regression functionalities in a different way. The RMSprop optimizer receives the output of this model. The performance of the diagnosis is improved by training the model iteratively using optimization. 1600 datasets from the top hospital in Bangalore's urban area were gathered for this research.

VII. CONCLUSION AND FUTURE REFERENCE

This paper addresses women's fertility problems caused because of polycystic ovary syndrome (PCOS). PCOS is a very common problem faced by women of the reproductive age. These risk factors causing this problem are unhealthy food habits, lack of exercise, hereditary, diabetes, prolonged medications, etc. The ML technique is applied to 1800 datasets that were collected from ESIC hospital, Bangalore. Various classification algorithms like, SVM, Decision Tree, Logistic Regression, Random forest, NB, Adaboost, and KNN were applied to the datasets collected. Accuracy, RMSE, ROC, Precision, Recall, and F-measure were calculated for the classification algorithm. Optimizing improves the performance of the algorithm and thus improves the overall performance. It is observed that the optimized hybrid SVLR performs well when compared with all other classification algorithms. The Optimized SVLR algorithms give an accuracy of 89.03, which is comparatively better than other classification algorithms. Also this new proposed optimized SVLR Algorithm could be applied to other healthcare domains to obtain better results in both medical as well as ML domains.

ACKNOWLEDGMENT

We declare that this manuscript is original, has not been published before, and is not currently being considered for publication elsewhere.

CONFLICTS OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding**:** The authors received no specific funding for this study.

Availability of Data and Material: Not applicable.

Code Availability: Not applicable.

Authors' Contributions: The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Ethics Approval: This material is the authors' own original work, which has not been previously published.

Elsewhere: The paper reflects the authors' own research and analysis in a truthful and complete manner.

REFERENCES

[1] Stein, Irving F. Amenorrhea associated with bilateral polycystic ovaries,"Am J Obstet Gynecol 1935",vol. 29, pp. 181-191.

[2] Kirschner MA,"Obesity, androgens, oestrogens, and cancer risk", Cancer Res 1982,Vol. 42, pp. 3281–3285.

[3] Ehrmann, A .David,"Polycystic ovary syndrome," New England Journal of Medicine2005, 352, no. 12, pp. 1223-1236.10.1056/NEJMra041536.

[4] Professor Cindy Farquhar , Associate Professor Neil Johnson, Department of Obstetrics and Gynaecology, University of Auckland Understanding polycystic ovary syndromeIssue 12 April 2008, ISSN 1177-5645 (Print) ISSN 2253-1947 (Online).

[5] N. Pise and P. Kulkarni,"Algorithm selection for classification problems", SAI Computing Conference (SAI), London 2016, pp. 203-211, 10.1109/SAI.2016.7555983.

[6] Jiawei Han and kamber"Data Mining Concepts and techniques".

[7] S. Taneja, C. Gupta, K. Goyal and D. Gureja,"An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering," Fourth International Conference on Advanced Computing & Communication Technologies, Rohtak 2014, pp. 325-329,10.1109/ACCT.2014.22.

[8] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of Improving KNearest-Neighbor for Classification.Proc", 4rth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007), vol. 1, Aug. 2007, pp. 679-683, 10.1109/FSKD.2007.552.

[9] N.Cristianini, J. Shawe-Taylor,"Support Vector Machines. In An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge: Cambridge University Press 2000, pp. 93-124, 10.1017/CBO9780511801389.008.

[10] Burges, JC. Christopher,"A tutorial on support vector machines for pattern recognition", Data mining and knowledge discovery,vol. 2, no. 2, 1998, pp. 121-167, https://doi.org/10.1023/A:1009715923555.

[11] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho, and A. A. Freitas, "A Survey of Evolutionary Algorithms for Decision-Tree Induction," Systems Man & Cybernetics Part C Applications & Reviews IEEE Transactions on2012, vol. 42, pp. 291-312,https://doi.org/10.1023/A:1009715923555.

[12] L. Breiman, "Random Forests, Machine Learning".vol. 45, no. 1, 2001 pp. 5–32, https://doi.org/10.1023/A:1010933404324.

[13] Mohammed, Mohssen & Khan, Muhammad & Bashier, Eihab. (2016). Machine Learning: Algorithms and Applications. 10.1201/9781315371658.

[14] MuhamadHariz B. Muhamad Adnan, Wahidah Husain and Nur`Aini"Abdul Rashid, Parameter Identification and Selection forChildhood Obesity Prediction Using Data Mining", 2nd International Conference on Management and Artificial Intelligence 2012, Vol.3.

[15] L. Liu,"Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning", International Conference on Robots & Intelligent System (ICRIS), Changsha 2018, pp. 157-160, 10.1109/ICRIS.2018.00049.

[16] G K, G. KesavarajandSukumaran, Surya,"A study on classification techniques in data mining. 2013 4th International Conference on Computing, Communications and Networking Technologies,ICCCNT 2013, Pp. 1-7, 10.1109/ICCCNT.2013.6726842.

[17] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar,"Prediction of Heart Disease Using Machine Learning", Second International Conference on Electronics, Communication and Aerospace Technology (ICECA) Coimbatore 2018, pp. 1275-1278, 10.1109/ICECA.2018.8474922.

[18] Z. Wang, J.W. Chung, X. Jiang, Y. Cui, M. Wang, A. Zheng, "Machine learning-based prediction system for chronic kidney disease using associative classification technique", International Journal of engineering & Technology 2018,Vol. 7,Pp. 1161-1167, 10.14419/ijet.v7i4.36.25377.

[19] Çığşar, BegümandUnal, Deniz,"Comparison of Data Mining Classification Algorithms Determining the Default Risk," Scientific Programming 2019,Pp. 1-8, 10.1155/2019/8706505.

[20] W. Yue, Z. Wang, H. Chen, A. Payne, X. Liu, "Machine learning with applications in breast cancer diagnosis and prognosis", Designs2018, vol. 2, no. 2, pp. 13.

[21] Daqqa, Khaled AS Abu, Ashraf YA Maghari, and Wael FM Al Sarraj. "Prediction and diagnosis of leukemia using classification algorithms," 2017 8th international conference on information technology (ICIT), IEEE 2017, pp. 638-643.

[22] I. Nguyen, E. Frank, and M. Hall, "Data Mining Practical Machine Learning Tools and techniques", Morgan Kaufmann, Burlington, MA, USA, 2011.

[23] J. Stefanowski, Data Mining- Evaluation of Classifiers, "Institute of Computing Sciences Poznan University of Technology," Poland, 2010.

[24] Deepti Sisodia, Dilip Singh Sisodia,"Prediction of Diabetes using Classification Algorithms",Procedia Computer Science, Vol. 132,2018,Pages 1578-1585,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2018.05.122.

[25] Setji TL, Brown AJ,"Polycystic ovary syndrome: update on diagnosis and treatment," Am J Med. 2014 Oct vol. 127, no.10, pp. 912-9, 10.1016/j.amjmed.2014.04.017.

[26] R. Azziz"Polycystic Ovary Syndrome. Obstet Gynecol," 2018 Aug; vol. 132, no. 2, pp. 321-336, 10.1097/AOG.0000000000002698.

[27] Silva IS, Ferreira CN, Costa LBX, Sóter MO, Carvalho LML, de C Albuquerque J, Sales MF, Candido AL, Reis FM, Veloso AA, Gomes KB. "Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models," JEndocrinol Invest. 2021 Sep 15, 10.1007/s40618-021-01672-8.

[28] Silva IS, Ferreira CN, Costa LBX, Sóter MO, Carvalho LML, de C Albuquerque J, Sales MF, Candido AL, Reis FM, Veloso AA, Gomes KB. "Polycystic ovary syndrome: clinical and laboratory variables related to new phenotypes using machine-learning models." J Endocrinol Invest. 2021 Sep 15, 10.1007/s40618-021-01672-8.

[29] Shaziya, Humera. "A Study of the Optimization Algorithms in Deep Learning,"2020, 10.1109/ICISC44355.2019.9036442.

[30] S, Chang and A.Dunaif,"Diagnosis of polycystic ovary syndrome: which criteria to use and when?",Endocrinology and Metabolism Clinics 2021, vol. 50, no.1, pp.11-23.

[31] Mustapha, Aatila&Lachgar, Mohamed and Ali, Kartit. "Comparative study of optimization techniques in deep learning: Application in the ophthalmology field Comparative study of optimization techniques in deep learning: Application in the ophthalmology field," Journal of Physics: Conference Series2021, 1743. 10.1088/1742-6596/1743/1/012002.

[32] K.Soucie, T.Samardzic, K.Schramer, C. Ly and R.Katzman, "The diagnostic experiences of women with polycystic ovary syndrome (PCOS) in Ontario", Canada. Qualitative Health Research 2021, vol. 31, no.3, pp.523-534.

[33] S. Sun, Z. Cao, H. Zhu and J. Zhao, "A Survey of Optimization Methods From a Machine Learning Perspective," in IEEE Transactions on Cybernetics, vol. 50, no. 8, pp. 3668-3681, Aug. 2020, doi: 10.1109/TCYB.2019.2950779.

[34] P.Rao, and P.Bhide, "Controversies in the diagnosis of polycystic ovary syndrome", Therapeutic Advances in Reproductive Health 2020, vol. 14, p.2633494120913032.

[35] E.Khashchenko, E.Uvarova, M.Vysokikh, T.Ivanets, L.Krechetova, N.Tarasova, I.Sukhanova, F.Mamedova, P.Borovikov, I.Balashov, G.Sukhikh,"The relevant hormonal levels and diagnostic features of polycystic ovary syndrome in adolescents," Journal of Clinical Medicine2020, vol. 9, no. 6, p.18-31.