Feature Selection Strategies for Enhancing the Accuracy for Detecting Polycystic Ovary Syndrome (PCOS) Health Problem

Ayobami Ekundayo[1], John Kolo Alhassan[2 *] , Enesi Femi Aminu [3] , Solomon Adelowo Adepoju [4] , Hamzat Olanrewaju Aliyu [5] , Oluwaseun Adeniyi Ojerinde[6], Mudathir Ayomide Ekundayo[7]

[1, 2, 3,4,5,6] Department of Computer Science, School of Information and Communication Technology

[1, 2, 3, 4, 5, 6] Federal University of Technology, Minna, Nigeria.

[7] Department of Electrical Electronics Engineering, Faculty of Engineering, Eskisehir Osmangazi Universitesi, Eskisehir, Turkey.

Abstract.

A hormonal condition called Polycystic Ovarian Syndrome (PCOS) results in larger ovaries with tiny cysts on the margins. Although the exact etiology of Polycystic Ovary Syndrome is unknown, it may be a result of both hereditary and environmental factors. One of the endocrine diseases that most frequently affect women of reproductive age is Polycystic Ovary Syndrome (PCOS). Artificial intelligence (AI)-based machine learning models has the capacity to classify and predict the potential for PCOS condition. The dataset used in this study was obtained from Kaggle repository which consists of 45 features (attributes) and 541 data points. This dataset was balanced using the Synthetic Minority Oversampling Technique (SMOTE) and features were selected by employing firefly and fruitfly optimization algorithms. The firefly optimized algorithm with Random Forest obtained an accuracy score of 95.205% with 18 selected features. The KNN with firefly algorithm used 13 features and obtained an accuracy of 91.096%. The SVM with firefly algorithm uses 14 features and obtained an accuracy of 93.151%. The fruitfly algorithm with KNN, SVM and RF obtained and accuracy of 86.986%, 90.411% and 93.151% respectively.

Keywords: Data balancing, Firefly, Fruitfly, Polycystic Ovary Syndrome, Synthetic Minority Oversampling Technique.

1. Introduction

The failure of the ovary to release the egg is referred to as infertility. One of the many causes of infertility is the development of an unusually high volume and number of follicles during the ovulation phase, which is thought to be the first sign of Polycystic Ovarian Syndrome (Alamoudi *et al.,* 2023). Polycystic Ovarian Syndrome (PCOS) is a hormonal disorder causing enlarged ovaries with small cysts on the outer edges. The cause of Polycystic Ovary Syndrome is not well understood, but may involve a combination of genetic and environmental factors. One in ten women of reproductive age may have Polycystic Ovary Syndrome (PCOS), a hormonal disease. When a woman's ovulation is inconsistent or nonexistent, this condition is suspected. An excess of androgen is what defines PCOS both biochemically and clinically (Bharati *et al.,* 2022).

Ovarian dysfunction and an excess of androgen are the two main signs of PCOS. Many factors are thought to cause this syndrome, including genetics, puberty, physiological changes, mental state, and environmental influences. Menstrual abnormalities, hirsutism,

obesity, insulin resistance, and cardiovascular disease are all common among PCOS patients (Lv *et al.,* 2022).

In order to overcome this difficulty, some clinical methods, including in vitro fertilization (IVF), have been created and used. Infertility has become a global health concern due to the rising number of couples seeking in vitro fertilization (IVF) worldwide, which highlights how devastating the problem is. Some couples still have no children even after numerous IVF cycles due to misdiagnosis. IVF for women entails increased risks and expenses. One of the most common disorders affecting premenopausal women is Polycystic Ovary Syndrome (PCOS), a complicated condition with a variety of phenotypes that involves abnormalities in metabolism, reproduction, and endocrine function (Hussein & Karami, 2023). However, many research efforts have been focused on exploitation of the most cutting-edge technology in order to assure precise and proper identification of the illness indicators.

An innovative use of these technologies is machine learning models based on artificial intelligence (AI), which have the potential to categorize and forecast the possibility of PCOS condition. Examples of these models include Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), Multi-Layer Perceptron (MLP) and XGBoost (Denny *et al.,* 2019 ; Liu *et al.,* 2020 ; Hassan & Mizra, 2020 ; Nilofer & Ramkumar, 2021 ; Madhumitha *et al.,* 2021; Inan *et al.,* 2021). Of course, many researchers in this field have made several efforts to use these models to effectively identify the condition based on the available dataset. However, the level of accuracy has been limited and warped as a result of the unbalanced nature of datasets or a lack of resilience in carrying out the feature engineering process effectively. In support of this advancement, Nandipati *et al.,* (2020) used SMOTE as a data sampling approach. On the aforementioned datasets, seven machine learning (ML) techniques were employed to evaluate the performance of the classifiers, followed by five feature selection methods from Python (Spyder as an IDE) - Scikit Learn package and RapidMiner. Despite the many techniques utilized, Random Forest (93.12%, RapidMiner) has the highest accuracy with the entire dataset. KNN and SVM, on the other hand, exhibit comparable accuracy (90.83%, RapidMiner) with 10 selected characteristics.

Furthermore, based on literature there are other feature engineering techniques that could be used. Feature engineering techniques can be feature selection, feature extraction or data preprocessing. The process of selecting which relevant characteristics and qualities to include in predictive modeling, as well as which ones to leave out, is known as feature selection (FS). It is an essential task that supports machine learning classifiers in lowering error rates, processing time, overfitting, and enhancing classification accuracy (Alyasiri *et al.,* 2022). Finding the most pertinent features from provided data with a broad feature space is the process of feature selection (Bashir *et al.,* 2022).

Yang introduced the Firefly algorithm in 2008 as a Meta-heuristic, biologically inspired swarm intelligence. The flashing activities of the firefly population serve as attracting partners, communication, and warning from predators. Based on previous work, Aziz et al. (2015) developed the Firefly algorithm, which was inspired by the flashing light of a firefly. This flashing light aids in luring prey or a mate. Yang built classic FA using attraction behavior. There is a collection of preliminary solutions obtained from the population. Every firefly has the potential to be useful in the search space. Consider the following: 'N' is the

population size, and Xi is the ith population solution with i = 1,2,..., N. Light intensity (I) decreases with increasing distance (Maheswari *et al.,* 2020).

The LF strategy approach can ensure algorithm diversity during the optimization process and boost the convergence rate. The modified FOA approach, LFFOA, was used to optimize the two important parameter pairs in the ANN method, including weight and bias, and achieve the best model (Nilofer & Ramkumar, 2021).

This manual counting may lead to problems of variability, reproducibility and low efficiency. Automating this mechanism will resolve these problems. The objective of the present work is to propose an automated ovarian classification method for classifying an ovary as normal or not in an ovarian ultrasound image by using the fruit fly and fire fly based feature selection technique. Therefore, the remaining sections of the research are organized as follows: section 2 presents the account of the related works, and the proposed methodology is presented by section 3. Others are results discussion and conclusion, which are presented by sections 4 and 5 respectively.

2. Related Works

Enormous attempts have been made to diagnose and examine ailments utilizing diverse classical machine learning models and feature engineering techniques. This section provides an extensive review of effort made in this field to identify the contribution of this investigation with PCOS diagnosis perspective. To this end, the literature of Sreejith *et al.,* (2022) aims to create a classification framework for the creation of a clinical decision support system that will help doctors keep track of PCOS. The selection of features was done via a wrapper method which comprises of Red Deer Algorithm (RDA) and a RF classifier. The RDA was used by the researchers to identify the best characteristics, and the random forest classifier was used to assess them. From the Kaggle repository, a total of 541 data points with 37 instances were obtained. The preprocessing subsystem uses the z-score statistical measure to deal with outliers and noise that are present in the dataset. The evaluation matrices used in this study are accuracy (89.81%), specificity (90.43%) and sensitivity (89.73%).

The research by Dutta *et al.,* (2021) aims to identify and predict PCOS using SMOTE and classical machine learning models. A total of 541 items with 14 attributes from the UCI collection was obtained. This study employed five machine learning models (RF, LR, DT, SVM, and KNN). Principal Component Analysis (PCA) was employed as the feature extraction technique. Standard scalar preprocessing was utilized to prepare the data, and SMOTE was employed as the optimization method. In terms of accuracy, training time, F1 score, recall, prediction, and area under the ROC, the SMOTE-based LR scored best (97.11%, 0.010 sec., 98%, 98%,98%, and 95.6%).

By prioritizing the most promising oocyte for implantation and evaluating the ability of follicular fluid (FF) Raman Spectra to predict oocyte development and pregnancy outcome, Huang *et al.,* (2021) want to lessen the physical and financial constraints that PCOS patients experience. Artificial neural networks were used in this study as the machine learning method. Between October 2017 and December 2019, 300 participants (150 without PCOS and 150 with PCOS) who underwent in vitro fertilization (IVF) at Shanghai First Maternity

and Infant Hospital were enrolled in the study. The feature extraction method used in this investigation was an unsupervised Principal Component Analysis method. Additionally, Raman Spectroscopy was used in conjunction with multivariate statistical techniques by the researchers to identify metabolic alterations in FF from PCOS-afflicted women. The machine-learning algorithms using the fully connected artificial neural network (ANN) properly assigned oocyte developmental potential and clinical pregnancy based on the Raman spectra of PCOS FF with overall accuracies of 90% and 74%, respectively.

The work of Prapty & Shitu,(2020) aim at detecting the presence of PCOS in a patient by developing an efficient decision tree. The researchers employed four different traditional machine learning models in this study which are SVM, RF, KNN and Naïve Bayes classifiers A total of 542 dataset with 31 features was obtained from ten different hospitals out of which 177 patients are PCOS positive and 365 patients are PCOS negative. The Principal Component Analysis (PCA) was employed as the feature selection technique in this study. Out of these algorithms Naïve Bayes and Random Forest performed best with accuracies of 93% and 93.5% respectively.

Due to the bad effects of PCOS in women infertility, Thomas & Kavitha, (2020) aim to identify PCOS before it worsens. Doctors from several clinics and hospitals in and around Thodupzha Municipality provided the clinical diagnostic and prescription for this study. The Nguyen-Widrow initialization strategy was adopted by the researchers as their optimization method since it boots the weights and biases of each layer. Additionally, MATHLAB R2018a was used as the tool for data mining together with NB, ANN, and a hybrid of the two classifiers. In terms of accuracy (95%), recall, and f-measure, the hybridized classifier fared better than the NB and ANN.

Thufailah *et al.,* (2018) research aims to use ultrasound scans to classify PCOS using Elman Neural Network (ENN). PCO and non-PCO were categorized using SVM and ENN. The feature extraction approach employed was the Gabor wavelet method**.** Gray scalability, histogram equalization, picture binarization, morphology, invert image, and data cleaning were used for data preprocessing. The Canny operator was used to segment the image into its component parts for edge detection and follicle cutting. It may be inferred that the Gabor wavelet's extraction results utilised 16, 24, and 32 features. The feature set examined three different settings (layer delays, hidden sizes, and training function). For 16, 24, and 32 features, the accuracy of the set data is 68.8%, 66.7%, and 78.1%, respectively.

There are many features which helps in determining the PCOS status of a patient to this end, Nandipati *et al.,* 2020 discover which machine learning model is more effective at detecting the PCOS disease, as well as the features that contribute to the disease's prognosis. In this study a dataset with a total of 541 from surveys, doctor consultations, and clinical examinations was taken into account. Eight features were taken from a total of 23 features, which included clinical and metabolic data using SPSS V.2.0. KNN, SVM, RF, NB, NN, Bagging and Adaboost are the seven machine learning models used in this study to assess the disease. For Rapidminer's accuracy (93.12%), random forest performed best. The data were preprocessed using Synthetic Minority Oversampling Technique (SMOTE).

In conclusion, the suggested PCOS data prediction is promising based on the relevant literature evaluated in this study, since the accuracy result outperforms the related literature of the same dataset and those with close range.

3. The Methodology

This study used multi-stage strategy which started from data collection to model optimization. The dataset utilized in this experiment was obtained from the Kaggle repository, in accordance with the suggested methodology as shown in Figure 1, and it is the same dataset used in the research of Nandipati et al., (2020) It is evident from prior studies that researchers and physicians use various disease datasets to evaluate machine learning classification algorithms, much like the PCOS dataset. The original PCOS dataset consists of 541 instances with 42 attributes, including one attribute for the patient file number (which is ignored during data analysis). Finally, there are 41 attributes in all, with 40 of them being input attributes and PCOS serving as a class label [Positive (Yes) and Negative (No)]. The dataset demonstrates the unbalanced nature of the class labels (364 instances of class label = 0 and 177 instances of class label = 1) and missing values.

Due to the imbalanced nature of the data set SMOTE oversampling technique was employed to balance the dataset. Data preprocessing is necessary, including data cleansing, data scaling, and data balancing. Inconsistencies, missing values, and extraneous variables like patient file number and patient ID are removed from the dataset as part of the data cleaning process because they have no bearing on determining PCOS status. The conversion of some features' non-floating values into floating point values is similar. This is done to make sure that the highly scattered floating-point values in the data are scaled to match the value of the features. For this stage of the process, the Min-Max Scaling (value between 0 and 1) technique is used.

$$Scaled\_value = (X-Min)/ (Max-Min) \quad (1)$$

Where, X is the original value of the data point.

Min is the minimum value of the features in the dataset.

Max is the maximum value of features in the dataset.

This research uses the Synthetic Minority Oversampling Technique at this phase to balance the data before features are chosen. With this method, data samples can be generated for the minority class (the "Yes" class). As a result, with 364 samples from each class, both classes would have the identical collection of data points. More importantly, the lack of balance across the data classes would make it difficult for the firefly and fruitfly algorithms' feature selection techniques to function correctly.

The feature selection process includes selection of essential features or attributes that have high contribution to the final prediction class. Therefore, in this study Fruitfly and Firefly Optimization techniques were considered. The firefly algorithm is first used to pick key features from the raw dataset by assigning a penalty score or threshold to each feature. Similar to how the optimal feature for a dataset with numerical input and output were chosen , it might utilize the Fruit Fly feature technique. It is used to choose the best features that can forecast the output, or dependent variable. Finally, the datasets preprocessed and

chosen features are divided into training and testing sets in proportions of 80% to 20% for the purposes of training and evaluating the model, respectively. While the testing data is analyzed using the metrics of accuracy, precision, recall, and f1-score, the training data is supplied into the classifiers.

Google Colaboratory forms the experimental environment for the implementation of this methodology. It is crucial to note that the goal of this research is to increase the PCOS model's accuracy by using optimization strategies for feature selection. This is better represented by the flow chart below.

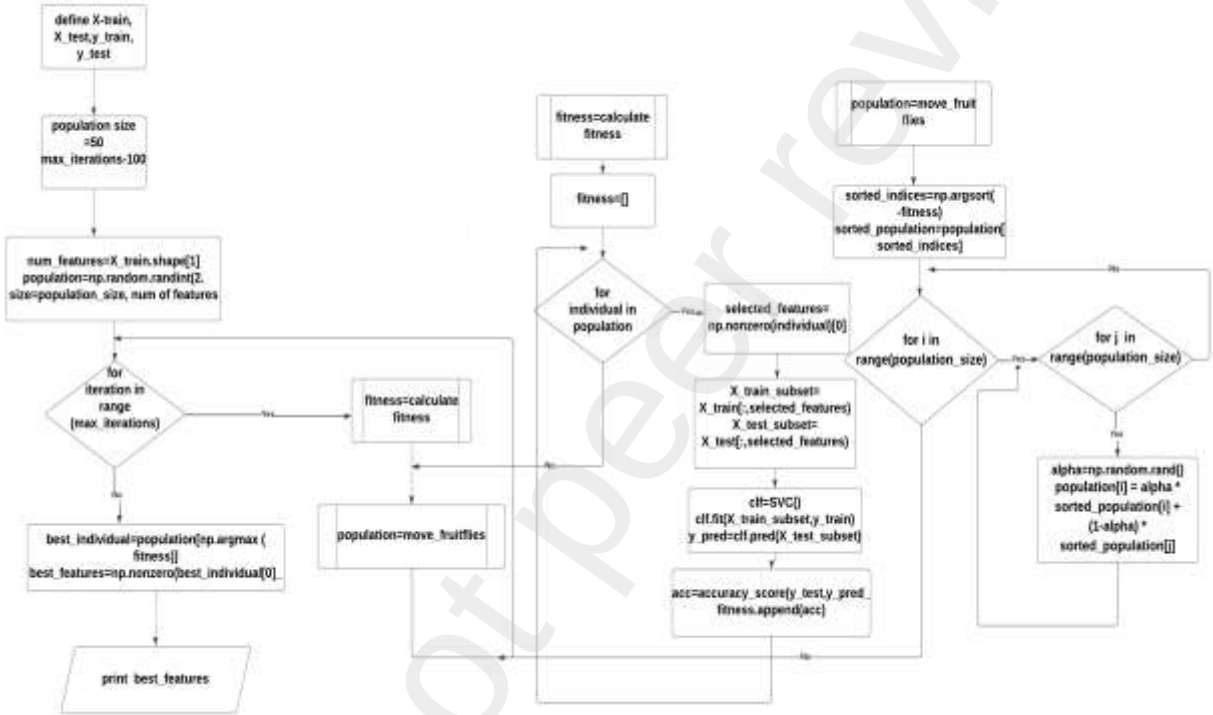

Figure 1 : Fruitfly Optimum Feature Subset Flowchart

Where:

Population_size: Number of fireflies (individuals) in the population.

Max_iterations: Maximum number of iterations the algorithm will run.

Alpha: The initial attractiveness coefficient for the movement of fireflies.

Beta: Coefficient for determining the attractiveness of brighter fireflies.

Gamma: Coefficient controlling the light absorbtion rate (distance inflluence).

Min_alpha: The minimum number of which alpha will be decayed over iterations.

The population of fireflies is initialized randomly with binary values (0 or 1), where each bit represents whether a specific feature is selected or not. The calculate_fitness module in the flow chart calculates the fitness (accuracy) of each firefly's feature selection using a machine learning (ML) support vector classifier (SVC) model. It trains specific ML model on the selected features and calculates the accuracy on the test data. The move_fireflies

module in the flow chart implements the movement of fireflies towards brighter fireflies. It initializes the population, iteratively calculates the fitness and moves the fireflies. The alpha value decays over iterations to control the exploration-exploitation trade-off. After the algorithm finishes running, it selectes the best individual (firefly) based on the highest fitness and returns the selected features as best_features.

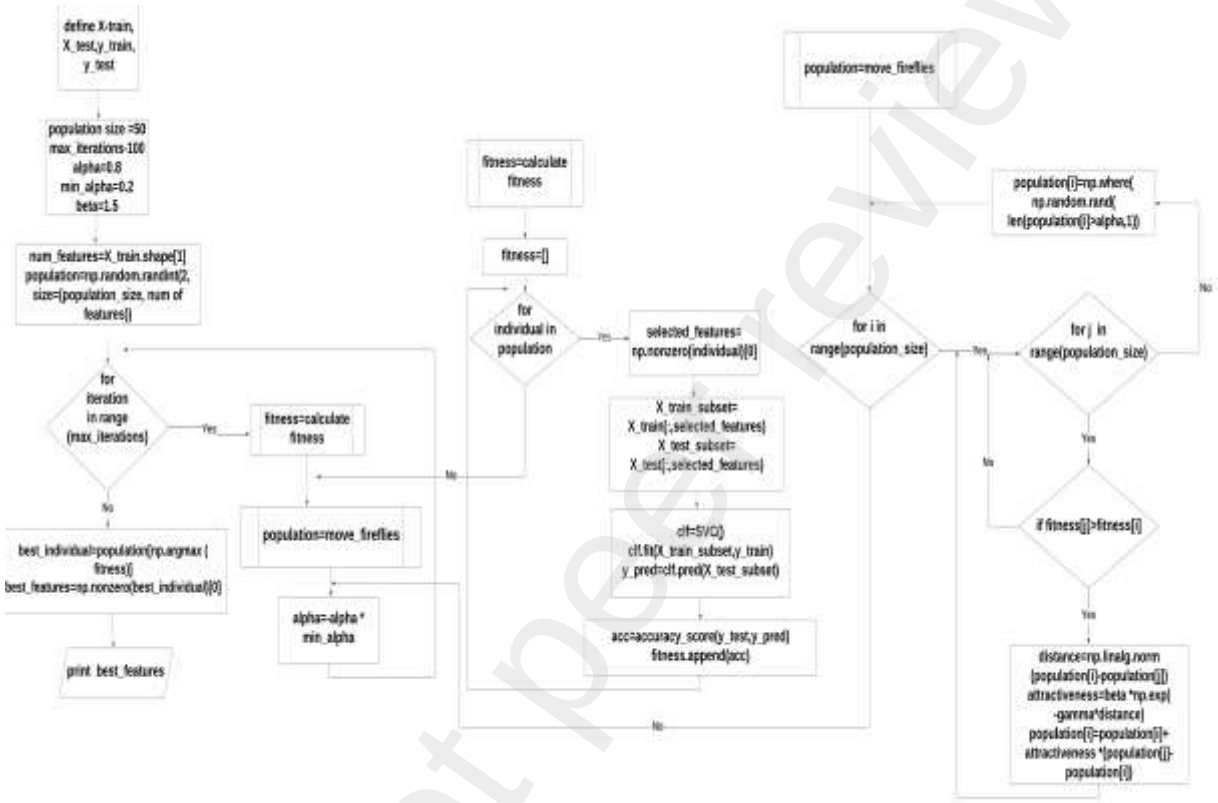

Figure 2 : Firefly Optimum Feature Subset Flowchart

Population_size : number of fruitflies (individuals) in the population.

Max_iterations: Maximum number of iterations the algorithm will run.

Num_features: The total number of features available in the dataset.

The population of fruitflies is initialized randomly with binary values (0 or 1), where each bit represents whether a specific feature is selected or not. The calculate_fitness module in the flow chart calculates the fitness (accuracy) of each fruit fly's feature selection using a ML (SVC) model. It trains the specific ML model on the selected features and calculates the accuracy on the test data. The move_fruitflies module performs the movement of fruit flies in the population based on their fitness. It sorts the fruit flies based on their firtness (accuracy), and then each fruit fly moves towards the more fit ones. When the fruit fly algorithm is used for feature selection. It initializes the population, iteratively evaluates fitness, and updates the position of fruit flies. The algorithm repeats this process for the specified number of iterations. After the algorithm finishes running, it selects the best fruit fly (individual) and returns the selected features as best_features.

4. Result and Discussion

The selected optimized feature subsets provided from these algorithms are used as input features into three different machine learning models namely KNN, SVM and RF. The key model performance metric used for comparison across the three different ML algorithms is the accuracy score. The study spans across seven main sections starting with data source, data analysis, data preprocessing, data balancing, feature selection, model building and model optimization.

There was an invalid value of 'a' in one of the numerical columns and it was cleaned by replacing with the mean of the data values in the column. Also, there was a data column with an invalid value which was fixed by replacing with the corresponding valid float value. The Feature engineering technique applied on the dataset include Data balancing using oversampling of the minority class (class 1: PCOS positive patients) by the synthetic minority oversampling technique. The number of PCOS positive patients is 177 while for negative patients is 364, which infers the need to balance the dataset in order to avoid model bias during training. Another feature engineering technique that was leveraged is feature scaling on the numerical features using Normalization technique to bring the values between 0 and 1 as the units of the numerical columns were in different ranges.

The feature selection is done with aid of the fruitfly and firefly algorithms. Fruitfly algorithm works by representing a subset of features as a position in a search space. Fruitfly agents then move within this search space, where the movement is influenced by the fitness (quality) of the solution they represent. Firefly is inspired by the flashing behavior of fireflies, where the brightness of a firefly attracts other fireflies. The algorithm aims to find the most relevant features by simulating this light attraction behavior. The selected features are then passed into the three different ML algorithms used in this research. The table below summarizes the results obtained in terms of the precision, recall, accuracy and F1 score. Training and testing were done with the balanced dataset with 80% as the training dataset and 20% as the test dataset.

Table 1: Model Result Comparison

Optimization Algorithms	Firefly Algorithm			Fruitfly Algorithm		
Machine learning Algorithms	KNN	SVM	RF	KNN	SVM	RF
Accuracy	91.096	93.151	95.205	86.986	90.141	93.151
Precision	91.107	93.151	95.277	86.988	90.434	93.188
Recall	91.096	93.151	95.205	86.986	90.141	93.151
F1-Score	91.097	93.151	95.201	86.983	90.406	93.152

As shown in table 1, for all three ML algorithms experimented with, the firfefly algorithm was a better feature selection tool than the fruitfly algorithm as it has shown significant higher performance across the used performance metrics in comparison to the fruitfly algorithm. The firefly optimized algorithm with Random Forest gave the accuracy score of

95.205%. With reference to the bench paper, highest accuracy for KNN (90.83%) was obtained using RapidMiner algorithm with 10 features selected, while the KNN with firefly algorithm uses 13 features and obtained an accuracy of 91.096%. With regards to the SVM algorithm, the highest accuracy for SVM (90.83%) was obtained using RapidMiner algorithm with 10 features selected, while the SVM with firefly algorithm uses 14 features and obtained an accuracy of 93.151%. With regards to the RF algorithm, the highest accuracy for RF (93.12%) was obtained using RapidMiner algorithm with the complete dataset, while the SVM with firefly algorithm uses 18 features and obtained an accuracy of 95.205%. It can be inferred that the ML algorithms combined with the firefly algorithm have surpassed the corresponding counter parts in the bench mark paper.

Table 2: Selected Number of Features by Firefly and Fruitfly Algorithms

Optimization Algorithms	Firefly Algorithm			Fruitfly Algorithm		
Machine learning Algorithms	KNN	SVM	RF	KNN	SVM	RF
Number of Features	13	14	18	22	11	21

Table 2 summarizes the numbers of features selected by the fruitfly and firefly algorithms. The lowest number of features selected was 11 by the fruitfly algorithm using the SVM model while the highest number of features selected was 22 by the fruitfly algorithm using the KNN model. There were two relevant features selected by the two optimization algorithms for all the three ML algorithms used which are the TSH (mIU/L) and Follicle No. (R). There are also other significant features like Skin darkening, No.of aborptions, Fast food and endometrium.

The line graph below summarizes the performance metrics for the three different ML algorithms categorized by fruitfly and firefly optimization technique.

Figure 3:Performace matric linegraph

5. Conclusion

It is undeniable that ongoing attempts have been made in recent years to improve the accuracy of PCOS data based on machine learning techniques. That does not, however, exclude future efforts to raise the accuracy. In light of this development, the goal of this research is to rigorously evaluate the proposed multi layers methodology in order to increase the precision of PCOS's data prediction model. The approach is divided into seven parts, starting with the data source, exploratory data analysis, data preprocessing, data balancing, feature selection, model training and model evaluation. The comma separated value dataset is imported via Google Drive after being downloaded from the Kaggle repository. It initially included 541 datapoints and 45 features. The data was scaled using the minmax scaling method, and the synthetic minority oversampling technique was also used to balance the classes. The combination strategy for efficient feature selection utilized Fire Fly and Fruit Fly optimization techniques.

More so, owing to the sizeable numbers of learning models that the benchmark research had earlier employed, which include SVM, KNN, RF the proposed model of this research was trained based on fruitfly and firefly algorithms. The result of accuracy of RF with firefly performed best with an accuracy of 95.20%. However, this research experiment is still work in progress; interested researchers could in future, consider rule-based models and/or combination to test the results.

References

Alamoudi, A., Khan, I. U., Aslam, N., Alqahtani, N., Alsaif, H. S., Dandan, O. Al, Gadeeb, M. Al, & Bahrani, R. Al. (2023). A Deep Learning Fusion Approach to Diagnosis the Polycystic Ovary Syndrome (PCOS). *Applied Computational Intelligence and Soft Computing*, *2023*, 1–15. https://doi.org/https://doi.org/10.1155/2023/9686697

Alyasiri, O. M., Cheah, Y. N., Abasi, A. K., & Al-Janabi, O. M. (2022). Wrapper and Hybrid Feature Selection Methods Using Metaheuristic Algorithms for English Text Classification: A Systematic Review. *IEEE Access*, *10*, 39833–39852. https://doi.org/10.1109/ACCESS.2022.3165814

Bashir, S., Khattak, I. U., Khan, A., Khan, F. H., Gani, A., & Shiraz, M. (2022). A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches. *Hindawi Complexity*, *2022*, 1–12. https://doi.org/10.1155/2022/8190814

Bharati, S., Podder, P., & Mondal, R. H. (2022). Ensemble Learning for Data-Driven Diagnosis of Polycystic Ovary Syndrome. *Springer Nature Switzerland*, 1250–1259. https://doi.org/10.1007/978-3-030-96308-8

Denny, A., Raj, A., Ashok, A., Ram, M., & George, R. (2019). i-HOPE : Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques. *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 673–678. https://doi.org/10.1109/TENCON.2019.8929674

Dutta, P., Paul, S., & Majumder, M. (2021). An Efficient SMOTE Based Machine Learning

classification for Prediction & Detection of PCOS. *Research Square*, 1–14. https://doi.org/https://doi.org/10.21203/rs.3.rs-1043852/v1

Hassan, M. M., & Mizra, T. (2020). Comparative Analysis of Machine Learning Algorithms in Diagnosis of Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. *International Journal of Computer Applications*, *175*(17). https://doi.org/10.5120/ijca2020920688

Huang, X., Hong, L., Wu, Y., Chen, M., Kong, P., Ruan, J., Teng, X., & Wei, Z. (2021). Raman Spectrum of Follicular Fluid: A Potential Biomarker for Oocyte Developmental Competence in Polycystic Ovary Syndrome. *Frontiers in Cell and Developmental Biology*, *9*, 1–23. https://doi.org/10.3389/fcell.2021.777224

Hussein, K., & Karami, M. (2023). Association between insulin resistance and abnormal menstrual cycle in Saudi females with polycystic ovary syndrome. *Saudi Pharmaceutical Journal*, *31*(6), 1104–1108. https://doi.org/10.1016/j.jsps.2023.03.021

Inan, K., Ulfath, R., Alma, F., Bappee, F., & Hasan, R. (2021). Improved Sampling and Feature Selection to Support Extreme Gradient Boosting For PCOS Diagnosis. *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*, 1046–1050. https://doi.org/10.1109/CCWC51732.2021.9375994

Liu, L., Jiao, Y., Li, X., Ouyang, Y., & Shi, D. (2020). Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor. *Computer Methods and Programs in Biomedicine*, *196*, 1–8. https://doi.org/10.1016/j.cmpb.2020.105624

Lv, W., Song, Y., Fu, R., Lin, X., Su, Y., Jin, X., Yang, H., Shan, X., Du, W., Huang, Q., Zhong, H., Jiang, K., Zhang, Z., Wang, L., & Huang, G. (2022). Deep Learning Algorithm for Automated Detection of Polycystic Ovary Syndrome Using Scleral Images. *Frontiers in Endocrinology*, *12*(789878), 1–8. https://doi.org/10.3389/fendo.2021.789878

Madhumitha, J., Kalaiyarasi, M., & Sakthiya Ram, S. (2021). Automated Polycystic Ovarian Syndrome Identification with Follicle Recognition. *3rd International Conference on Signal Processing and Communication (ICPSC)*, 98–102. https://doi.org/10.1109/ICSPC51351.2021.9451720

Maheswari, K., Baranidharan, T., Karthik, S., & Sumathi, T. (2020). Modelling of F3I based feature selection approach for PCOS classification and prediction. *Journal of Ambient Intelligence and Humanized Computing*, 1–14. https://doi.org/10.1007/s12652-020-02199-1

Nandipati, S. C. R., XinYing, C., & Wah, K. K. (2020). Polycystic Ovarian Syndrome (PCOS) Classification and Feature Selection by Machine Learning Techniques. *Applied Mathematics and Computational Intelligence*, *9*, 65–74.

Nilofer, N. S., & Ramkumar, R. (2021). Follicles Classification To Detect Polycystic Ovary Syndrome Using Glcm And Novel Hybrid Machine Learning. *Turkish Journal of Computer and Mathematics Education*, *12*(7), 1062–1073.

Prapty, A. S., & Shitu, T. T. (2020). An Efficient Decision Tree Establishment and Performance Analysis with Different Machine Learning Approaches on Polycystic Ovary Syndrome.

2020 23rd International Conference on Computer and Information Technology (ICCIT), 19–21. https://doi.org/10.1109/ ICCIT51783.2020.9392666

Sreejith, S., Nehemiah, H. K., & Kannan, A. (2022). A clinical decision support system for polycystic ovarian syndrome using red deer algorithm and random forest classifier. *Healthcare Analytics*, *2*, 1–9. https://doi.org/10.1016/j.health.2022.100102

Thomas, N., & Kavitha, A. (2020). Prediction Of Polycystic Ovarian Syndrome With Clinical Dataset Using A Novel Hybrid Data Mining Classification . *International Journal of Advanced Research in Engineering and Technology (IJARET)*, *11*(11), 1872–1881. https://doi.org/10.34218/IJARET.11.11.2020.174

Thufailah, I. F., Adiwijaya, Wisesty, U. N., & Jondri. (2018). An implementation of Elman neural network for polycystic ovary classification based on ultrasound images. *International Conference on Data and Information Science*, 1–9. https://doi.org/10.1088/1742-6596/971/1/012016