Improving the Accuracy of PCOS' Data Prediction Model Based on Data Balancing and Multilayer Feature Selection Strategies.

Ayobami Ekundayo[1], Enesi Femi Aminu[2,*] and Uchenna Cosmas Ugwuoke[3]

[1,2,3]Department of Computer Science, School of Information & Communication Technology

[1,2,3]Federal University of Technology, Minna Nigeria

*Corresponding author

Abstract.

The use of machine learning algorithms to design a model for PCOS prediction or diagnosis in a recent time is attracting an impressive magnitude of research attentions. Of course, the rationale behind this development is not farfetched as the disease is common in women of reproductive age, which causes infertility. Also the algorithms as top notch techniques for classification, are highly promising. However, attention has to be paid to the raw dataset used for training the models; this is because the strategies adopted for feature engineering processes have direct proportionate effect on the robustness state of the model. Meanwhile, significant efforts have been advanced towards this development but that does not foreclose adoption of more strategies for better accuracy of the model. To this end, this research aims to adopt multilayers strategies in terms of class balancing and feature selection to improve the accuracy of the existing PCOS's data. The strategies include the use oversampling and LASSO-Pearson's correlation techniques for class balancing and feature selection respectively for the proposed ensemble random classifier based model. The 97.80% accuracy result of the proposed model outperforms the rest of the other seven models used in the benchmark work. Therefore, careful attentions have to be constantly advanced towards the process of feature engineering, which include data preprocessing, data exploratory, data balancing and feature selection strategies for optimal result.

Keywords: PCOS model, LASSO-Pearson technique, Oversampling, Data balancing, Classifier

1. Introduction

Preterm abortions, infertility, anovulation and other issues have a significant impact on the world's female population today. It has been noted that Polycystic Ovary Syndrome (PCOS), a disorder common in women of reproductive age, plays a significant role in the causes of infertility. PCOS affects more than five million women globally who are of reproductive age. It is an endocrine condition defined by alterations in the levels of female hormones and aberrant in male hormone synthesis (Denny *et al.,* 2019). One of the most prevalent reasons of female infertility is Polycystic Ovary Syndrome (PCOS), which affects many women of reproductive age and even persists well after the child bearing years (Inan *et al.,* 2021).

To this effect, some clinical efforts such as in vitro fertilization (IVF) have been developed and exploited to address this challenge. In order to show how devastated the challenge is, the increasing number of couples seeking in vitro fertilization (IVF) globally has made infertility a global health concern. After multiple IVF treatments, some couples are still without children as a result of inaccurate diagnosis. IVF for women comes with additional dangers and costs. One of the most prevalent types of endocrine disorders in women of reproductive age is Polycystic Ovary Syndrome (PCOS). Consequently, anovulation and infertility could occur from this problem. Therefore, the clinical and metabolic markers that serves as an early indicator of the disease are part of the diagnostic or prediction criteria (Mehrotra *et al.,* 2019). However, in order to ensure accurate and proper diagnosis of the disease indications; a lot of research efforts have been geared towards exploitation of the cutting edge technologies.

A ground breaking example of these technologies include the artificial intelligence (AI) specifically, machine learning models, which have the potentials to classify and predict the likelihood of PCOS disorder. Examples of these models include Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Multi-Layer Perceptron (MLP) and XGBoost (Abrar et al., 2010; Cinar & Koklu, 2019; Khan et al., 2021; Mustaqeem et al., 2017; Sathe & Adamuthe, 2021; Zhang et al., 2019). Of course, so many researchers in this field have been putting up several efforts in using these models to accurately diagnose the disease based on the given dataset. However, the level of accuracy has been hampered and distorted as a result of the imbalance nature of datasets or lack of robustness in decisively carrying out the process of feature engineering. In furtherance to this development, Inan, et al, 2021 employed SMOTE and ENN as two strategies for data sampling; also, Chi square and ANOVA tests were similarly used as techniques for feature selection. All these efforts were carried out on the dataset obtained from kaggle repository for PCOS diagnosis model. The researchers trained and evaluated the PCOS dataset on seven classifiers where in the final analysis, extreme gradient boosting classifier is adjudged to be the best with accuracy rate of 95.83%.

Furthermore, based on literature; there are other similar techniques for class balancing that have been constantly employed for training models. Resampling techniques (such as random oversampling or random under sampling), sample weight and SMOTE are part of the most popular techniques to achieve class balancing (Arora *et al.,* 2020; Bae *et al.*, 2021; Pecorelli *et al., 2019)*. Similarly, lasso regularization is one of the prominent feature selection techniques that have been increasingly used for training models (Afrin *et al.,* 2021; Ghosh *et al.,* 2021; Muthukrishnan & Rohini, 2016). However, the technique is not sufficiently efficient to carry out optimum selection of features. In view of this development, there have been constant efforts to extend the use of Lasso algorithm either by improving it by manipulating variables or hybridizing with other reliable techniques as noted in some literature (Chen *et al.,* 2019; Luo *et al.,* 2021; Zhang et al., 2021).

As a result of this feat, this research aims to improve the accuracy of the same dataset (both in features and data points) by employing the simple but reliable random over sampling data balancing strategy. Similarly, Least Absolute Shrinkage and Selection Operator (LASSO) regularization and Pearson's correlation feature engineering techniques are equally employed. Finally, ensemble random classifier is used as for the training of the PCOS' data

model and the accuracy performs better than the seven model classifiers used in the benchmark research. 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 Work

This section gives a brief related literature on Polycystic Ovary Syndrome and the application of machine learning models to proffer viable solutions in addressing the challenges of accurately predicting PCOS disorder in a given dataset. In the light of this, the literature of Denny *et al.*, (2019) aim at detecting and predicting PCOS at an early stage for an optimal and minimal but promising metabolic and clinical parameter. The research employed six traditional machine learning models which include but not limited to Naïve Bayes classifier method, Logistic Regression, K-Nearest Neighbor (KNN), and Random Forest classifier (RFC). In order to obtained optimal results, Principal Component Analysis (PCA) is used for the feature selection strategy. A total of 541 patient samples were collected for this study, 364 cases were PCOS negative and 177 cases of PCOS positive and 23 features were used which include the reports on TVS, hormone profile and patients way of life. Evaluation was done in terms of Accuracy, Precision, Sensitivity/Recall, Specificity, F1-Score. The RFC outperformed other learning models with an accuracy of 89.02%.

The research of Mehrotra *et al.*, (2019) also aim to develop an automated screening system that would help in the early detection of the disease. Two classical machine learning models were used in their work, these are Bayesian and Logistic Regression classifiers. Feature selection was done via two sample t-tests to select a significant subset of original feature. Nine parameters were considered; and a total of 200 patients were considered for the study among where 50 patients were normal and 150 suffer from the disorder. The overall accuracy of Bayesian classifier is 93.93% as compared with logistic regression i.e. 91.04%. Similarly, Bharati *et al.,* (2020) focused on diagnosis of PCOS data in women based on some models which include Extreme Gradient Boost, Random Forest, Logistic Regression and a hybrid of Random Forest and Logistic Regression. A Univariate Feature Selection algorithm was employed as the feature selection technique to identify the important features and find their scores. A total of 541 samples with 43 attributes were gotten from Kaggle repository. Thus, the hybridized model performed better than other models in terms of accuracy (for 10 feature: 90.01%, for 14 feature: 89.27%) and recall using 40-cross fold validation (91.01%).

Furthermore, the literature Cheng & Mahalingaiah, (2019) carried out evaluation performance of two machine learning text algorithms (rule base classifier and gradient boost tree model) in the classification of PCOM in pelvic ultrasound. Feature extraction was done via Porter's Snowball Algorithm. The result of the evaluation shows that rule base classifier slightly outperforms the gradient boost tree model with an accuracy of 97.6% while the latter has 96.1%. The two classifiers estimated prevalence of PCOM with our population ultrasounds to be about 44% for PCOM absent, 32% for PCOM unidentifiable and 24% for PCOM present.

Similarly, five learning models, which include RF, LR, DT, SVM and KNN along with SMOTE were used in the research of (Dutta *et al.,* 2021) who proposed to detect and predict PCOS in a give set of data. Principal Component Analysis (PCA) was used for feature extraction with a total of 541 records having 41 attributes obtained from the UCI repository. The SMOTE based LR performed best in terms of accuracy, training time, F1 score, recall, prediction and area under the ROC of 97.11%, 0.010 sec., 98%, 98%,98% and 95.6% respectively.

In addition, the literature of Madhumitha *et al.,* (2021) proposed SVM, LR and KNN also a stacking ensemble of these three algorithms were employed to identify follicles and edges in images. Gabor wavelet was used as their feature extraction technique; thus, the hybridized model outperformed other models in terms of accuracy, recall and precision. Also, an automated ovarian classification method for classifying ovary into normal or not in an ovarian ultrasound image was proposed by (Nilofer & Ramkumar, 2021). The researchers made use of ANN, KNN, SVM and IFFOA-ANN; while Grey -level co-occurrence matrix was used in the extraction of features. The IFFOA-ANN outperformed other machine learning models used in terms of accuracy with 97.50%.

Hassan & Tabasum, (2020) employed the learning algorithms which include Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression (LR), and Random Forest (RF) to diagnose PCOS based on clinical data of the patients. R-language was used for the implementation of algorithms, and RF performed best with an accuracy of 96%. The researchers thus submitted that further work can be done by making use of a larger size of data sets in the diagnosis of the disease. Also, the literature of Nandipati *et al.*,(2020) aim to know the features that play a role in the prediction of the PCOS disease and to determine which machine learning model performs better in the detection of the disease. A dataset consisting of a total of 541 women through survey from consultations from doctors and clinical examinations were considered in this study. SPSS V.2.0 was used to extract 8 features from a total of 23 features consisting of both clinical and metabolic parameters. The Synthetic Minority Oversampling Technique (SMOTE) was used as the preprocessing technique in this study. Seven machine learning models were employed to evaluate the disease. Consequently, Random Forest performed best with an accuracy of 93.12%. Future study suggests to build a better model for the PCOS data set using 10 features and complete attributes.

The use of this particular PCOS dataset of 540 data samples with 41 features (columns) from kaggle repository is very prominent for developing PCOS model. The research of Inan *et al.*, (2021) premised their diagnostic model on the dataset through the use of seven learning models. They are SVM, KNN, RF, MLP, NB, adaboost and XGBoost; SMOTE along with ENN and chi square with analysis of variance (ANOVA) tests were used as data sampling and feature selection techniques respectively. Based on the results, XGBoost outperformed other models with an accuracy rate of 95.83%.

In conclusion, based on the related literature reviewed in this research, the proposed PCOS's data prediction is promising as the accuracy result outperforms the related literature of the same dataset and those with close range.

3. The Methodology

The proposed methodology in this research is described as multi-stages approach, which ranges from data source to model evaluation. Figure 1 presents the detail stages involved along with the explanation.

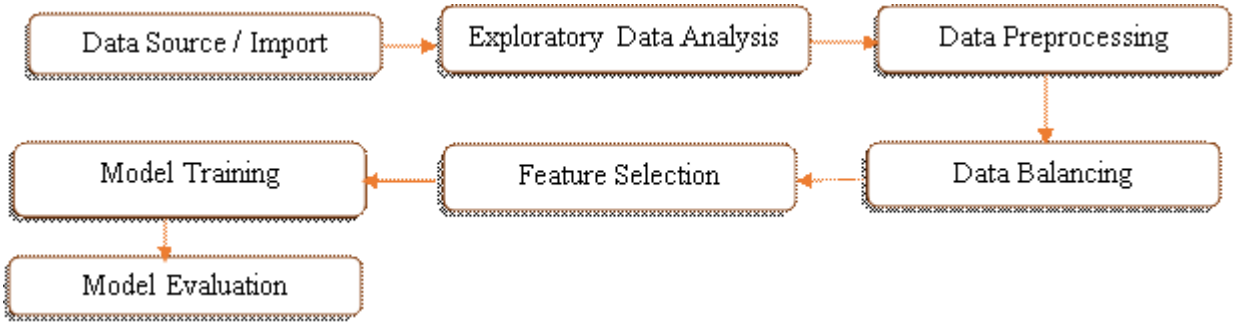

Figure 1: Multi-Stages Methodology

Following the proposed methodology as schematically presented by Figure 1, the dataset used in this experiment is obtained from kaggle repository, which is the same dataset used by the research of Inan *et al.*,(2021). The PCOS raw dataset consist of 45 features (attributes) and 541 data points (data samples); and they are imported from google drive since Google Collaboratory is the choice of environment for implementing the model. The dataset is in comma separated values (csv) format; and based on the data exploration and visualization, the dataset include 177 patients with the PCOS disorder and 364 with no disorder. That is, 177 Yes data points and 364 No data point respectively. However, with this analysis, it is obvious that the dataset is highly imbalanced; which can drastically reduce the overall prediction's accuracy or causes bias during predictions. More so, most of the features consist of floating point value with some features containing missing value that are not closely distributed. Therefore, data preprocessing such as data cleaning, data scaling and data balancing is required.

The data cleaning includes removing of inconsistency, missing value and irrelevant features such as patient file No, Patient ID from the dataset since they have zero impact on predicting PCOS status. Similarly, some features with non-floating values are converted into floating point values. This is to ensure data standardization (scaling) of features' value with highly distributed floating-point value. The Min-Max Scaling (value between 0 and 1) approach is adopted for this stage of the work. At this point, before features are selected, this research employs the technique of Random over Sampling for data balancing. This approach enables generation of data samples for the minority class (the 'Yes' class). Consequently, both classes would contain the same set of data points with 364 samples from each class. More importantly, this stage is necessary because if the classes of the data are not balanced, it would be difficult for the feature selection technique of Pearson correlation feature filtering to work properly.

The feature selection process includes extracting essential features or attributes that have high contribution to the final prediction class. Therefore, in this research Lasso Regularization and Pearson's Correlation techniques are considered. The raw dataset is first fed into the lasso regularization to extract important features by adding a penalty score or

threshold on all the features. Similarly, Pearson's correlation feature technique is used for selecting best feature for dataset with numeric input and output. It's used in selecting the best features that can predict the dependent variable (output). Finally, the preprocessed and selected features of the datasets are split into training set (for training the model) and testing set (for evaluating the model) in the ratio of 75% and 25% respectively. While, the training data is fed into the ensemble random forest classifier, the testing data are evaluated considering the metrics of precision, recall, f1-score and support.

Python 3.9.0 forms the experimental environment for the implementation of this methodology. It is important to mention that this research focuses on techniques of data balancing and feature selection to improve the accuracy of the PCOS model. This is better represented by the algorithmic design in Table 1.

Table 1: Optimum PCOS's Feature Subset Algorithm

Algorithm3.1*: Optimum Feature Subset Algorithm*
Input*: CSV Based PCOS Dataset*
Output*: Returns t_{ff}*
Parameters: *Python Packages and Classes such as edu.stanford.nlp.tagger.maxent.…, py.io.BufferedReader, py.io.FileReader (list); input PCOS data (t); preProcessFxn(); candidateTerms (t_c); featuresSubset (t_f) LASSO(t_c); minimalClass (c); maximalClass (n); randOverSamp(c); pearsCorr(t_f); optimumFeaturesSubs (t_{ff}); LSTM(t_{ff})*

Procedure:

```
1    Input t
2    forAll t preProcessFxn (t)
3        tc ← preProcessFxn (t)
4        LASSO(tc)
5        tf ← LASSO(tc)
6        while c != n
7            execute randOverSamp(c);
8            c++;
9        end
10       pearsCorr(tf)
11       tff ← pearsCorr(tf)
12       return tff
13       LSTM(tff)
14   end
```

From the algorithm as shown in Table 1, the raw PCOS dataset is inputted where each and all the inputted text is preprocessed to eliminate unwanted text; consequently, the text produced is known as candidate terms (t_c) as shown by steps 1 to 3. Thereafter, LASSO regularization techniques is employed as feature selection strategy where it acted as function on the t_c as input parameter. It thus produce result called feature subset (t_f) as presented from steps 4 to 5.

At this point, the second feature selection strategy, which is Pearson's correlation cannot be used due to the imbalance of the classes. From steps 6 to 9, attention has to be paid to the minimal class (c) by bringing to it a data resampling strategy herein named *randOverSamp* to populate the class. Thus, the algorithm execute the resampling method

randOverSamp(c) and the number of the minimal classes start to increase until it is equal to the number of maximal classes. From steps 10 to 13, the second feature selection function called *pearsCorr(t_f)* which accept the output of the first feature selection strategy as input is invoked. Finally, it return an output described in this research as optimum feature subset (t_{ff}) where 75% of the dataset is used for the LSTM based model training.

4. Results and Discussion

This section discusses the results of the proposed techniques for dataset balancing and the multilayer feature selection. Finally, the results of the PCOS prediction model using ensemble random forest classifier and comparison with the related classifiers are also presented. The appraoch for the results analysis is premised on the proposed methodology sequentially, and the optimum feature subset algorithm. Thus, Figure 2 which denotes the exploratory data analysis is presented.

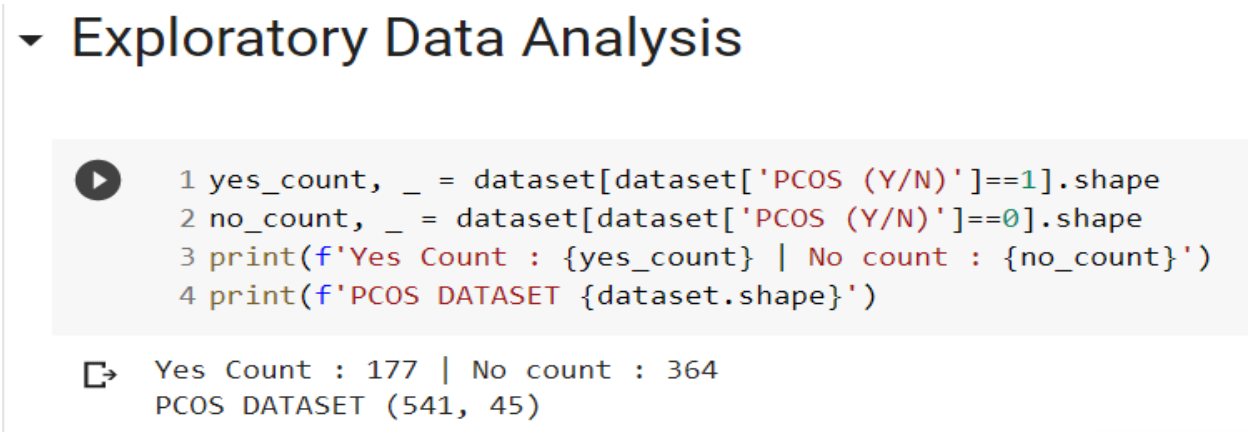

Figure 2: Exploratory of the Data

The essence of this Figure 2 that presents the exploratory view of the data analysis for each class count is to put forward a strong argument why there is need for a more strategic data balancing approach. From the Figure, patients with PCOS disorder are 177, while patients with no disorder are 364. This clearly establish the need to balance the data (class); thus, the strategy of random over sampling with imblearn. It randomly populate the minority class, which is the Yes class for the class to maintain balancing level. Furthermore, in order to have a well selected features, this research proposed the combined efforts of lasso regularization to extract very relevant features and Pearson's correlation technique for optimal feature selection. In view of this, the dataset is fed into the lasso regularization technique by continuously adding a threshold on all the features. Consequently, and based on experiments the best threshold identify during the series of trials is 0.004 as shown by Figure 3.

LASSO REGULARIZATION

```
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import SelectFromModel
# from sklearn import selecte

# set the regularization paramter c=1
logistic = LogisticRegression(C=1, penalty='l1', solver='liblinear', ran
model = SelectFromModel(logistic, prefit=True)

x_new = model.transform(X)
print(x_new)

select_colum = X.columns[X.var() > 0.004]
print('Number of column: ' , len(select_colum))
print('colums :  ' , select_colum)

# Dropped colums have value of all 0s, keep other columns
# select_columns = selected_features.columns[selected_features.var() != 
# select_colums
```

Figure 3: The Implementation of Lasso Regularization

With the threshold of 0.004 as shown in Figure 3 to implement the effective selection of features by the technique, the result of the implementation is presented by Figure 4.

```
Number of column:  33
colums :   Index([' Age (yrs)', 'Weight (Kg)', 'Height(Cm) ', 'BMI', 'Blood
       'Pulse rate(bpm) ', 'RR (breaths/min)', 'Hb(g/dl)', 'Cycle(R/I)',
       'Cycle length(days)', 'Marraige Status (Yrs)', 'Pregnant(Y/N)',
       'No. of aborptions', '  I   beta-HCG(mIU/mL)', 'II    beta-HCG(mIU/ml
       'Hip(inch)', 'Waist(inch)', 'Waist:Hip Ratio', 'AMH(ng/mL)',
       'PRL(ng/mL)', 'RBS(mg/dl)', 'Weight gain(Y/N)', 'hair growth(Y/N)',
       'Skin darkening (Y/N)', 'Hair loss(Y/N)', 'Pimples(Y/N)',
       'Fast food (Y/N)', 'Reg.Exercise(Y/N)', 'Follicle No. (L)',
       'Follicle No. (R)', 'Avg. F size (L) (mm)', 'Avg. F size (R) (mm)',
       'Endometrium (mm)'],
      dtype='object')
```

Figure 4: The Result of Lasso Regularization

With the implementation of Lasso algorithm, 33 features were selected out of the original features of 45. Based on the result, it was observed that the features obtained are yet to be optimal; hence, the need for further action is required, which necessitated for Pearson's correlation.

Pearson's correlation feature technique is used for selecting best feature for dataset with numeric input and output. It's used in selecting the best features that can predict the dependent variable (output). The PC feature filtering allows the option of selecting numbers of features that are highly significant in predicting the output. After series of trials, the best 24 features were found to yield the best result.

After series of feature engineering processes, which include preprocessing, features selection, and data balancing where 24 optimum features were obtained. The dataset is split

into 75% and 25% for training and testing respectively. The training set was fed into the random forest classifier tuning out an accuracy of 97.80% as shown by the model in Figure 5.

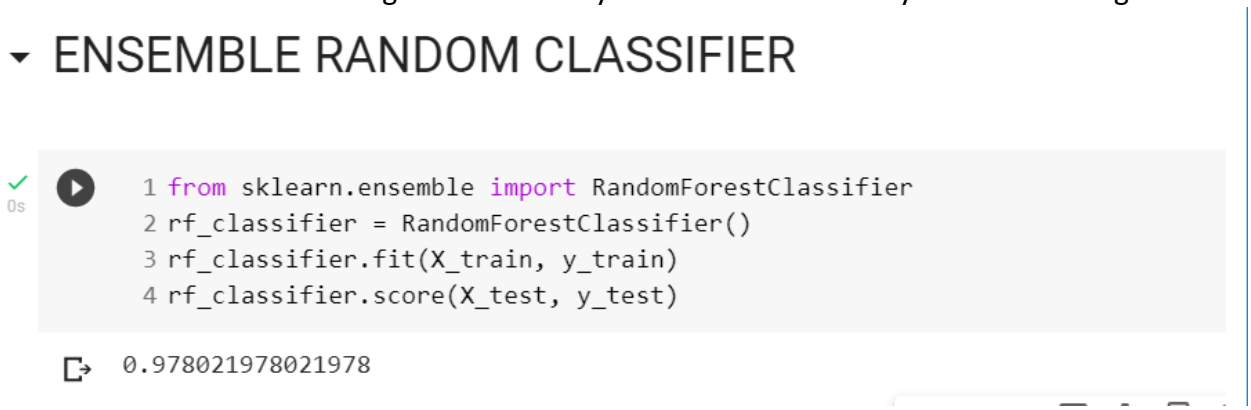

Figure 5: Accuracy Result of the Model

Based on figure 5, an accuracy of 97.80% is achieved using the ensemble random forest classifier. Similarly, the result of the model evaluation for precision, recall and f1-score were equally obtained as 99%, 96% and 98% respectively. Consequently, the accuracy result obtained was compared with the seven models used in the benchmark literature based on the same dataset. Table 1 presents the comparison results.

Table 1: Model Result Comparison

	Benchmark Models							***Adopted Model***
Models	*SVM*	*KNN*	*RF*	*AdaB*	*XgB*	*NB*	*MLP*	*Ensemble RF*
Accuracy (%)	*93.75*	*87.50*	*92.71*	*92.71*	*95.83*	*93.75*	*92.71*	*97.80*

Obviously, as shown by Table 1 the result of ensemble random classifier as the adopted model performs better than the seven models that were previously considered in the benchmark paper. For better representation, Figure 6 presents the summary result graphically.

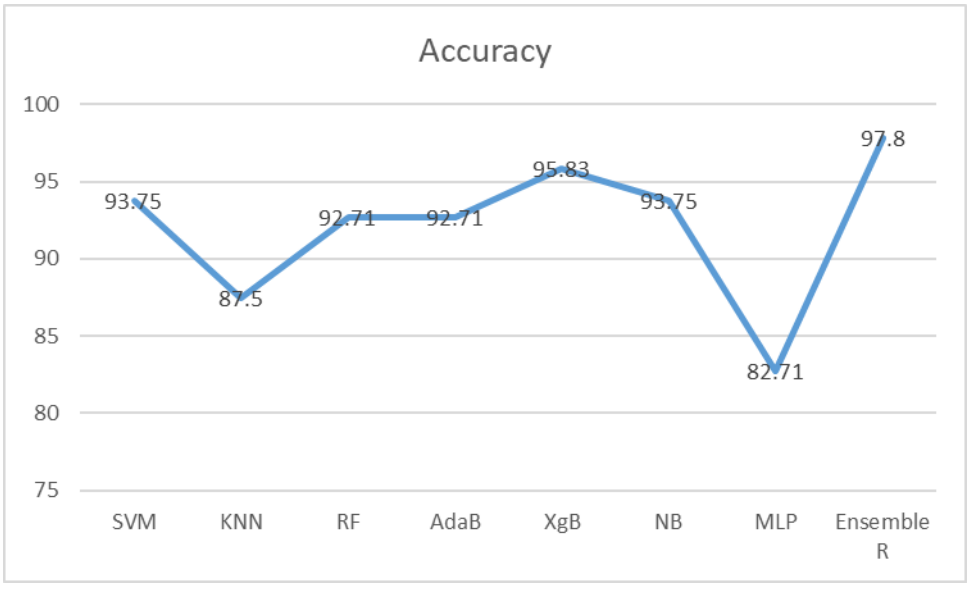

Figure 6: Model Accuracy Graphical Representation

As shown by the figure, the adopted ensemble random classifier (ensemble R) has the highest accuracy of 97.80% over the remaining models earlier used to classify the same dataset from the benchmark literature.

5. Conclusion

It is unarguable fact that in recent time there have been constant efforts geared towards having a better result accuracy for PCOS' data based on machine learning models. However, that does not foreclose a further attempt to improve on the accuracy. In view of this development, this research aims to improve the accuracy of PCOS's data prediction model by rigorously considering the proposed multi layers methodology. The methodology consist of seven stages ranges from data source, exploratory data analysis, data preprocessing, data balancing, feature selection, model training and model testing. The comma separated value dataset, which originally consist of 541 datapoints and 45 features is obtained from kaggle repository and imported via google drive. The data is scaled via min-max scaling approach; and random over data sampling was equally employed for class balancing. Lasso regularization and Pearson's correlation were the combined strategy for effective feature selection.

More so, owing to the sizeable numbers of learning models that the benchmark research had earlier employed, which include SVM, KNN, RF, MLP, NB, adaboost and XGBoost; the proposed model of this research was trained based on ensemble random classifier. The result of accuracy which is 97.80% performs better than all the other seven training models. However, this research experiment is still work in progress; interested researchers can exploit other robust feature engineering mechanisms to ascertain further, the reliability of the mechanisms deployed on this research. This is because researchers hope to employ word2vec or WordNet as data augmentation strategy to cushion balance for the given dataset to derive contextual knowledge. In other words, this anticipation would pave way for the deployment of robust networks (deep learning) models against the use of traditional learning models. Furthermore, in future it is intended to employ the mechanisms on other related datasets to validate their strengths.

References

Abrar, I., Ayub, Z., Masoodi, F., & Bamhdi, A. (2010). A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset. *Proceedings of the International Conference on SMart Electronics and Communication (ICOSEC 2020)*, 919–924.

Afrin, S., Javed Mehedi Shamrat, F. M., Nibir, T. I., Muntasim, M. F., Moharram, M. S., Imran, M. M., & Abdulla, M. (2021). Supervised machine learning based liver disease prediction approach with LASSO feature selection. *Bulletin of Electrical Engineering and Informatics*, *10*(6), 3369–3376. https://doi.org/10.11591/eei.v10i6.3242

Arora, M., Dhawan, S., & Singh, K. (2020). Data Driven Prognosis of Cervical Cancer Using Class Balancing and Machine Learning Techniques. *EAI Endorsed Transactions on Energy Web*, *7*(30), 1–9. https://doi.org/10.4108/eai.13-7-2018.164264

Bae, S. Y., Lee, J., Jeong, J., Lim, C., & Choi, J. (2021). Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular

fingerprints. *Computational Toxicology*, *20*, 1–6. https://doi.org/10.1016/j.comtox.2021.100178

Bharati, S., Podder, P., & Mondal, M. R. H. (2020). Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms. *2020 IEEE Region 10 Symposium (TENSYMP)*, 1486–1489.

Chen, S., Zhang, Y., Ding, C. H. Q., Zhang, J., & Luo, B. (2019). Extended Adaptive Lasso for Multi-class and Multi-label Feature Selection. *Knowledge-Based Systems*, 1–27. https://doi.org/10.1016/j.knosys.2019.02.021

Cheng, J. J., & Mahalingaiah, S. (2019). Data mining polycystic ovary morphology in electronic medical record ultrasound reports. *Fertility Research and Practice*, *5*(13), 1–7. https://doi.org/10.1186/s40738-019-0067-7

Cinar, I., & Koklu, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(3), 188–194. https://doi.org/10.1039/b000000x

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*. 1–14. https://doi.org/10.21203/rs.3.rs-1043852/v1

Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. M. J. M., Ignatious, E., Shultana, S., Beeravolu, A. R., & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and lasso feature selection techniques. *IEEE Access*, *9*, 19304–19326. https://doi.org/10.1109/ACCESS.2021.3053759

Hassan, M. M., & Tabasum, M. (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), 1–13. https://doi.org/10.5120/ijca2020920688

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

Khan, B., Naseem, R., Shah, M. A., Wakil, K., Khan, A., Uddin, M. I., & Mahmoud, M. (2021). Software Defect Prediction for Healthcare Big Data : An Empirical Evaluation of Machine Learning Techniques. *Journal of Healthcare ENgineering*, 1–16. https://doi.org/https://doi.org/10.1155/2021/8899263

Luo, M., Wang, Y., Xie, Y., Zhou, L., Qiao, J., Qiu, S., & Sun, Y. (2021). Combination of feature selection and catboost for prediction: The first application to the estimation of aboveground biomass. *Forests*, *12*(216), 1–21. https://doi.org/10.3390/f12020216

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

Mehrotra, P., Chatterjee, J., Chakraborty, C., Ghoshastidar, B., & Ghoshdastidar, S. (2019). *Automated Screening of Polycystic Ovary Syndrome using Machine Learning Techniques*. *2*(1), 1–5.

Mustaqeem, A., Anwar, S. M., Majid, M., & Khan, A. R. (2017). Wrapper method for feature selection to classify cardiac arrhythmia. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 3656–3659. https://doi.org/10.1109/EMBC.2017.8037650

Muthukrishnan, R., & Rohini, R. (2016). LASSO:A Feature Selection Technique In Predictive Modeling For Machine Learning. *2016 IEEE International Conference on Advances in Computer Applicatio s (ICACA)*, 18–20.

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.

Pecorelli, F., Nucci, D. Di, & Lucia, A. De. (2019). On the Role of Data Balancing for Machine Learning-Based Code Smell Detection. *MaLTeSQuE '19*, 19–24. https://doi.org/https://doi.org/10.1145/3340482.3342744

Sathe, M. T., & Adamuthe, A. C. (2021). Comparative study of supervised algorithms for prediction of students' performance. *International Journal of Modern Education and Computer Science*, *13*(1), 1–21. https://doi.org/10.5815/ijmecs.2021.01.01

Zhang, B., Qi, S., Monkam, P., Li, C., Yang, F., & Yao, Y. D., & Qian, W. (2019). Ensemble learners of multiple deep CNNs for pulmonary nodules classification using CT images. *IEEE Access*, *7*, 110358–110371.

Zhang, S., Zhu, F., Yu, Q., & Zhu, X. (2021). Identifying DNA-binding proteins based on multi-features and LASSO feature selection. *Biopolymers*, *112*(2), 1–11. https://doi.org/10.1002/bip.23419