INVESTIGATING THE WORKABILITY OF RANDOM FOREST IN POLYCYSTIC OVARY SYNDROME DETECTION: A PROPOSED FRAMEWORK.

A. Ekundayo*, E. F. Aminu, J. K. Alhassan, S. A. Adepoju, O. A. Ojerinde
Department of Computer Science, Federal University of Technology, Minna, Niger State, Nigeria.
*Email: a.ekundayo@futminna.edu.ng

ABSTRACT

Recently, there has been a significant increase in the quantity of research focusing on the use of machine learning techniques to develop a model for PCOS detection. Given that the illness frequently affects women who are of reproductive age and results in infertility, the reasoning for this development is understandable. PCOS symptoms can fluctuate over a patients' lifetime, making it difficult to diagnose. Irregular menstrual cycles, acne, hirsutism, and alopecia are common symptom, but their severity and presence can vary widely among patients. This study proposes investigating the workability of Random Forest (RF) in Polycystic Ovary Syndrome detection. The study utilized five models of random forest based on these parameters; estimators, bootstrap, depth, maximum features, maximum leaf nodes which was examined during the experimentation to ascertain the viability of the model. Model 3 of the random forest outperformed other models with an accuracy of 87.96% at 300 estimators, a true bootstrap, zero depth, maximum features is square root, and zero maximum leaf nodes.

KEYWORDS: Polycystic Ovary Syndrome (PCOS), Random Forest (RF), Bootstrap, Depth, maximum features, maximum leaf nodes.

DOI: 10.63748/FETiCON2025.v3i2p30

1. INTRODUCTION

Stein-Leventhal Syndrome, often known as Polycystic Ovarian Syndrome, is an endocrine disorder that affects 5-10% of women of reproductive age. (Wang *et al.,* 2022; Roy & Alvi, 2022; Chen *et al.*, 2023; Harish *et al.*, 2023 ;Mojahed *et al.,* 2023). Today's female population is significantly impacted by problems such as preterm abortions, infertility, anovulation, and other factors. Infertility have been shown to be significantly influenced by the condition known as Polycystic Ovary Syndrome (PCOS), which is frequent in women of reproductive age (Ekundayo *et al.*, 2022). Almost 60-70% of women with PCOS have hypothyroidism, 60-80% have raised testosterone levels, and 70-80% have metabolic problems such as obesity and an increased BMI (Hassan & Mizra, 2020).

Patients suffer from physical and psychological difficulties such as menstrual disruption, infertility, hirsutism, acne, obesity, insulin resistance, hypertension, mental problems, lipid disorders, diabetes, cancer, hair loss, metabolic syndrome, excessive levels of androgen, cardiovascular disease, bipolar disorder, hormonal imbalance, sexual dysfunction, extreme weight gain, depression, anxiety disorders are all symptoms of the chronic and diverse endocrine condition known as Polycystic Ovary Syndrome (Kaushik and Mishra, 2023; Ganie *et al.,* 2023; Wu *et al.,* 2023; Mogos *et al.,* 2024; Khushal and Fatima, 2024). A variety of endocrine and metabolic disorders that impair fertility are present in Polycystic Ovary Syndrome, a condition with an unclear cause. In addition to having trouble getting pregnant, PCOS women are more likely to miscarry, develop gestational diabetes, preeclampsia, pregnancy-induced hypertension, and greater fetal morbidity than other women. The pancreatic hormone insulin, widely known for its crucial function in glucose metabolism, also plays a key part in reproduction. PCOS patients with concomitant insulin resistance (and resulting hyperinsulinemia) experience poor pregnancy outcomes and problems with fertility (Ionescu *et al.,* 2023; Rahman *et al.,* 2024).

Several clinical techniques, such as in vitro fertilization (IVF), have been developed and employed to get around this problem. The increasing number of couples seeking in vitro fertilization (IVF) globally has made infertility a global health concern, highlighting the devastation caused by this issue. Due to incorrect diagnosis, some couples have multiple IVF cycles and are still childless. There are more dangers and costs associated with IVF for women.

Polycystic Ovary Syndrome (PCOS), a complicated disease with a range of phenotypes involving problems in metabolism, reproduction, and endocrine function, is one of the most common illnesses affecting premenopausal women (Hussein & Karami, 2023). However, in order to ensure accurate and suitable identification of disease indicators, various research projects have focused on the use of cutting-edge technologies based on artificial intelligence and machine learning.

Artificial intelligence (AI), particularly machine learning models, is a pioneering example of such technologies. They have the ability to categorize and forecast the probability of PCOS condition. A few of these models are Support Vector Machine (SVM), XGBoost, Logistic Regression (LR), Adaboost (AD), Gradientboost (GB) Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), and Multi-Layer Perceptron (MLP) (Kumari *et al.*, 2023 ; Mou *et al.,* 2023; Prasher *et al.,* 2023; Jaiswal *et al.,* 2024 ; Kumar *et al.,* 2024). Of course, numerous scholars in this field have attempted on multiple occasions to apply these models to efficiently detect the condition given the facts at hand. Unfortunately, because of the unbalanced nature of the datasets, the precision has been distorted and constrained.

This study will look into the practicality of employing random forest to diagnose polycystic ovarian syndrome. This would be achieved by severally tuning the hyperparameters of the dataset obtained from the work of Sreejith *et al.*, (2022) via kaggle repository. The experiment to conduct the viability of random forest in this given dataset leverages on five parameters that is, bootstrap, depth, maximum features, maximum leaf nodes, and more importantly, the estimators. The choice of random forest is motivated by Simi *et al.,* 2017; Mehr & Polat, 2021; Sreejith *et al.,* 2022; Tiwari *et al.*, 2022; Mridul *et al.,* 2024 in these literatures it has been proven that random forest is effective in terms of reducing overfitting and improves overall accuracy, suitable for datasets with many features, less sensitive to noisy data and can handle missing values in data.

In this research, Random Forest assists clinicians in making faster and more accurate diagnoses, thereby reducing the time and cost associated with the diagnosis. Five models of random forest based on the five parameter of estimators, bootstrap, depth, maximum features, and maximum leaf nodes are examined during the experimentation to ascertain the viability of the model. 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. LITERATURE REVIEW

Based on literatures, there are different approaches to PCOS prediction. However, machine learning approach has been identified to have a significant contribution in mitigating the flaws associated with PCOS prediction based on other approaches. In view of this development, this review article has investigated systematically the application of optimization algorithms to predict PCOS but considering textual datasets in this context. Several approaches such as optimization algorithms to aid the prediction of the disease at an early stage have been reported in literatures. 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 random forest classifier. The RDA was utilized by researchers to identify the best characteristics and the random forest was used to assess them. From the kaggle repository, a total of 541 data points with 37 instance were obtained. The preprocessing subsystem handles the dataset's outliers and noise using the z-score statistical metric. The evaluation matrices used in this study are accuracy, specificity and sensitivity.

Early detection and prognosis of PCOS are extremely important Denny *et al.*, (2019) used a sample size of 541 to detect and predict PCOS at an early stage, focusing on an ideal and minimal but promising metabolic and clinical parameter.. Six traditional machine learning models were used which are Naïve Bayes classifier method, Logistic Regression, K-Nearest Neighbor (KNN), Classification and Regression Tree (CART), Random Forest classifier, Support Vector Machine (SVM). The RFC outperformed other learning models used in terms of accuracy (89.02%).

Out of the 541 samples gotten from the city of Thrissur 364 cases were non-PCOS and 177 cases had PCOS. 23 features were listed out of which 8 features were finally considered after PCA (Principal Component Analysis) was used for feature transformation. Spyder Python was used for model formation, HTML with SQL for designing a proper interface for inputting of patient data into the system and PCOS status is being obtained as an output. SPSS V22.0 was used for establishing the relevance of the features. Further studies can be done on impact of PCOS on preterm abortions, effects of vitamin D on PCOS patients, attempts to unveil the number of lean PCOS patients. However, it was admitted by the researchers that the accuracy obtained in this study is relatively low compared to other literatures reviewed in this study a better accuracy can be obtained by enhancing the weight parameters of classifiers performance.

The work of Prapty & Shitu, (2020) aim at detecting the presence of PCOS in a patient by developing an efficient decision tree and also by comparing the performance analysis of different traditional machine learning models. 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. The researchers employed four different traditional machine learning models in this study which are SVM,RF,KNN and Naïve classifiers. Out of these algorithms Naïve classifiers and Random Forest performed best with accuracies of 93% and 93.5% respectively. This study considered Accuracy, Precision, Recall and F1 score as the confusion matrix employed.

The literature of Inan *et al.,* (2021) aim at integrating XGBoost algorithm to classify future instances of PCOS data. A total of 540 samples with 41 features was gotten from Kaggle repository. Seven machine learning algorithms such as: SVM, KNN, RF, MLP, NB, adaboost and XGBoost were used in this study. It can be deduced that XGBoost outperformed other ML algorithms used in terms of accuracy, recall, precision and F-score for both PCOS and non-PCOS cases. SMOTE and and ENN (Edited Nearest Neighbor) was used in this study as the technique for data preprocessing. Further studies can be done by including CNN in an optimized form to improve our model and also an extensive hyper parameter tunning of ML models and more improved feature selection can be worked on to obtain a better performance. However, the researchers equally admitted that the size of the sample data set used in this research is relatively small.

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 sedven machine learning techniques utilized 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). Future research recommends using 10 features and all available attributes to create a better model for the PCOS data set. Rapidminer, cannot however, be employed as a general rule in machine learning algorithms, it was also acknowledged, as performance depends on dataset, sampling and preprocessing procedures.

Bharati *et al.,* (2020) focuses on the data driven diagnosis of PCOS in women. From the Kaggle library, a total of 541 samples with 43 attributes were obtained. Different machine learning models, including, GB, RF, LR and a hybrid of RF and LR were used. The feature selection method utilized in this study to determine the importance of the characteristics and their scores was a Univariate Feature Selection algorithm. Using 40-cross fold validation, the hybridized model outperformed other models in terms of recall and accuracy. Data preprocessing steps in this study include data labeling and encoding non-numeric columns. The classification accuracy of machine learning models can be improved in the future research by creating a variety of hybrid techniques. The literature of Raut *et al.,* (2022) literature aims to help in early identification and prediction of PCOS therapy from an optimal and minimal collection of criteria that have been statistically examined.

It was reported that the researchers made use of seven machine learning models (RF, DT, SVC, LR, K-NN, XGBRF, CatBoost classifier). The data set used in this study was gotten from Kaggle repository. The platform used for the machine learning models implementation in this study is the Jupyter Notebook and python programming language. The data used in this study was cleaned using the Pandas and Numpy2. Out of all the seven machine learning classifiers used in this study the CatBoost classifier performed best with an accuracy of 92.64%.

The literature of Duttal *et al.*, 2021 aim at detecting and predicting PCOS using machine learning models and SMOTE. 541 records having 14 attributes were gotten from the UCI repository. Five machine learning models (RF, LR, DT, SVM and KNN) were used in this study. Principal Component Analysis (PCA) was used for feature extraction. Data was preprocessed using the standard scalar technique and SMOTE was used as the optimization technique. The SMOTE based LR performed best in terms of accuracy, training time, F1 score, recall, prediction and area under the ROC (97.11%, 0.010 sec., 98%, 98%,98% and 95.6%). Future studies can be done by embedding IOT technology with the proposed model that so patient's health parameters can be incidentally call the shots for developing a compelling health care system. However, the researchers admitted that the size of dataset used in the course of the study is small.

Nowadays, Polycystic Ovary Syndrome (PCOS) affects many women of reproductive age, making it a prevalent concern. It is a hormonal disorder that causes irregular, delayed, or absent menstrual cycles in the female body. However, the study of Rahman *et al.*, (2024) aim to employ machine learning models to identify patterns in this disorder. The dataset used for this research contains records for 541 patients. The information learned is then inputted into various algorithms to assess accuracy, specificity, sensitivity, and precision using different ML models, such as Logistic Regression (LR), Decision Tree (DT), AdaBoost (AB), Random Forest (RF), and Support Vector Machine (SVM). The research utilized the Mutual Information model for feature selection and compared the models to determine the most accurate one. The pitfall of this study is that the data set used is obtained from a single source. Early detection and treatment are critical for avoiding long-term problems from polycystic ovarian syndrome. The study of Kumar *et al.*, (2024) proposes a thorough data preprocessing, feature engineering, and exploratory data analysis (EDA). The refined dataset was used to classify PCOS using several default machine learning (ML) techniques such as LR, LDA, GNB, SVM, XGB, DT, AB, RF, and KNN. To improve the model's performance, this study combined all of the ML models with Particle Swarm Optimization (PSO). Remarkably, the proposed LR+PSO model attained the best accuracy of 96.30%, indicating excellent competency with an 80:20 train-test ratio. It increased sensitivity to 94.44%, indicating better detection of positive cases, while maintaining the greatest specificity (97.22%) and accuracy (94.44%) among other models. Overfitting is possible, especially if the model is well tuned for training data but performs poorly on unseen data, which is a common problem with complex swarm intelligence algorithms.

Polycystic Ovary syndrome (PCOS) is the most common serious health problem in women worldwide. The study of Kumari *et al.*, (2023) proposes a SmS: SMOTE-Stacked hybrid model, for the early diagnosis of PCOS. This study combined the ensemble technique 'stacking' with oversampling technique Synthetic minority oversampling technique. In the Stacking ensemble technique, training is performed at two levels: base level and meta level. This study utilized six machine learning classifiers LR, SVM, DT , RF, NB, and AdaB as base leaners at the base level and consider each of these classifiers separately at the *meta*-level to implement SmS hybrid models: Stack-LR, Stack-SVM, Stack-DT, Stack-RF, Stack-NB and Stack-AdaB. Standardization, SMOTE and Backward Feature Elimination (BFE) are performed before training the classifiers. The performance of the individual classifiers was evaluated and hybrid models on PCOS dataset (Imbalanced) taken form the Kaggle Repository and observe that Stack-AdaB shows the most encouraging results. Its key drawback is its dependency on SMOTE-based synthetic data augmentation, which can introduce noise into the dataset even though it is useful for addressing class imbalance. The model's capacity to effectively generalize to real-world situations may be hampered by this noise.

3. METHODOLOGY

The proposed methodology in this study is characterized as a multi-stage strategy that starts with the data source and ends with the model validation. The explanation and the stages involved are shown in Figure 1.

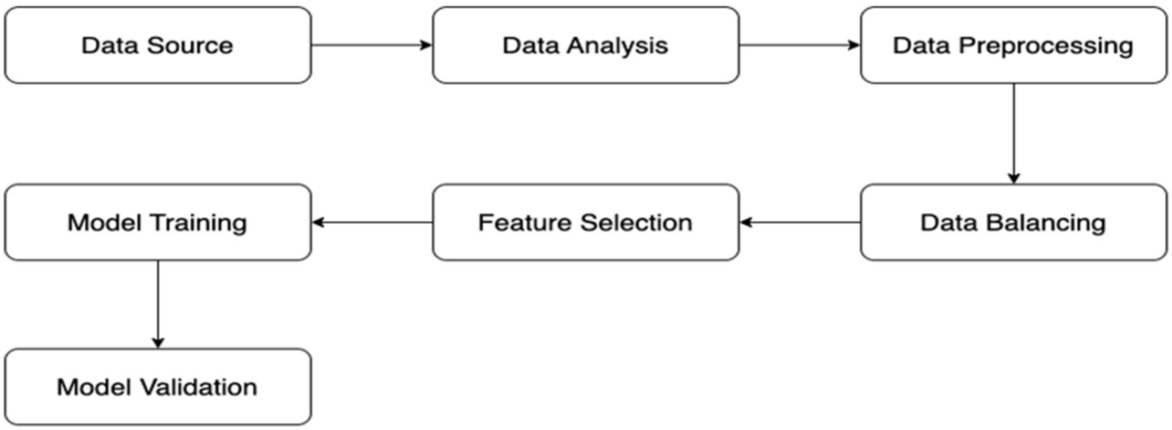

Figure 1: Multi-Stages Methodology

The multi-stage approach adopted in this study began with data collection which was obtained from the Kaggle Repository with a size of five hundred and thirty-eight (538) data points and forty-one (41) features and ended with model validation. Basically, this study focuses on the model training. Figure 2 depicts the numbers of random forest's models considered in this work. Figure 2 depicts the five models used for experimenting the random forest (RF) and the number of estimators used by each of the models. Models 1 and 3 made used of 300 estimators each. Models 2 and 4 employed 200 estimators each. Model 5 made use of 100 estimators. The depth is another critical parameter considered in this study during the hyperparameter tuning of the algorithm. Models 1 and 3 had a zero depth, models 2, 4, and 5 had a depth of 20 respectively. All the models had a true bootstrap, maximum leaf nodes for all models are none, and the maximum features used are square root. The algorithmic representation for the best model of RF is shown in Table 1.

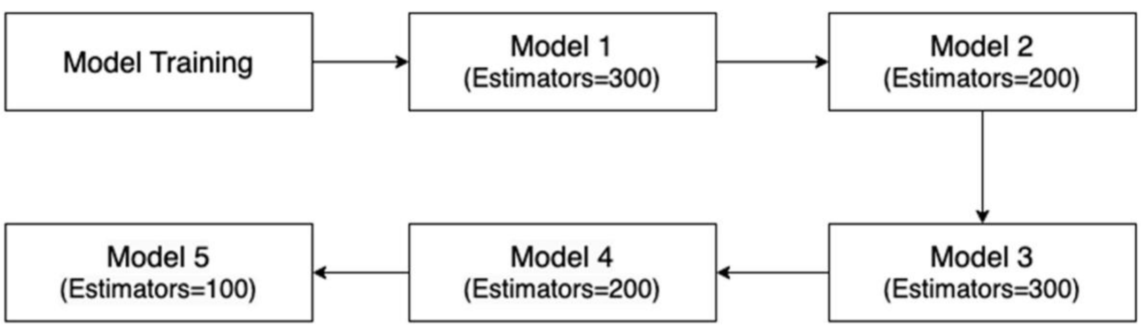

Figure 2: Random Forest and the Number of Estimators Used

From the algorithm as shown in Table 1, the raw PCOS dataset is inputted where each and all of the inputted text is preprocessed to eliminate unwanted text; consequently, the text produced is known as candidate terms (t_c) as shown by lines 1 to 3. Thereafter, the candidate terms (t_c) were fed into the enhanced red deer algorithm, and the selected feature text (t_f) was produced as shown in line 4. Initialization of the counter (i) and number of models (n) was done as shown in line 5. Then a while loop was introduced in line 6. Then a mathematical model to calculate the model output was formulated in line 7, where δ_r is the bootstrap, δ_k is the number of estimators, and $\int_{acc}$ is the accuracy. If the model counter is greater than M_{i+1} it prints the best model as shown in line 9. From lines 11 to 13 the if, while and for loops were terminated.

Table 1: Algorithmic Design for a Five Based Model of Random Forest.

Algorithm 1: A Five Based Model of Random Forest
Input: t
Output: Returns M_b
Parameters: Python Packages and classes such as edu. Standford.nlp.tagger.maxent…., py.io.BufferedReader,py,io.FileReader (list); PCOS dataset (t), Preprocess Fxn ();Accuracy (a), candidate term (t_c), best model (M_b), counter (i), number of models (n),seledted feature text (t_f), model output (M_o), bootstrap (δ_b), estimators (δ_k),max_leaf (m_l),max_leaf_node(m_n), depth (m_d), enhanced red deer algorithm (eRDA), accuracy (acc), model counter (m_i)

1. *Input t*
2. *For all t preprocess Fxn (t)*
3. $t_c \longleftarrow$ *preprocessTxn (t)*
4. $t_f \longleftarrow$ *eRDA(t_c)*
5. *i=0, n= 5*
6. *While (i< = n)*
7. $M_o = \delta_b / \delta_k \,.\, \int_{acc}$
8. *If $m_i > m_{i+1}$ Then*
9. *Print M_b*
10. *Else If i++*
11. *End If*
12. *WEND*
13. *End*

4. SUMMARY AND DISCUSSION

This section discusses the results obtained from the experiment carried out on random forest based on the five major parameters that is, estimators, bootstrap, depth, maximum features, and maximum leaf nodes. Table 2 shows the results of hyperparameter tuning for the five models of the algorithm.

Table 2: Random Forest' Models Results

Model	Estimators	Bootstrap	Depth	Maximum features	Maximum Leaf Nodes	Accuracy (%)
Model 1	300	True	None	Square Root	None	87.04
Model 2	200	True	20	Square Root	None	86.11
Model 3	300	True	None	Square Root	None	87.96
Model 4	200	True	20	Square Root	None	87.07
Model 5	100	True	20	Square Root	None	87.04

Table 2 shows the result of the random forest models. Models 1 and 3 used a total of 300 estimators, had a true bootstrap, they had no depth and no maximum leaf nodes and the maximum features used by both models is square root. Models 2 and 4 used 200 estimators, they both had a true bootstrap, a depth of 20, no maximum leaf nodes and a maximum features of square root. Model 5 used a total of 100 estimators, a true bootstrap, a depth of 20, no maximum leaf node, and a maximum feature of square root. Models 1, 2, 3, 4, and 5 achieved and accuracy of 87.04%, 86.11%, 87.96%, 87.07%, and 87.04% respectively. However, surprisingly model 3 gains a better accuracy compared to the previous models. This gain would have been attributed to the number of estimators but conversely, the accuracy of model 1 does not suggest it; this is because both models used the same number of estimators. Thus, model 3 outperformed the other models. Figure 3 shows the number of estimators used each of the models. Figure 3 shows the number of estimators used by each model of the algorithm. At model 1 and model 3 a total of 300 estimators were used.

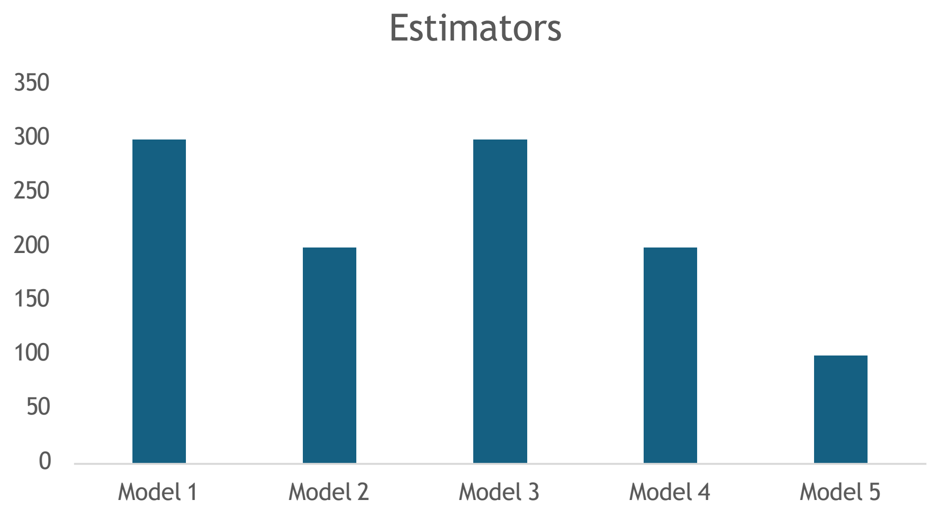

Figure 3: The Number of Estimators used by each Models of the Algorithm

Models 2 and 4 used 200 estimators each and model 5 used 100 estimators. Each of the models presents it result in confusion matrix, hence for emphasis sake; Figure 4 shows a graph of the accuracies obtained by each of the models. Figure 4 depicts the accuracies obtained by each random forest models. Models 1 aand 5 had an accuracy of 87.04%, model 2 had an accuracy of 86.11%, model 4 had an accuracy of 87.07%, and model 3 achieved an accuracy of 87.96%. The model 3 outperforms other models of random forest in terms of accuracy. Figure 5 shows the random forest's confusion matrix model 3 being the best model that reported the highest accuracy.

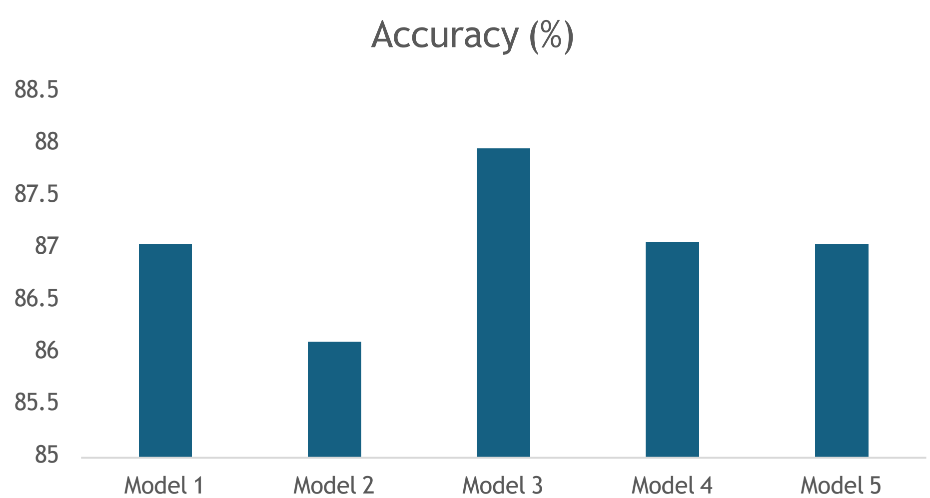

Figure 4: Accuracies obtained by each of Random Forest Models

Figure 5 shows the random forest confusion matrix for model 3. The experiment is conducted to determine the performance of random forest which leverages on five parameters that is, estimators, bootstrap, depth, maximum features, and maximum leaf nodes. Model three is the model that performed best in terms of accuracy 87.96% utilizing 300 estimators, a true bootstrap, nil depth, maximum features as square root, and a nil maximum leaf node.

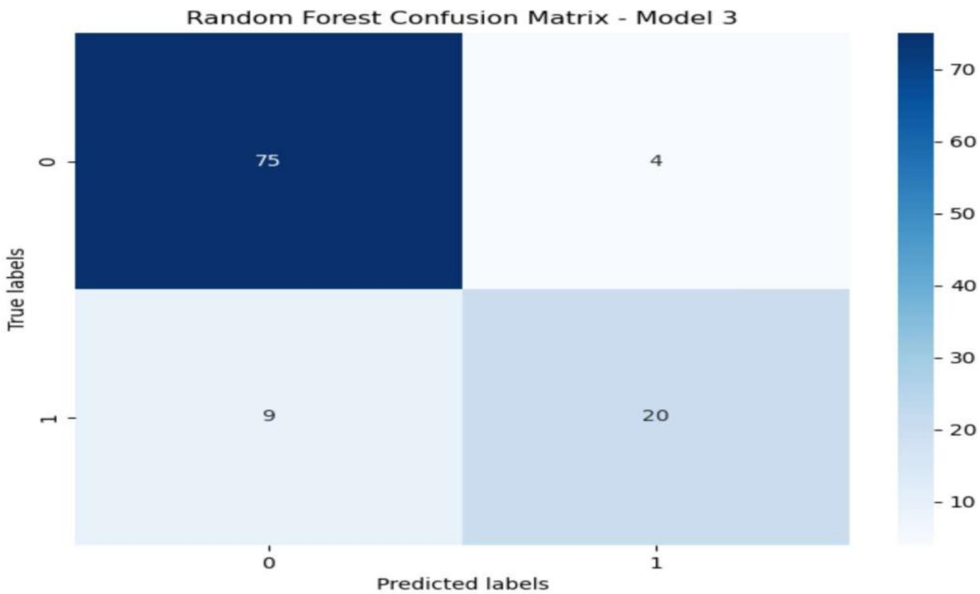

Figure 5: Random Forest Confusion Matrix for Model 3

5. CONCLUSION

Polycystic Ovary Syndrome (PCOS) is a common cause of infertility in women, affecting many of them at reproductive age and sometimes continuing long after they reach childbearing age. It is an indisputable fact that there have been ongoing efforts in recent times to utilize different traditional machine learning techniques in early detection of Polycystic Ovary Syndrome. To this end, this research aim to experiment the performance of random forest in Polycystic Ovary Syndrome Detection by rigorously considering the five models of random forest based on the five parameters that is the estimators, bootstrap, depth, maximum features, and maximum leaf nodes. Models 1 and 3 used a total of 300 estimators, had a true bootstrap, they had no depth and no maximum leaf nodes and the maximum features used by both models is square root. Models 2 and 4 used 200 estimators, they both had a true bootstrap, a depth of 20, no maximum leaf nodes and a maximum features of square root. Model 5 used a total of 100 estimators, a true bootstrap, a depth of 20, no maximum leaf node, and a maximum feature of square root. Models 1, 2, 3, 4, and 5 achieved and accuracy of 87.04%, 86.11%, 87.96%, 87.07%, and 87.04% respectively. However, surprisingly model 3 gains a better accuracy compared to the previous models. This gain would have been attributed to the number of estimators but conversely, the accuracy of model 1 does not suggest it; this is because both models used the same number of estimators. Thus, model 3 outperformed the other models.

REFERENCES

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

Chen, W., Yang, Q., Hu, L., Wang, M., Yang, Z., Zeng, X., & Sun, Y. (2023). Shared diagnostic genes and potential mechanism between PCOS and recurrent implantation failure revealed by integrated transcriptomic analysis and machine learning. *Frontiers in Immunology*, *14*, 1–15. https://doi.org/10.3389/fimmu.2023.1175384

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

Ekundayo, A., Aminu, E. F., & Cosmas, U. U. (2022). Improving the Accuracy of PCOS ' Data Prediction Model Based on Data Balancing and Multilayer Feature Selection. *International Conference on Informaion Systems and Emerging Technologies (ICISET-2022)*, 736–746.

Ganie, M. A., Rashid, A., Baba, M. S., Zargar, M. A., Wani, I. A., Nisar, S., & Sreenivas, V. (2023). Pre-polycystic ovary syndrome and polymenorrhoea as new facets of polycystic ovary syndrome (PCOS): Evidences from a single centre data set. *Clinical Endocrinology*, *99*(6), 566–578. https://doi.org/https://doi.org/10.1111/cen.14964

Harish, K. P., Dhivyanchali, M. N., Devi, K. N., Krishnamoorthy, N., Sree, R. D., & Dharanidharan, R. (2023). Smart Diagnostic System For Early Detection And Prediction Of Polycystic Ovary Syndrome. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6. https://doi.org/10.1109/ICCCI56745.2023.10128560.

Hassan, M. M., & Mizra, T. (2020). Comparative Analysis of Machine Learning Algorithms in Diagnosis of Polycystic Ovarian Syndrome. *International Journal of Computer Applications*, *175*(17), 42–53. https://doi.org/10.5120/ijca2020920688

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

Ionescu, O., Frincu, F., Mehedintu, A., Plotogea, M., Cirstoiu, M., Petca, A., Varlas, V., & Mehedintu, C. (2023). Berberine — A Promising Therapeutic Approach to Polycystic Ovary Syndrome in Infertile / Pregnant Women. *Life*, *13*(125), 1–13. https://doi.org/https://doi.org/10.3390/life13010125

Jaiswal, G., Bhardwaj, G., Tarushi, S. A., & Rani, R. (2024). Predictive Modeling to Identify Syndrome Patterns. *Cham: Springer Nature Switzerland.*, *11*, 67–91. https://doi.org/https://doi.org/10.1007/978-3-031-65434-3_4

Kaushik, S., & Mishra, S. K. (2023). Prediction of PCOD using Machine Learning Algorithms. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT56998.2023.10307450.

Khushal, R., & Fatima, U. (2024). Fuzzy machine learning logic utilization on hormonal imbalance dataset. *Computers in Biology and Medicine*, *174*, 108429. https://doi.org/https://doi.org/10.1016/j.compbiomed.2024.108429

Kumar, A., Sungh, J., & Khan, A. A. (2024). A comprehensive machine learning framework with particle swarm optimization for improved polycystic ovary syndrome (PCOS) diagnosis. *Engineering Research Express*, 1–30. https://doi.org/https://doi.org/10.1088/2631-8695/ad76f9

Kumari, R., Singh, J., & Gosain, A. (2023). SmS: SMOTE-stacked hybrid model for diagnosis of polycystic ovary syndrome using feature selection method. *Expert Systems with Applications*, *225*, 120102. https://doi.org/https://doi.org/10.1016/j.eswa.2023.120102

Mehr, H. D., & Polat, H. (2021). Diagnosis of polycystic ovary syndrome through different machine learning and feature selection techniques. *Health and Technology*, 1–15. https://doi.org/10.1007/s12553-021-00613-y

Mogos, R., Gheorghe, L., Carauleanu, A., Vasilache, I. A., Munteanu, I. V., Mogos, S., Solomon-Condriuc, I., Baean, L. M., Socolov, D., Adam, A. M., & Preda, C. (2024). Predicting Unfavorable Pregnancy Outcomes in Polycystic Ovary Syndrome (PCOS) Patients Using Machine Learning Algorithms. *Medicina (Lithuania)*, *60*(1298), 1–12. https://doi.org/10.3390/medicina60081298

Mojahed, B. S., Ghajarzadeh, M., Khammar, R., & Shahraki, Z. (2023). Depression , sexual function and sexual quality of life in women with polycystic ovary syndrome (PCOS) and healthy subjects. *Journal of Ovarian Research*, *16*(105), 10–13. https://doi.org/https://doi.org/10.1186/s13048-023-01171-9

Mou, T. H., Jyoti, O., Ahmed, T., & Imam, R. (2023). A Comparative Study for Detecting Polycystic Ovary Syndrome Using A Machine Learning Framework. *26th International Conference on Computer and Information Technology (ICCIT)*, 1–7. https://doi.org/10.1109/ICCIT60459.2023.10441348

Mridul, A. H., Ahsan, N., Alam, S. S., & Afrose, S. (2024). Polycystic Ovary Syndrome (PCOS) Disease Prediction Using Traditional Machine Learning and Deep Learning Algorithms. *International Journal of Computer Information SYstems and Industrial Management Applications*, *16*, 322–346.

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.

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

Prasher, S., Nelson, L., & Sharma, A. (2023). Evaluation of Machine Learning Techniques to Diagnose Polycystic

Ovary Syndrome Using Kaggle Dataset. *In International Conference on Emerging Trends in Expert Applications & Security*, 279–287. https://doi.org/https://doi.org/10.1007/978-981-99-1946-8_25

Rahman, M. M., Islam, A., Islam, F., Zaman, M., Islam, R. M., Sakib, A. S. M., & Babu, H. H. M. (2024). Empowering early detection : A web-based machine learning approach for PCOS prediction. *Informatics in Medicine Unlocked*, *47*, 1–16. https://doi.org/10.1016/j.imu.2024.101500

Raut, K., Katkar, C., & Itkar, P. S. A. (2022). PCOS Detect using MAchine Learning ALgorithms. *International Research Journal of Engineering and Technology (IRJET)*, *9*(1), 1214–1218.

Roy, D. G., & Alvi, P. A. (2022). Artificial intelligence in diagnosis of polycystic ovarian syndrome. *In Contemporary Issues in Communication, Cloud and Big Data Analytics: Proceedings of CCB 2020*, 453–463. https://doi.org/https://doi.org/10.1007/978-981-16-4244-9_37

Simi, M. S., Nayaki, S. K., Parameswaran, M., & Sivadasan, S. (2017). *Exploring Female Infertility Using Predictive Analytic*. 1–6.

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

Tiwari, S., Kane, L., Koundal, D., Jain, A., Alhudhaif, A., Polat, K., & Althubiti, S. A. (2022). SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. *Expert Systems with Applications*, *203*, 117592. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117592

Wang, R., Gu, Z., Wang, Y., Yin, X., Liu, W., Chen, W., & Qian, K. (2022). A "one-stop shop" decision tree for diagnosing and phenotyping polycystic ovarian syndrome on serum metabolic fingerprints. *Advanced Functional Materials*, *32*(45), 2206670. https://doi.org/doi.org/10.1002/adfm.202206670

Wu, X., Wang, C. C., Cao, Y., Li, J., Li, Z., Ma, H., Gao, J., Chang, H., Zhang, D., Cong, J., Wang, Y., Wu, Q., Han, X., Chung, P. W. J., Li, Y., Zheng, X., Chen, L., Zeng, L., Borchert, A., … Shi, Y. (2023). Novel Genetic Risk and Metabolic Signatures of Insulin Signaling and Androgenesis in the Anovulation of Polycystic Ovary Syndrome. *Engineering*, *23*, 103–111. https://doi.org/10.1016/j.eng.2022.08.013