Comparative Study on Learning Based Oversampling Model for Prediction of PCOS

Pijush Dutta[1*], Arindam Sadhu [2], Gour Gopal Jana[3] , Suchismita Maiti [4], Soumya Bhattacharyya[5], Shambhu Nath Saha[6], Sourav Saha[7]

[1,2, 3]Department of Electronics & Communication Engineering, Greater Kolkata College Engineering & Management, Baruipur, West Bengal , India, Pin: 743387

[4,5,6]Department of Information Technology, Narula Institute of Technology, West Bengal, India, Pin: 700109

[7]Department of Computer Science and Engineering, Narula Institute of Technology, West Bengal, India, Pin: 700109

Abstract:

From the survey it has been seen that upto 18% of woman all around the world of reproductive age suffer by a well-known reproductive endocrinopathies disease known as polycystic ovarian syndrome (PCOS). The earliest possible diagnosis and treatment of this condition has drawn research interest. Considering the serious imbalance of PCOS detection datasets will result in low classification performance and difficulty to detect the disease accurately and efficiently. In this study, the performance of two oversampling methods (SMOTE and ADASYN) is examined in conjunction with the RF classifier model in order to considerably increase the model's performance and evaluation metrics. The suggested PCOS prediction model's framework is made up of three distinct levels. The most important features are chosen in the first layer utilizing correlation and the principle component analysis (PCA) technique. The suggested model is trained in the second layer, and in the third layer, its performance is assessed in terms of classification accuracy (CA), precision, recall (sensitivity), Matthews' correlation coefficient (MCC), and area under the ROC curve (AUROC).The proposed RF+ADASYN algorithm clearly outperforms its counterparts and achieves a remarkable accuracy of CA, F1 Score, Sensitivity, AUROC, MCC, Training time and Prediction time are 97.94, 93.02, 94.19, 92.89, 0.90825, 0.7714, 0.139 sec and 0.005 sec respectively. The acquired simulation results demonstrate the excellence and efficacy of our suggested model.

Keywords: PCOS, Machine learning, Classification, SMOTE, ADASYN

1. Introduction

A hormonal imbalance condition known as polycystic ovary disease (PCOD) or polycystic ovary syndrome (PCOS) affects women of reproductive age [1-2]. Based on the symptoms and by preventing long-term issues, early diagnosis and therapy can be used to control. A doctor can use ultrasonography to diagnose PCOS by counting the number and size of follicles in the ovaries [3]. However, this procedure requires a long time, good imaging quality, and great accuracy to identify PCOS. Examining biological variables like hormone levels is another method for detecting PCOS [4]. This study compares appropriate evaluation metrics for models of unbalanced data that have undergone pre-processing & over sampling using ADASYN [7–9] and SMOTE [2-6]. PCOS data sets that show varying levels of class inequality have been chosen. The random forest classifier, a canonical and widely used statistical learning model, will be used to categories the processed data sets. The following are the primary contributions of this work:

i) Using the feature selection approach with SMOTE & ADASYN, identification of most crucial PCOS patient characteristics.

(ii) Using Random Forest ML techniques to analyze the key characteristics of the PCOS dataset.

(iii) Comparing the testing accuracy and recall of the three algorithms RF, RF-SMOTE, and RF-ADASYN. The remainder of the essay is structured as follows. The literature assessment of prior work is introduced in Section 2 along with the data used in the paper. A theoretical underpinning of oversampling techniques, a dataset description, and the classifiers are described in Section 3. The procedure and the evaluation metrics are described in Section 4. Section 5 presents the findings of the present research, while Section 7 presents a conclusion.

2. Literature Review:

PCOS is detected and predicted using a variety of machine learning models [10]. The highest accuracy recorded by the best model, Random Forest, is 89.02% [11]. Datasets are resampled using a mix of SMOTE (Synthetic Minority Oversampling Techniques) & ENN (Edited Nearest Neighbor) [12] to provide an effective classification performance for PCOS. According to the experimental findings, the Extreme Gradient Boosting classifier outperformed all other classifiers for a 10 fold cross validation [13]. PCOS on ultrasound pictures can be classified using a system that uses feature extraction and Competitive Neural Networks (CNN) [14]. The proposed model's highest accuracy and testing time were 80.84% and 60.64 seconds respectively. Five distinct machine learning techniques were used to diagnose PCOS data [15]. The utmost accuracy that random forest may achieve, according to the result analysis, is 96%.

Navies Bayes and an artificial neural network technique are combined in a unique hybrid structure to predict the likelihood of PCOS [16]. Rapid Miner and Python were used in a comparison study, and it was discovered from the results analysis that Random Forest performed better than the other categorization when used with Rapid Miner [17]. For the PCOS dataset, four classification algorithms and five feature selection approaches were used to predict the disease [18–20].

3. Descriptions of Data

The data sets and the elimination of observations and variables are explained in this section. The number of observations, number of variables, and level of class imbalance are all included in the description of the data sets. Single data sets with binary response variables make up the data. They were taken from the 541 records and 41 attributes of the UCI PCOS without Fertility dataset [1]. The minority class accounts for 32.90% of all observations in this data set. By applying the data cleaning and dimensioning technique the numbers of variables of PCOS dataset (without fertility) were shrunk from 41 to 12 variables. Table 1 summarizes the number of observations, number of variables & class imbalance ratio.

Table 1. Description of the Dataset

Dataset	Description of the Response variable	Number of observation	Number of variables	Class imbalance ratio
PCOS without fertility	Predicting the status of PCOS	541	12	32.90%

3.1 Over-Sampling Techniques

The theory underlying the over-sampling methods SMOTE and ADASYN is discussed in this subsection [21–23]. This involves concise explanations of their operation as well as knowledge of how they provide novel observations for the minority class.

3.1.1 Synthetic Minority Over-sampling Technique

Chawla [24] proposed SMOTE, which produces synthetic observations for the minority class. Synthetic observations are created between a particular minority class observation and its k-nearest minority class neighbors for that observation. For each observation of a minority class, this process is followed. The k-nearest neighbor count for SMOTE is set to 5. Prior to the

method, the quantity of generated synthetic observations is specified, and it should reflect the level of imbalance [25].

3.1.2 Adaptive Synthetic sampling approach

He et al. [26] put up the idea of ADASYN, which functions similarly to SMOTE in that it produces synthetic observations for the minority class. However, it is predicated on producing more synthetic data for observations for a particular model that are harder to learn than those that are easy to learn. A minority class observation and its k-nearest minority class neighbors are connected in a straight line by ADASYN, much like with SMOTE. The number of k-nearest neighbors is set to 5, just like SMOTE. But ADASYN produces more synthetic observations for minority class observations when there are more majority class observations in the region of the nearest neighbors [27]. On the other hand, no synthetic observations will be produced for a minority observation if there are no majority observations inside its k-nearest neighbor range. The justification for this is because learning from these observations is more difficult than from minority observations that are located far from the majority observations.

3.2 Learning models

3.2.1 Random Forest Classifier model

The random forest classifier is based on classification trees that are decision trees used to predict qualitative responses [28-31]. According to decision trees are made by dividing the predictor space $x_1, \ldots\ldots x_n$ into n distinct and no overlapping regions $R_1, \ldots. R_n$. Then, the same prediction is made for every observation that falls into the region R_n, which in the classification setting is the majority group that occupies that specific region, which again can be regarded as Bayes classifier. The rule by which the predictor space is partitioned is called recursive binary splitting, in which the predictor space is split iteratively based on the highest reduction of some measure of classification error. More formally, consider the predictor $x_n$and the cut point s. Then recursive binary splitting is done by splitting the predictor space into the regions. In the classification setting one measure that is often used for splitting is the Gini index, which is defined as

$$G = \sum_{k=1}^{K} \widehat{p_{mk}} \left(1 - \widehat{p_{mk}}\right)$$

Here, $\widehat{p_{mk}}$ is the proportion of training observations in the mth region that belong to the kth class Applying decision trees in the learning setting will likely lead to over fitting the data.

4 Proposed model

The numerous ML principles that are employed to address the aforementioned issues with the current PCOS prediction system were covered in the section prior. Pre-processing layer, dimensionality reduction layer, training layer, and performance evaluation layer make up the general architecture of our suggested prediction model. In the subsections below, we discuss how these layers work.

4.1 Dimensionality reduction layer

The input variables affect how well a machine learning algorithm performs. The performance of machine learning algorithms suffers when there are more input variables. The output of ML algorithms that fit on data with a variety of input properties may be dramatically impacted by this. Using correlation, feature selection is used in this layer to eliminate unimportant characteristics. This can be done with the help of GA [32-34], PSO [35], FPA [34,35], PCA, [36-38] and other bio-inspired algorithms. PCA is employed in this study to replicate the high-quality dataset. Due to this decrease in dataset dimension, over fitting problems are resolved by reducing the complexity of training.

4.2 Training layer

The DT-based prediction model is trained in this layer utilizing various splits of the training data set. The training dataset consists of several characteristics and their corresponding class labels. The gain ratio notion is used to preprocess a set of training data before the model training phase even begins. Each sample in a training set is an n-dimensional vector that contains the sample's feature values as well as the class to which it belongs.

4.3 Performance evaluation layer

This layer is employed to assess the model's efficacy. Various measures, including CA, accuracy, sensitivity, MCC, and AUROC, are used to assess the performance of the proposed prediction model on the PCOS dataset [28-30]. The structure of the suggested prediction model is shown in Figure 1.

4.4 Evaluation metrics

The methods and metrics employed to assess the model outputs are described in this section. The chosen assessment criteria take into account both a classifier's general classification performance and its ability to accurately categories minority data.

4.5 Matthews correlation coefficient

A metric used to assess the output of classification models, frequently in the binary classification situation, is the Matthews correlation coefficient (MCC) [39]. In a contingency table, such as the

confusion matrix, it is a method for determining the Pearson correlation coefficient between actual and predicted values. MCC should be preferred to the Fmeasure by all scientific communities when evaluating binary classification models [39–41]. The reason is that the MCC generates results that reflect the overall predictions made by a model, which is not the case for the F-measure. Although it is not as widely used as the F-measure, its usefulness has been shown in different scientific fields when it comes to evaluating predictive models.

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)+(TN+FN)}} \quad (1)$$

As the Pearson's correlation coefficient ranges between (-1, +1). For misclassification value of MCC is -1 , for perfect classification it will be 1 & coin tossing classification It is 0. The MCC will give an idea of the overall classification performance of the model.

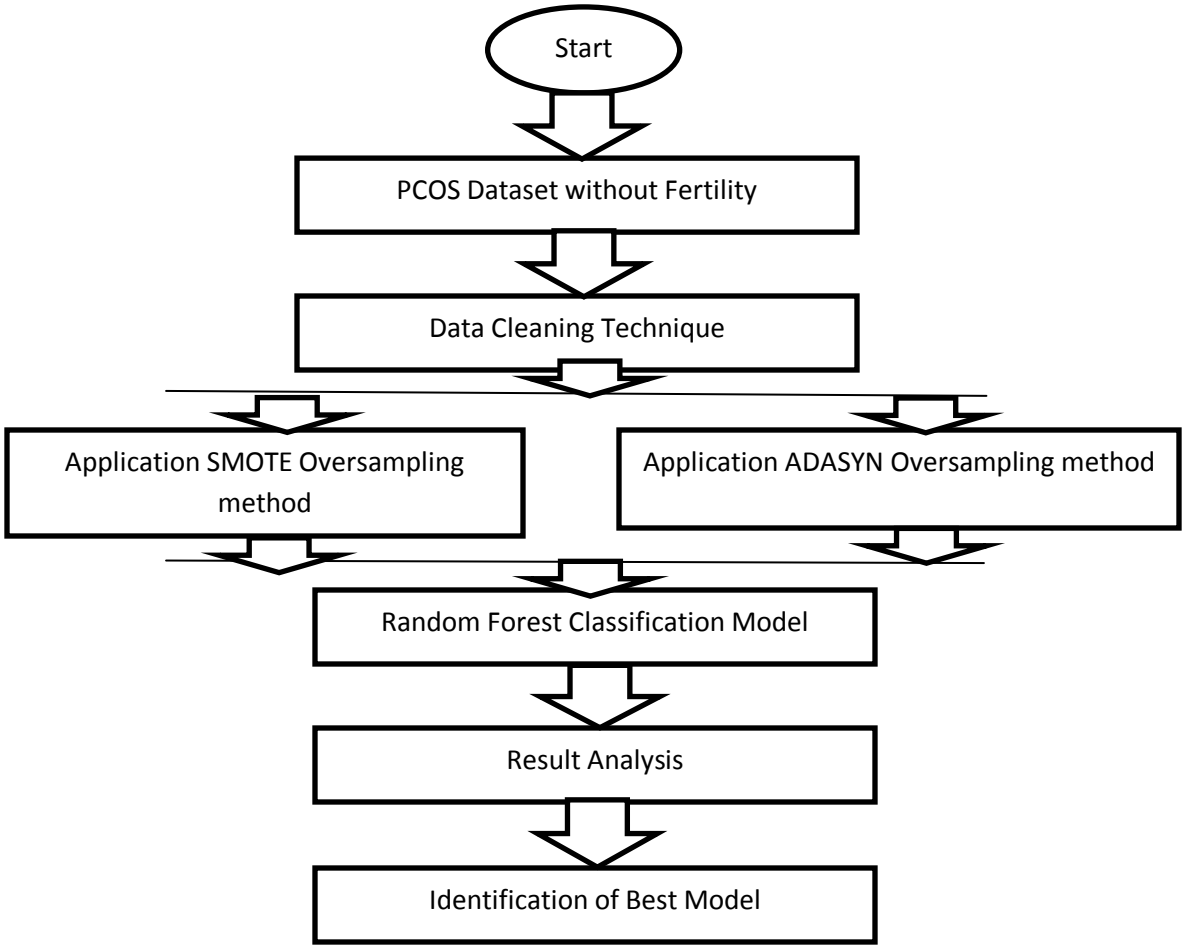

Figure 1. Proposed model for PCOS prediction

5 Results

The findings of applying pre-processing techniques to the minority class proportions of the training sets are reported in this section. Additionally, utilizing a variety of pre-processing techniques and models for the various data sets, the sensitivity, F-measures, and MCCs are shown along with the confidence intervals. Table 2 displays the class imbalance for each data set with and without pre-processing, after implementing SMOTE, and ADASYN.

Table 2. Description of the Dataset

Name	No Preprocessing	Using after SMOTE	Using after ADASYN
PCOS without fertility	32.90	50.21%	50.71%

Table 3 displays the sensitivity, F-measure, and MCC for each combination of pre-processing technique and classification model. We evaluated the suggested model for 80%–20% of the split test. The first component in this equation denotes the size of the training set, and the second component denotes the size of the testing set. The best five results are collected for each data set after the procedure is simulated ten times for each split. For training and testing, the three dataset types PCOD+RF, PCOD+SMOTE+RF, and PCOD+ ADASYN+RF are employed.

The CA is a popular performance metric for PCOD research in ML. Due to the class imbalance in the diabetic dataset (like PCOD), CA alone is insufficient to assess the system's effectiveness. According to the related work, CA alone is insufficient for assessing efficiency. The three following simulation scenarios are run in order to evaluate and compare the proposed prediction model. the trained model is assessed using many measures, including CA, F1 Score, Sensitivity, MCC, and AUROC. The suggested prediction model's results are contrasted with those of other traditional, already-in-use methods in terms of CA, F1 Score, Sensitivity, MCC, and AUROC.

Table 3. 80-20 Training testing result analysis using different pre-processing methods

Method	CA	F1 Score	Sensitivity	AUC	MCC
Random Forest (RF)	88.99	89	90.17	0.81521	0.5471
Hybrid RF+SMOTE	97.11	92.96	92.56	0.89741	0.77256
Hybrid RF+ADASYN	**97.94**	**93.02**	**94.19**	**0.90825**	**0.7714**

Table 4. Training & Prediction time for the different model

Method	RF	RF+MOTE	RF+ADASYN
Training Time (Sec)	0.139	0.061	0.063
Prediction Time(sec)	0.009	0.005	0.005

Table 5 Performance evaluation with other existing methods

Source	CA	Precision	Sensitivity	AUC
(Denny et al., 2019)	89.02	88.20	87.96	0.7851
(Satish et al., 2020)	93.12	91.56	91.23	0.8324
(Dutta et al.,2022)	97.11	93.85	93.79	0.8725
Proposed Work	97.94	94.12	94.19	0.9082

Following observations are noted in this simulation strategy with respect to accuracy.

- The RF+ADASYN combination produces the best results. The results of the performance metrics CA, F1 Score, Sensitivity, AUROC, MCC, Training time, and Prediction time, respectively, are 97.94, 93.02, 94.19, 92.89, 0.90825, 0.7714, 0.139 sec, and 0.009 sec.
- On RF+SMOTE, the second-best result is seen. The results of the performance metrics CA, F1 Score, Sensitivity, AUROC, MCC, Training time, and Prediction time are, respectively, 97.11, 92.96, 92.56, 0.899741, 0.77256, 0.061sec, and 0.005sec.
- On RF , the third best result is obtained. The results of the performance metrics CA, F1 Score, Sensitivity, AUROC, MCC, Training time, and Prediction time are, respectively, 88.99, 89, 90.17, 0.81521, 0.5471, 0.063 sec & 0.005 sec respectively.

5.1 Performance evaluation with existing systems

The value of TP and TN here means that they have the most significant influence on the resulting accuracy. The TP and TN values of the ADASYN-RF method are higher than that of SMOTE-RF[42] . The SMOTE-RF model in predicting the FN value also has more errors. The metrics like CA, accuracy, sensitivity, and AUROC of the suggested technique are compared with existing methods indicated in Table 5, the suggested model outperforms the other existing methods in terms of results. The RF+ADASYN data set shows the best results. 97.94, 94.12, 94.119, and 0.9082, respectively, are the results of the performance metrics CA, precision, sensitivity, and AUROC that were taken into consideration. The outcome from RF+SMOTE shows the second-best result. The results of the performance metrics CA, precision, sensitivity,

and AUROC, which were taken into consideration, are 97.11, 93.85, 93.79, and 0.8725, respectively.

6. Conclusion

The results imply that there is no pre-processing technique that consistently enhances the sensitivity, F-measure, and MCC performance of all the models. In this study, a brand-new RF+ADASYN prediction model is put out for the categorization of PCOS that also takes into account the issues with data imbalance and the curse of data dimensionality. Dealing with unbalanced data sets is challenging since most AI algorithms ignore the minority class, producing unreliable findings. In this regard, the proposed model makes use of correlation and PCA to extract important features while using ADASYN & SMOTE to oversample the minority class in its pre-processing step. The training and testing sets are created based on the results of feature selection. The suggested prediction model is trained using the training set, and its effectiveness is evaluated using the testing set. In terms of a number of measures, including CA, F1 Score, Sensitivity, AUROC, MCC, Training time, and Prediction time, the suggested model performs better than the existing models. The best results obtained by the suggested system are 97.94, 93.02, 94.19, 92.89, 0.90825, 0.7714, 0.139 sec, and 0.009 sec, respectively, in terms of CA, F1 Score, Sensitivity, AUROC, MCC, Training time, and Prediction time.

Future research can assess the proposed model's accuracy for automatic PCOS analysis and prediction. A fascinating area for future research could be to refine the rule sets of the suggested model. Furthermore, the implementation of various nature-inspired optimizations may be looked into in order to improve accuracy, decrease the size of the dataset, and minimize time complexity.

References:

1. Dutta, P., Paul, S., & Majumder, M. (2021). An Efficient SMOTE Based Machine Learning classification for Prediction & Detection of PCOS.
2. Dutta, P., Paul, S., Sadhu, A., Jana, G. G., & Bhattacharjee, P. (2023, February). Performance of Automated Machine Learning Based Neural Network Estimators for the Classification of PCOS. In Doctoral Symposium on Human Centered Computing (pp. 65-73). Singapore: Springer Nature Singapore.
3. Pathak, G. (2015). Polycystic ovary syndrome in contemporary India: An ethnographic study of globalization, disorder, and the body (Doctoral dissertation, The University of Arizona).

4. Hahn, S., Kuehnel, W., Tan, S., Kramer, K., Schmidt, M., Roesler, & Janssen, O. E. (2007). Diagnostic value of calculated testosterone indices in the assessment of polycystic ovary syndrome.
5. Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012). SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and information systems*, *33*(2), 245-265.
6. Alsawalqah, H., Faris, H., Aljarah, I., Alnemer, L., & Alhindawi, N. (2017, April). Hybrid SMOTE-ensemble approach for software defect prediction. In *Computer science on-line conference* (pp. 355-366). Springer, Cham.
7. Brandt, J., & Lanzén, E. (2021). A comparative review of SMOTE and ADASYN in imbalanced data classification.
8. Kurniawati, Y. E., Permanasari, A. E., & Fauziati, S. (2018, August). Adaptive synthetic-nominal (ADASYN-N) and adaptive synthetic-KNN (ADASYN-KNN) for multiclass imbalance learning on laboratory test data. In *2018 4th International Conference on Science and Technology (ICST)* (pp. 1-6). IEEE.
9. Lu, C., Lin, S., Liu, X., & Shi, H. (2020, May). Telecom fraud identification based on ADASYN and random forest. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)* (pp. 447-452). IEEE.
10. Boomidevi, R., & Usha, S. (2021). Performance Analysıs of Polycystıc Ovary Syndrome (PCOS) Detectıon System Usıng Neural Network Approach. In *Data Engineering and Communication Technology* (pp. 449-459). Springer, Singapore.
11. 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*, 117592.
12. Inan, M. S. K., Ulfath, R. E., Alam, F. I., Bappee, F. K., & Hasan, R. (2021, January). Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 1046-1050). IEEE.
13. Rathod, Y., Komare, A., Ajgaonkar, R., Chindarkar, S., Nagare, G., Punjabi, N., & Karpate, Y. (2022, July). Predictive Analysis of Polycystic Ovarian Syndrome using CatBoost Algorithm. In *2022 IEEE Region 10 Symposium (TENSYMP)* (pp. 1-6). IEEE.
14. Dewi, R. M., & Wisesty, U. N. (2018, March). Classification of polycystic ovary based on ultrasound images using competitive neural network. In *Journal of Physics: Conference Series* (Vol. 971, No. 1, p. 012005). IOP Publishing.
15. Hassan, M. M., & Mirza, T. (2020). Comparative analysis of machine learning algorithms in diagnosis of polycystic ovarian syndrome. *Int. J. Comput. Appl*, *975*, 8887.
16. Thomas, N., & Kavitha, A. (2020). Prediction of polycystic ovarian syndrome with clinical dataset using a novel hybrid data mining classif ication technique. *Int J Adv Res Eng Technol (IJARET)*, *11*(11), 1872-1881.

17. Nandipati, S. C., Ying, C. X., & Wah, K. K. (2020). Polycystic Ovarian Syndrome (PCOS) classification and feature selection by machine learning techniques. *Appl Math Comput Intell*, *9*, 65-74.
18. Boomidevi, R., & Usha, S. (2021). Performance Analysıs of Polycystıc Ovary Syndrome (PCOS) Detectıon System Usıng Neural Network Approach. In *Data Engineering and Communication Technology* (pp. 449-459). Springer, Singapore.
19. Sumathi, M., Chitra, P., Prabha, R. S., & Srilatha, K. (2021, February). Study and detection of PCOS related diseases using CNN. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1070, No. 1, p. 012062). IOP Publishing.
20. Wagh, P., Panjwani, M., & Amrutha, S. (2021). Early detection of PCOD using machine learning techniques. In *Artificial Intelligence and Speech Technology* (pp. 9-20). CRC Press.
21. Mansourifar, H., & Shi, W. (2020). Deep synthetic minority over-sampling technique. *arXiv preprint arXiv:2003.09788*.
22. Tarawneh, A. S., Hassanat, A. B., Almohammadi, K., Chetverikov, D., & Bellinger, C. (2020). Smotefuna: Synthetic minority over-sampling technique based on furthest neighbour algorithm. *IEEE Access*, *8*, 59069-59082.
23. Huang, M. W., Chiu, C. H., Tsai, C. F., & Lin, W. C. (2021). On combining feature selection and over-sampling techniques for breast cancer prediction. *Applied Sciences*, *11*(14), 6574.
24. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
25. Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009, April). Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 475-482). Springer, Berlin, Heidelberg.
26. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE.
27. Gnip, P., Vokorokos, L., & Drotár, P. (2021). Selective oversampling approach for strongly imbalanced data. *PeerJ Computer Science*, *7*, e604.
28. Zigarelli, A., Jia, Z., & Lee, H. (2022). Machine-Aided Self-diagnostic Prediction Models for Polycystic Ovary Syndrome: Observational Study. *JMIR Formative Research*, *6*(3), e29967.
29. Inan, M. S. K., Ulfath, R. E., Alam, F. I., Bappee, F. K., & Hasan, R. (2021, January). Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 1046-1050). IEEE.

30. Aggarwal, S., & Pandey, K. (2021). An Analysis of PCOS Disease Prediction Model Using Machine Learning Classification Algorithms. *Recent Patents on Engineering*, *15*(6), 53-63.
31. Pratap, N. L., Vallabhuni, R. R., Babu, K. R., Sravani, K., Kumar, B. K., Srikanth, A., ... & Mohan, K. S. K. (2020). A Novel Method of Effective Sentiment Analysis System by Improved Relevance Vector Machine. *Australian Patent AU*, *2020104414*, 31.
32. Guo, Q., Wu, W., Massart, D. L., Boucon, C., & De Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, *61*(1-2), 123-132.
33. Tao, Z., Huiling, L., Wenwen, W., & Xia, Y. (2019). GA-SVM based feature selection and parameter optimization in hospitalization expense modeling. *Applied soft computing*, *75*, 323-332.
34. Dutta, P., Paul, S., Jana, G. G., & Sadhu, A. (2023, May). Hybrid Genetic Algorithm Random Forest algorithm (HGARF) for improving the missing value Imputation in Hepatitis Medical Dataset. In *2023 International Symposium on Devices, Circuits and Systems (ISDCS)* (Vol. 1, pp. 01-05). IEEE.
35. Chuang, L. Y., Chang, H. W., Tu, C. J., & Yang, C. H. (2008). Improved binary PSO for feature selection using gene expression data. *Computational Biology and Chemistry*, *32*(1), 29-38.
36. Zawbaa, H. M., & Emary, E. (2018). Applications of flower pollination algorithm in feature selection and knapsack problems. In *Nature-Inspired Algorithms and Applied Optimization* (pp. 217-243). Springer, Cham.
37. Sen, S., Saha, S., Chaki, S., Saha, P., & Dutta, P. (2021). Analysis of PCA based AdaBoost machine learning model for predict mid-term weather forecasting. *Computational Intelligence and Machine Learning*, *2*(2), 41-52.
38. Dutta, P., Shaw, N., Das, K., & Ghosh, L. (2021). Early & accurate forecasting of mid term wind energy based on PCA empowered supervised regression model. *Computational Intelligence and Machine Learning*, *2*(1), 53-64.
39. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1-13.
40. Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, *14*(1), 1-22.
41. Zhu, Q. (2020). On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognition Letters*, *136*, 71-80.
42. Dutta, P., Paul, S., & Majumder, M. (2021). Intelligent SMOTE Based Machine learning Classification for Fetal State on Cardiotocography Dataset.